THE BELL LABS GERMAN TEXT-TO-SPEECH SYSTEM: AN OVERVIEW

Bernd Möbius, Richard Sproat, Jan P. H. van Santen, Joseph P. Olive

Bell Labs – Lucent Technologies
600 Mountain Avenue, Murray Hill, NJ 07974, USA
{bmo,rws,jphvs,jpo}@research.bell-labs.com

ABSTRACT

In this paper we present an overview of the German version of the Bell Labs text-to-speech system, a high-quality concatenative synthesis system with extensive text analysis capabilities. We discuss problems of text analysis, and our solutions to these problems, including: the integration of text normalization tasks into linguistic text analysis; the capability to morphologically analyze compounds and unseen words; name analysis and pronunciation. We briefly describe the prosodic components of the text-to-speech system and their underlying duration and intonation models. Finally, the phonetically motivated structure of the acoustic inventory is presented.

1. INTRODUCTION

The Bell Labs multilingual text-to-speech (TTS) system can be characterized as consisting of one single set of language-independent modules. Any language-specific information is represented in, and at run time retrieved from, precompiled tables, models and finite-state transducers. In this paper we present an overview of the German version of the Bell Labs TTS system. We will first discuss aspects of text analysis and our solutions to the problems they pose. Some of these problems, such as the expansion of numbers and abbreviations, and proper name pronunciation, occur in many languages while others, such as productive compounding, are specific to German and several related languages. We will then report on the construction of models for segmental duration and intonation. Finally, we will explain the structure of the acoustic inventory for concatenative synthesis and the criteria and procedures that were used to build it.

2. TEXT ANALYSIS

2.1. Technological Framework

Weighted finite state transducer (WFST) technology is the computational framework for text analysis in the multilingual TTS effort at Bell Labs [11]. All linguistic descriptions, such as morphologically annotated lexica, word and syllable models, language models, or phonological rewrite rules, are compiled into finite-state transducers (FST). Assigning weights, or costs, to certain paths through a finite-state machine is a convenient way to describe and predict linguistic alternations. While these descriptions are typically hand-built by experts, it is also possible to compile into FSTs data-based linguistic prediction models, such as decision trees. For the German TTS system, weights were derived from three types of information sources: (a) frequency distributions in specific databases (see the section on name analysis); (b) a model of productive word formation processes (see unknown word analysis); and (c) linguistic knowledge and intuition.

2.2. Integrated Text Analysis

By definition, TTS systems start by converting text written in the standard orthography of the language into linguistic representations from which synthesis can then proceed. But written language is at best an imperfect representation of linguistic structure in the sense that it is both ambiguous and incomplete and lacks information that is crucial for the proper pronunciation of words. Tokenization into sentences and words and expansion of numeral expressions and abbreviations are some of the problems one encounters while analyzing ordinary text. Performing these text normalization tasks in pre-processing steps, as it is done in conventional systems, often leads to incorrect analyses because sufficient context information is not available at the time the expansion is performed.

Consider, for example, the German sentence *Die Konferenz soll am 22.9.1997 beginnen.* ('The conference is supposed to begin on 9-22-1997.'). The numeral expression has to be recognized as a date, in which case the four digits *1997* should be expanded to *neunzehnhundert* (not *eintausend neunhundert*), and the numbers representing the day and month have to be interpreted as ordinals. A conventional pre-processor would then expand the ordinal numbers to their default word forms, which most likely is the nominative singular masculine: *zweiundzwanzigster neunter*. Without context information, this is the best guess text normalization can take; unfortunately, the expansion would be wrong. The correct solution *zweiundzwanzigsten neunten* can only be found if a special local grammar or language model is applied that enforces number, case, and gender agreement between the numeral expression and the preceding preposition, and rules out all non-agreeing alternatives. The example illustrates that the so-called text normalization tasks can best be handled by embedding them into linguistic analysis.

In the Bell Labs multilingual TTS systems, the generalized text analysis component integrates normalization tasks with other aspects of linguistic analysis, such as lexical analysis, morphology, and phonology. Lexical information is represented by mostly morphological annota-

tions of the regular orthography. The following is an example of an annotated lexicon entry, viz. a past tense form of the verb *veranschlagen* ('to estimate'), with labels for morphological properties, accent status, syllabic stress ('), morpheme boundaries (+), and word boundaries (#):

[#] [ACC:+] ver [+] 'an [+] schlag [verb] [+] test [sg] [2per] [past] [indi] [#]

Different types of lexical source files can be compiled into finite-state machines. For example, for words with complex inflectional morphology, such as nouns, adjectives and verbs, we first specify classes of inflectional paradigms in terms of sets of possible affixes and their linguistic features; the paradigms can be represented by a finite-state acceptor. We then list the stems of words that belong to each of the paradigm classes; the mapping between the classes and the lexical items is performed by an FST. The complete FST for inflected words results from the composition of the two machines. Uninflected and underived words can simply be compiled into finite-state acceptors. A special FST maps digit strings onto appropriate expansions as number names. Similarly, abbreviations and acronyms are expanded, and it is also possible to incorporate sub-lexica for specific domains, such as geographical names or foreign loan words.

