Computer Speech and Language (1999) **13**, 319–358
Article No. csla.1999.0127
Available online at http://www.idealibrary.com on IDEAL

The Bell Labs German text-to-speech system

Bernd Möbius†

Institute of Natural Language Processing, University of Stuttgart, Azenbergstrasse 12, D-70174 Stuttgart, Germany

Abstract

The Bell Labs multilingual text-to-speech system can be characterized as consisting of a 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 a detailed description of the German version of the Bell Labs text-to-speech 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 design and structure of the acoustic inventory for concatenative synthesis and the criteria and procedures that were used to build it.

1. Introduction

Two defining characteristics of the Bell Labs TTS system (Sproat, 1998) are *modularity* and *multilinguality*. The architecture of the system is entirely modular. This design has a number of advantages for system development and testing, and research. First, although the division of the TTS conversion problem into subproblems is always arbitrary to some extent, each module still corresponds to a well-defined subtask in TTS conversion. Second, from the system development point of view, members of a research team can work on different modules of the system, and an improved version of a given module can be integrated anytime, as long as the communication between the modules and the structure of the information to be passed along is defined. Third, it is possible to interrupt and (re-)initiate processing anywhere in the pipeline and assess TTS information at that point, or to insert tools or programs that modify TTS parameters.

In the Bell Labs system, the linguistic text analysis component is followed by the prosodic component, which is then followed by the synthesis component. Information flow is unidirectional, and each module adds information to the data stream. Communication between the modules is performed by way of a single data structure; each module reads and writes the information relevant to it.

The multilingual design of the Bell Labs TTS system is achieved by the use of common algorithms for multiple languages. The system consists of one single set of

†Author to whom correspondence should be addressed: E-mail: moebius@ims.uni-stuttgart.de

0885–2308/99/040319 + 40 $30.00/0

language-independent software modules. The multilingual character of the TTS system can be compared to a text processing program that allows the user to edit text in a number of languages by providing language-specific fonts, whereas the same underlying principles and options concerning text formatting or output are applied disregarding the language currently being processed.

Obviously, some language-specific information is necessary; there are acoustic inventories unique to each language and there are also special rules for linguistic analysis. These data, however, are stored externally in precompiled finite-state transducers, tables, models and parameter files, and are loaded by the TTS engine at run-time. It is therefore possible to switch voices and languages as desired at run-time. This capability is particularly useful in applications such as dialog or e-mail reading, where turns, quotations and foreign-language textual material could each be rendered in a distinctive way.

In this paper we present a detailed description of the German version of the Bell Labs TTS system. Following the general flow of information through the system, we first discuss aspects of linguistic text analysis and our solutions to the problems they pose (Section 2). 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 (Section 3) and intonation (Section 4). Finally, we will explain the design and structure of the acoustic inventory for concatenative synthesis and the criteria and procedures that were used to build it (Section 5).

The paper assigns considerably more space to the presentation of the text analysis component, compared to the prosodic and synthesis components. This imbalance reflects the complexity and diversity of problems that need to be addressed in the area of linguistic text processing. An additional consideration was that the duration and intonation components have been discussed in more detail in separate publications (Möbius & van Santen, 1996 and van Santen & Möbius, 1997, respectively). As for the synthesis component, only the inventory design considerations are specific to German, whereas the signal processing modules of the TTS system are language independent and have been presented elsewhere (Sproat, 1998, Chapter 7).

2. Text analysis

We use the term “text analysis” as a cover term for all the computations involved in converting input text into an internal symbolic linguistic representation. Text analysis thus comprises such tasks as end-of-sentence detection, tokenization of sentences into words, and expansion of abbreviations and numeral expressions, as well as lexical and morphological analysis, phonological modeling, phrasing, and accenting.

By definition, TTS systems start by converting text written in the standard orthography of the language into linguistic representations. However, written language is at best an imperfect representation of linguistic structure because it is ambiguous and incomplete and lacks information that is crucial for the proper pronunciation of words.

For instance, the input text has to be segmented (“tokenized”) into sentences and words. In German, as in most European languages, the period is ambiguous in that it delimits sentences but also marks abbreviations.[1] Abbreviations and acronyms, once recognized as such, have to be either expanded into regular words or spelled letter by letter (*3 kg, U.S.A.*) (see Section 2.1.4). Unlike Chinese or Japanese, German generally uses white space to separate

[1] Some writing systems, e.g. Chinese, use a special symbol (a period) to unambiguously mark the end of a sentence.

words from each other, but at the same time allows extensive compounding, by glueing together otherwise independent lexical units (Section 2.1.5), as well as complex expressions made of letters, digits, and other symbols. The character string *42%* actually consists of two distinct words, *zweiundvierzig* and *Prozent*. Numeral expressions, often occurring in combination with abbreviations and special symbols, have to be expanded into well-formed number names (Section 2.1.3).

