VieCtoS - The Vienna Concept to Speech System*

Kai Alter, Ernst Buchberger, Johannes Matiasek, Georg Niklfeld, Harald Trost

Abstract

VieCtoS is a system producing spoken utterances in German from information stored in a database. An important aspect of the system is the use of information relevant for intonation at all stages of the generation process. Prosodic features are derived from pragmatic information available in the strategic generator, and transformed by the tactical generator into features that can be interpreted by the synthesizer. The application domain is the generation of spoken weather forecasts from periodically incoming weather data.

VieCtoS ist ein System, das aus in einer Datenbank gespeicherter Information Äußerungen in gesprochener deutscher Sprache produziert. Ein wichtiger Aspekt ist die Integration intonationsrelevanter Information in alle Stufen des Generierungsprozesses. Prosodische Merkmale werden aus im strategischen Generator vorhandener pragmatischer Information abgeleitet und vom taktischen Generator in vom Synthetisator interpretierbare Merkmale übersetzt. Anwendungsbereich ist die Erzeugung von gesprochenen Wetterprognosen aus periodisch einlangenden Wetterdaten.

1 Introduction

A quickly growing number of applications employs synthesized speech. While at the moment most industrial applications use resynthesized speech in the form of templates, a large number of potential application fields requires the flexibility of fully synthesized speech. Unfortunately, the quality of the speech produced by today's synthesizers is limited. While segmental quality has improved considerably acceptable prosodic quality still poses a problem.

Mostly, current speech synthesizers are used in so-called text-to-speech systems where text in graphemic form is input to the system. This severely inhibits the systems' possibility to produce high quality prosody, since the necessary information is not available. The potential for improvement in these systems through reanalysis of the text is quite limited.

*This work has been sponsored by the Austrian Fonds zur Förderung der Wissenschaftlichen Forschung under grant no. P10822-MAT. Financial support for the Austrian Research Institute for Artificial Intelligence is provided by the Austrian Federal Ministry of Science, Transport, and the Arts.

A more promising approach to solve the problem are concept-to-speech systems where by definition the necessary pragmatic and semantic information is available from the generator and therefore directly accessible to the synthesizer.

Still, a number of problems have to be solved before a system with high quality prosody can be implemented. In this paper we give an overview of VIECTOS, a concept-to-speech system for German where prosody is explicitly determined by pragmatic features.

2 The basic generation process

The Austrian weather service produces periodic forecasts for essential weather data like temperature, humidity, wind direction and speed, and precipitation. These data are specified for equi-distant points evenly distributed over Austrian territory and are given in 6-hourly intervals for three days ahead. In VIECTOS' *conceptualizer*, these data are aggregated temporally and locally to come up with a description of the expected weather situation for the target area over the next 24 hours. This implies translating quantitative data like degrees of temperature or m/s of wind speed into terms like HOT or STRONG. The data are set into context by comparing them to the current weather situation and to the average values for the time of the year by means of a domain model in LOOM (MacGregor and Bates, 1987). This enables the generation of a conceptual representation of the weather situation, which forms the input for the generator.

The *strategic generator* breaks up this conceptual description into sentence-size chunks, provides for the correct linear ordering, and performs word selection. To achieve these goals, the strategic generator makes use of text plans (Garner, 1995b) formulated using the production rules of LOOM. The output of this component is a sequence of expressions in the sentence plan language SPL, forming the input for the tactical generator.

The *tactical generator* was originally developed for the use within a multilingual text generation system. It has been built by combining preexisting resources: FUF (Elhadad, 1991), a unification-based generation engine, an HPSG grammar for German, and X2MorF (Trost, 1991), an extended two-level morphology component. The integration of these resources is described in (Matiasek and Trost, 1996).

A distinctive feature of the tactical generator is the tight integration of syntactic and morphological processing within the same grammar. A word-level grammar accounts for the proper combination of stems and affixes. The affixes to be chosen are determined by the syntactic features of the root form established during generation on the syntactic level. FUF's linearization component is augmented by the two-level module transforming the concatenation of roots and affixes to the appropriate graphemic surface representation.

For the requirements of VIECTOS the morphological component had to be adapted to work with a phonemic description, i.e., lexicon, word grammar and two-level rules had to be replaced (see section 4).

3 Prosody and its integration into the generator

We have analyzed a number of actual weather reports both with regard to discourse structure and to intonation. Discourse structure analysis is based on the work of (Grosz and Sidner, 1986), intonation on (Pierrehumbert, 1980).

Based on these analyses, we have formulated an intonational model using different stress patterns (Garner, 1995a). We have implemented a number of phrasal stress rules depending on constituent type and focus type. These determine the placement of the stress feature. The creation of intonation markers has been integrated into the strategic generator's text plans.

The tight integration of syntactic and morphological processing in the tactical generator mentioned above allows to make use of syntactic and semantic features for the realization of intonation.