2.3. Unknown Word Analysis

A crucial aspect of the TTS system is its capability to analyze compounds and unseen words. Any well-formed text input to a general-purpose TTS system in any language is likely to contain words that are not explicitly listed in the lexicon. The inventory of lexicon entries is unbounded; every natural language has productive word formation processes, and the community of speakers of a particular language can, and in fact does, create novel words as need arises. German in particular is notorious for productive compounding as a means to create neologisms. Thus, in unlimited vocabulary scenarios we are not facing a memory or storage problem but the requirement for the TTS system to be able to correctly analyze unseen orthographic strings. Our solution to this problem is a particular type of linguistic description, a compositional model that is based on the morphological structure of words and the phonological structure of syllables. This model is implemented in the form of a finite-state grammar for words of arbitrary morphological complexity.

The core of the module is a list of approximately 5000 nominal, verbal, and adjectival stems, augmented by about 250 prefixes and 220 suffixes that were found to be productive in a study on productive word formation [5]. Also included are infixes (*Fugen*) that German word formation grammar requires as insertions between components within a compounded word in certain cases, such as *Arbeit+s+amt* ('employment agency') or *Sonne+n+schein* ('sunshine').

Probably the two most important aspects of this finite-state grammar are, first, the decision which states can be reached from any given current state and, second, which of the legal paths through the graph should be preferred over other legal paths. The first aspect can be regarded as an instantiation of a declarative grammar of the morphological structure of German words; the second aspect reflects the degrees of productivity of word formation, represented by costs on the transitions between states.

The word model contains a phonotactic syllable model of German. The syllable model enables the word model to further analyze orthographic substrings that are unaccounted for by the explicit list of morphological elements, such as lexical roots and affixes, in arbitrary positions within a morphologically complex word. By assigning high costs to all paths that involve the application of the syllable model, preference is given to those paths that decompose a word into as many known components as possible. The actual costs of the syllable model are a function of the number of syllables in the residual string and the number of graphemes in each syllable.

The finite-state transducer that the declarative grammar for unknown word analysis is compiled into is far too complex to be usefully diagrammed. For the sake of exemplification, Figure 1 shows a graphical representation of the transducer corresponding to a small sub-grammar that analyzes and decomposes the morphologically complex word *Unerfindlichkeitsunterstellung* ('allegation of incomprehensibility'); this compound is novel in the sense that it is unknown to the system.

2.4. Name Analysis

The general approach to unknown word decomposition described above is also applied to the analysis of names in our TTS system. In a recent evaluation of the system's performance on street names, which encompass many aspects of geographical and proper names, we reported a pronunciation error rate by word of about 12% [3]. The study shows that German street names are amenable to morphological decomposition, because they are often constructed by word formation and derivation processes that are similar to those for regular words. The observed errors can be attributed almost exclusively to inherent problems of proper name pronunciation. The system's performance compares favorably with results obtained in the European Onomastica project [10], which reports a by-word accuracy rate of 71% (or an error rate of 29%) for names that were transcribed by rule only. So far, we have opted against adding a name-specific set of pronunciation rules to the general-purpose one; while such an approach has been shown to achieve lower error rates [1], it presupposes an unrealistically reliable detection of names in arbitrary text. Proper name pronunciation remains one of the most problematic tasks for TTS systems.