Performing these text normalization tasks in pre-processing steps, as it is done in conventional systems for German, with the exception of the SVOX system (Traber, 1995), often leads to incorrect analyses because sufficient context information is not available at the time the expansion is performed. Such context information may comprise lexical and morphological analysis of surrounding words in the text, part-of-speech assignment, or syntactic parse trees.

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 first two digits of *1997* should be expanded to *neunzehnhundert* "nineteen hundred" (not *eintausend neunhundert* "one thousand nine hundred"), 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, and unfortunately, the expansion would be wrong. The correct solution *zweiundzwanzigsten neunten* (dative singular masculine) can only be found if a special grammatical constraint or language model is applied that enforces number, case, and gender agreement between the numeral expression and the preceding preposition (*am*), and rules out all non-agreeing alternatives. The example illustrates that the so-called text normalization tasks can best be handled by integrating them into other aspects of linguistic analysis, such as lexical analysis, morphology, and phonology. This is the design adopted by all our multilingual TTS systems.

Analogous to the modular internal structure of the TTS system as a whole, the text analysis component consists of a multitude of smaller modules. Besides making system development and testing more efficient, the internal modularity of the text analysis component also reflects the fact that certain aspects of linguistic analysis require their own specific representation. For instance, phonological processes and pronunciation rules are best formulated in terms of context-sensitive rewrite rules. Word formation processes are most appropriately represented in the form of inflectional paradigms in the case of inflectional morphology, whereas derivational processes call for the capability to decompose morphologically complex words into their constituents.

Despite the heterogeneous character of different levels of linguistic description and despite the variety of problems encountered across languages, a homogeneous approach to implementing the text analysis is possible if a more abstract view of the problems involved is taken. Sproat (1996) has shown that each subtask in linguistic analysis can be described as a transformation of one string of symbols (viz. orthographic characters) into another string of symbols (viz. linguistic representation). A flexible and, at the same time, mathematically elegant computational model for the conversion of symbol strings is finite-state transducer (FST) technology (van Leeuwen, 1990; Roche & Schabes, 1997). Sproat developed a toolkit, Lextools (Sproat, 1998, pp. 21–28), that provides programs to convert various forms of linguistic description into a weighted finite-state transducer (WFST) representation.

Compiled finite-state transducers tend to be large. Sizes of lexical analysis WFSTs for six languages are given in Table I. The highest computational complexity of the WFST searches

TABLE I. Sizes of lexical analysis WFSTs for six languages (Sproat, 1998, p. 74)

Language	States	Arcs
Spanish	31 601	71 260
Mandarin	45 806	277 432
Japanese	93 720	415 381
German	111 190	623 643
Italian	112 659	269 487
Russian	139 592	495 847

is observed when the system explores certain areas of the network, as in the case of disambiguating lexical analyses by applying local context models and syntactic agreements (see Section 2.2). Even then the performance on a standard computer (Sproat, 1998, p. 74 refers to a 100 MHz machine) is acceptably fast for a TTS application.

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 sublexica for specific domains, such as geographical names or foreign loan words.

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.

The toolkit was built on top of a C program library[2] that performs mathematical operations on WFSTs. Lextools includes the following programs that have been used to construct the text analysis of the German TTS system:

mapbuilder: compiles simple character mappings; used for, e.g. case conversion (Section 2.4), label deletion and insertion (Sections 2.4 and 2.3).

compwl: compiles word lists and grammatical constraints if represented as regular expressions; used for, e.g. lexicon construction (Section 2.1) and various filters and disambiguators (Sections 2.1.3 and 2.2).

paradigm: compiles morphological paradigms (Section 2.1).

arclist: compiles finite-state grammars; used for, e.g. word models and name analysis (Section 2.1.5).

numbuilder: converts digit strings to number names (Section 2.1.3).

[2]The C program library was written by Michael Riley and Mehryar Mohri, now with AT&T Research.

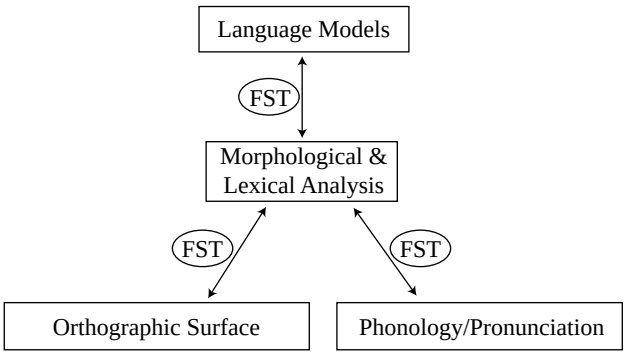

Figure 1. Text analysis component of German TTS based on weighted finite-state transducer technology.

rulecomp: compiles rewrite rules; used for, e.g. syntactic agreements (Section 2.2), phonological processes and pronunciation rules (Section 2.3).