4 The phonological component

For the use in a speech-producing system the two-level module of the tactical generator has been extended. Taking as its input the structural descriptions derived in tactical generation, the phonological component computes a string-based phonological data structure, which contains all the information needed in the final synthesis step to determine prosodically adequate speech. The phonological component is implemented in extended two-level phonology, and is characterized particularly by two properties:

First, the system employs strings and two-level string-processing to represent collapsed multilinear phonological structures comprising the dimensions of segmental quality, morphological and phonological boundaries, syllabification, stress, and tones. This integrated treatment is suggested by the numerous mutual dependences holding between the various dimensions in the phonology of German (Niklfeld, Pirker, and Trost, 1995).

Second, in extended two-level phonology the application of phonological rules may depend on filters on grammatical features. Thus, rules can inspect features from all levels of grammatical representation. Thereby, VieCtoS can take advantage of the availability of rich syntactic, semantic and pragmatic information for the assignment of prosody. Consider, e.g., the compound *Wétter+beruhigung*, which gets stressed *Wetter+berúhigung* in weather reports, where *Wetter* is *textually exophoric* (Benware, 1987).

As argued in (Niklfeld and Alter, 1996), a phonological component of this type is well suited for synthesizing intonation and timing in a demi-syllable based system. Mediating between the symbolic linguistic descriptions in tactical generation and speech processing, the architecture provides enough flexibility to account for contextual influences on speech by phenomena at different positions on the phonology-acoustics continuum. Such phenomena can be modeled symbolically transparent in phonology; or via context-dependent parameters in demi-syllable concatenation; or implicitly in the inventory of speech units.

5 The synthesizer

In our demi-syllable speech synthesis system, syllables are divided into initial and final demi-syllables, and suffixes. This technical approach gives us a particularly good output quality of articulatory conditioned phenomena (e.g., co-articulation). In our inventory which is optimized following Portele (1994) more than 2000 demi-syllables are registered. The demi-syllables have been recorded from a male speaker of standard Austrian German. After segmentation and normalization of the recorded samples these are stored in LPC-encoded form. Since the residual signal is stored as well, a lossless reproduction of the recorded samples is possible. By this encoding the accomodation of pitch and length adjustments needed for obtaining correct intonation during resynthesis is facilitated. Input to the synthesizer is an annotated string specifying the demi-syllables to be spoken and providing information on the smoothing parameters for demi-syllable concatenation as well as information on duration, amplitude and fundamental frequency changes needed for both suprasegmental and micro prosody. The signal processing parts of the synthesizer have been developed in close cooperation with the Technical University of Vienna.

References

Benware, W.A. 1987. Accent variation in german nominal compounds of the type (a(bc)). *Linguistische Berichte*, (108):102–187.

Elhadad, Michael. 1991. Fuf: The universal unifier user manual, version 5.0. Technical report, Dept.of Computer Science, Columbia University, NY.

Garner, Elizabeth J. 1995a. The interpretation of prosodic meaning using discourse-level features for the purposes of text generation. Technical Report TR-95-38, Austrian Research Institute for Artificial Intelligence, Vienna, Austria.

Garner, Elizabeth J. 1995b. Knowledge resources for the text planner: The domain model and plans for discourse. Technical Report TR-95-13, Austrian Research Institute for Artificial Intelligence, Vienna, Austria.

Grosz, Barbara J. and Candice L. Sidner. 1986. Attention, intention, and the structure of discourse. *Computational Linguistics*, 12:175–204.

MacGregor, R. and R. Bates. 1987. The loom knowledge representation language. Technical Report TR ISI/RS-87-188, Information Sciences Institute, University of Southern California, Marina del Rey, CA.

Matiasek, Johannes and Harald Trost. 1996. An hpsg-based generator for german—an experiment in the reusability of linguistic resources. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING-96)*, Copenhagen.

Niklfeld, Georg and Kai Alter. 1996. Covering prosody in concept-to-speech via an extended two-level phonology component. In *Proc. SIGPHON 96*. Association for Computational Linguistics.

Niklfeld, Georg, Hannes Pirker, and Harald Trost. 1995. Using two-level-morphology as a generator-synthesizer interface for concept-to-speech. In *Proc. Eurospeech '95*, pages 1223–1226.

Pierrehumbert, Janet. 1980. *The Phonology and Phonetics of English Intonation.* Ph.D. thesis, MIT, Dept. of Linguistics.

Portele, T. 1994. *Ein phonetisch-akustisch motiviertes Inventar zur Sprachsynthese Deutscher Äußerungen.* Ph.D. thesis, Univ. of Bonn.

Trost, Harald. 1991. X2morf: A morphological component based on augmented two-level morphology. In *Proceedings of the 12th International Joint Conference on Artificial Intelligence (IJCAI-91)*, Sydney, Australia.