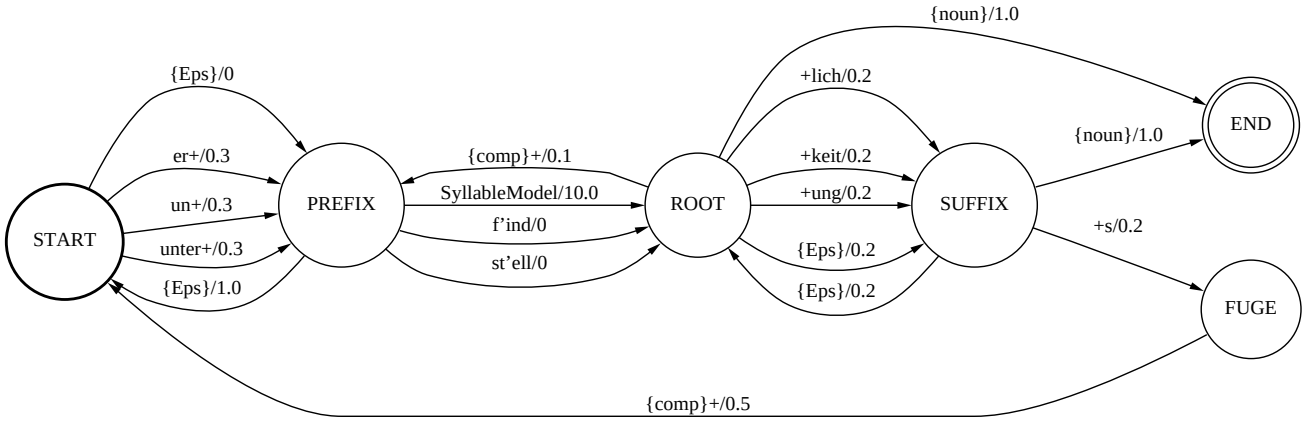

Figure 1: The transducer compiled from a sub-grammar that decomposes the morphologically complex word *Unerfindlichkeitsunterstellung*.

2.5. Language Models

Lexical disambiguation is performed by analyzing the local context. For instance, many abbreviations occurring after numbers are measure terms that can be expanded as either singular or plural forms, depending upon which number precedes the abbreviation (e.g., *1 t* $\rightarrow$ *eine Tonne* vs. *2 t* $\rightarrow$ *zwei Tonnen*). At the same time, the expanded word forms of both the number and the abbreviation have to agree in grammatical gender—and also in case, which requires further context analyses. Local language models of this type are usually implemented in the form of rules that describe which alternative lexical analysis is correct in a particular context.

2.6. Grapheme-to-Phoneme Conversion

The phonological component was implemented using an efficient algorithm for compiling weighted rewrite rules into FSTs [7]. The TTS system includes a set of approximately 200 general-purpose pronunciation rules, which draw upon the rich lexical and morphological annotation. The number of rules handling exceptions is minimal, at least as far as regular words of German are concerned.

3. PROSODY

In this section, we briefly discuss the duration and intonation components of the TTS system.

3.1. Segmental Duration

The task of the duration component of a TTS system is to reliably predict the duration of every speech sound depending upon a variety of contextual factors. Among the most important factors are the position of the word in the utterance, the accent status of the word, syllabic stress, and the segmental context. The solution proposed by van Santen [12] is to apply a class of arithmetic models known as sums-of-products models. This approach takes advantage of the fact that most interactions between factors are regular, which allows describing these interactions with equations consisting of sums and products. Addition and multiplication are sufficiently well-behaved mathematically to estimate parameter values even if the frequency distribution of feature vectors in the database is skewed. Following this general method, we constructed a model for segmental duration in German [6]. The procedure involved two phases, inferential-statistical analysis of a segmented speech corpus, and parameter estimation, and was made efficient by the use of an interactive statistical analysis package that is geared to van Santen's duration model. The results are stored in tables in a format that can be directly interpreted by the TTS duration module. The overall correlation between observed and predicted segmental durations is .896.

3.2. Intonation

The intonation component of the German TTS system computes a fundamental frequency (F_0) contour by adding up three types of time-dependent curves: a phrase curve, which depends on the type of phrase, e.g., declarative vs. interrogative; accent curves, one for each accent group (accented syllable followed by zero or more non-accented syllables); and perturbation curves, which capture the effects of obstruents on F_0 in the post-consonantal vowel. This approach shares some concepts with superpositional intonation models (e.g., [2, 4]) that have been applied to a number of languages. The key novelty in our approach is that we model in detail how accent curves depend on the composition and duration of accent groups. Earlier work [13] had shown that there is a relationship between accent group duration and F_0 peak location. Other important factors are the segmental structure of onsets and codas of stressed syllables. Based on these findings, the current model predicts F_0 peak location in a given accent group by computing a weighted sum of the onset and rhyme durations of the stressed syllable, and the duration of the remainder of the accent group; the model assumes

that the three factors exert different degrees of influence on peak location [14].

4. ACOUSTIC INVENTORY

Constructing acoustic inventories is a complex process. First, a speaker has to be selected according to a multitude of criteria. The second step, *inventory design*, is to set up a list of unit types that are to be recorded and excised, and the appropriate text materials. Next, for each unit type the best candidate token is selected (*unit selection*). Finally, the pertinent speech intervals are excised from the speech database and stored in the inventory.

4.1. Inventory Design

The majority of units in the acoustic inventory of the German TTS system are diphones. Inventory units are selected based on various criteria including spectral discrepancy and energy measures. Modeling certain contextual or coarticulatory effects can require the selection of context-sensitive units [8]; for instance, it may be necessary to have one /v-ɛ/ unit for intervocalic position as in *gewellt*, where the /v/ realization is fully voiced throughout, and a second /v-ɛ/ unit for the context of preceding voiceless obstruents as in *Schwester*, where the /v/ realization is entirely or partially devoiced. The current acoustic inventory consists of approximately 1250 diphonic units, including about 100 context-sensitive units. This inventory is sufficient to represent all phonotactically possible phone combinations for German. It was recently augmented by units representing speech sounds that occur in common foreign words or names, such as the interdental fricatives and the glide /w/ for English, and nasalized vowels for French.

4.2. Unit Selection

For acoustic inventory construction we use a procedure that performs an automated optimal unit selection and cut point determination [15]. The approach selects units such that spectral discrepancies between units as well as the distance of each sound to its 'ideal' are simultaneously minimized, and the coverage of required units is maximized. Diagnostic tools help reduce the amount of manual labor involved in the selection of inventory units.

5. SYNTHESIS

At run time, unit selection and concatenation modules retrieve the necessary units from the inventory, concatenate them and perform appropriate interpolation and smoothing operations, assign new durations, F_0 contours and amplitude profiles, and finally pass parameter vectors on to the synthesis module to generate the output speech waveform. Our TTS system uses standard LPC synthesis in conjunction with an explicit voice source model [9] that provides control of spectral tilt and the level of aspiration noise. This source generator is capable of modeling irregularities in the periodic component, such as vocal jitter, laryngealizations, and diplophonic double-pulsing. Thus, it opens up the possibility for the TTS system to vary the source parameters during an utterance and trigger voice quality changes according to the prosodic context.

6. CONCLUSION

We presented an overview of the German version of the Bell Labs multilingual TTS system. An interactive version of the system is accessible on the Web at `http://www.bell-labs.com/project/tts/german.html`.

7. REFERENCES

1. Karim Belhoula. A concept for the synthesis of names. In *ESCA Workshop on Applications of Speech Technology*, Lautrach, Germany, 1993.
2. Hiroya Fujisaki. A note on the physiological and physical basis for the phrase and accent components in the voice fundamental frequency contour. In Osamu Fujimura, editor, *Vocal physiology: Voice production, mechanisms and functions*, pages 347–355. Raven, New York, 1988.
3. Stefanie Jannedy and Bernd Möbius. Name pronunciation in German text-to-speech synthesis. In *Proc. 5th Conf. on Applied Natural Language Processing*, pages 49–56, Washington, DC, 1997.
4. Bernd Möbius. *Ein quantitatives Modell der deutschen Intonation—Analyse und Synthese von Grundfrequenzverläufen*. Niemeyer, Tübingen, 1993.
5. Bernd Möbius. Word and syllable models for German text-to-speech synthesis. Technical report, Bell Laboratories, 1996.
6. Bernd Möbius and Jan van Santen. Modeling segmental duration in German text-to-speech synthesis. In *Proc. ICSLP-96*, volume 4, pages 2395–2398, Philadelphia, PA, 1996.
7. Mehryar Mohri and Richard Sproat. An efficient compiler for weighted rewrite rules. In *Proc. 34th Annual Meeting of the ACL*, pages 231–238, Santa Cruz, CA, 1996.
8. Joseph P. Olive. A new algorithm for a concatenative speech synthesis system using an augmented acoustic inventory of speech sounds. In Gérard Bailly and Christian Benoit, editors, *Proc. ESCA Workshop on Speech Synthesis*, pages 25–29, Autrans, 1990. ESCA.
9. Luis C. Oliveira. Estimation of source parameters by frequency analysis. In *Proc. Eurospeech-93*, volume 1, pages 99–102, Berlin, 1993. ESCA.
10. Onomastica. Multi-language pronunciation dictionary of proper names and place names. Technical report, European Community, Ling. Res. Engin. Prog., 1995. Proj. No. LRE-61004, Final Report, 30 May 1995.
11. Richard Sproat, editor. *Multilingual text-to-speech synthesis*. Kluwer, Boston, 1997. Forthcoming.
12. Jan P. H. van Santen. Assignment of segmental duration in text-to-speech synthesis. *Computer Speech and Language*, 8:95–128, 1994.
13. Jan P. H. van Santen and Julia Hirschberg. Segmental effects on timing and height of pitch contours. In *Proc. ICSLP-94*, pages 719–722, Yokohama, 1994.
14. Jan P. H. van Santen and Bernd Möbius. Modeling pitch accent curves. In *ESCA Tutorial and Research Workshop on Intonation: Theory, Models and Applications*, Athens, Sept. 18–20 1997. Forthcoming.
15. Jan P. H. van Santen, Bernd Möbius, and Michael Tanenblatt. New procedures for constructing acoustic inventories. Technical report, Bell Laboratories, 1994.