The main modules of the text analysis components are displayed in Figure 1. First, input text is converted into a finite-state acceptor which is then composed sequentially with a set of transducers that go from the orthographic surface representation to lexical and morphological analysis. Since this yields all possible lexical analyses for a given input, a set of language models helps find the presumably correct or most appropriate analysis. The best path through the language model is then composed with a transducer that removes any labels that are not required by the phonological component. The resulting reduced annotation serves as input to phonological analysis and pronunciation rules.

As it should become obvious in the subsequent sections, the lexical and morphological analysis provides a rich linguistic annotation of the input text. At first glance this granularity might appear to be too fine for the purpose of text-to-speech. However, the pronunciation of words in German is sensitive to morphological structure. Inflected words such as nouns, adjectives and verbs have complex inflectional paradigms, and derivational and compositional processes are highly productive. Providing a detailed morphological analysis facilitates the implementation of an efficient set of pronunciation rules with a minimal number of rules that handle exceptions (see Section 2.3.1). Furthermore, retrieving the base forms of inflected or derived word forms would help to detect instances of coreference, which in turn might be used for prosodic deaccenting. Note that such a mechanism has not yet been implemented.

In the following we will describe the modules of the text analysis component in detail. Figure 2 displays the structure of the system and the dependencies between its modules, as well as the sections in which they are addressed in this paper.

2.1. Lexical and morphological analysis

Conventionally, German orthography writes nouns and names with initial capitals. This would appear to be a simple way of disambiguating strings that have multiple analyses in terms of their word class, e.g. *Regen* “rain (noun)” vs. *regen* “move (verb)”. Of course, sentence initial words are always capitalized, but as a rule, the probability of a capitalized word being a noun or name is higher than being a member of another word class. However, the TTS system should be able to handle orthographic input that slightly violates conventions, or follows new conventions such as the “all lower case” typing style often encountered in

Text analysis: Lexical/morphological analysis, Language models, Phonology
Lexical/morphological analysis: Lexicon, Numbers, Abbreviations, Alphabet, Compounds, Names; various maps from orthography to lexicon
Lexicon: nouns, adjectives, regular verbs, irregular verbs, uninflected words, special word lists (countries, cities, etc.) (Section 2.1)
Numbers: expansion of numeral expressions (Section 2.1.3)
Abbreviations: expansion of abbreviations and acronyms (Section 2.1.4)
Alphabet: spelling character by character (Section 2.1.4)
Compounds: morphological word model and phonotactic syllable model for compounded and unknown words (Section 2.1.5)
Names: morphological name models for city and street names, first names (Section 2.1.5)
Language models: Agreements, Prosodic models
Agreements: syntactic and grammatical agreements (Section 2.2.1)
Prosodic models: sentence mode and phrasing (Section 2.2.2), word accent status and syllabic stress (Section 2.2.3)
Phonology: phonological processes, pronunciation rules, syllabification; various maps from lexicon/morphology to phonology (Section 2.3)

Figure 2. Modules of the text analysis component and their dependencies.

e-mail messages. In our system, case conversion in both directions is performed by a simple character map file. The map is compiled into a WFST by means of the program mapbuilder and then composed with all lexical transducers as described in subsequent sections. Thus, the input <Regen> will be matched with the lexical entries *Regen* (noun) and *regen* (verb); in fact, any combination of upper and lower case letters will match both entries. A small penalty is assigned to each individual character conversion such that exact, case-sensitive matches between input and lexicon entry will be preferred.

Another reason to provide optional respelling are the umlauted vowels <ä, ö, ü, Ä, Ö, Ü> and the sharp-s <ß>. It is customary to replace these characters with the digraphs <ae, oe, ue, Ae, Oe, Ue, ss>, respectively, mostly for technical reasons like the 7-bit character coding of e-mail messages, but sometimes also for reasons of personal taste. A set of rules rewrites the single characters as their digraph equivalents. The transducer derived from this set of rules optionally transduces the digraph sequences into their umlauted equivalents. The lexical analysis component then decides whether or not to treat them as an umlauted vowel or as a sequence of vowels. Just as explained for case conversion, the digraph-to-umlaut conversion comes at a small cost. In those cases where actual homographs are created by such respellings the system would have to resort to homograph disambiguation techniques. But such cases are actually quite rare and mostly restricted to the pair <ß/ss>, as in *Maße* “measure (noun, pl.)” vs. *Masse* “mass (noun, sg.)”.

2.1.1. Inflected word classes

In German, there are three inflected word classes, nouns, adjectives and verbs, each of which displays a rich set of quite diverse inflectional paradigms. An explicit, if not exhaustive, grammatical and morphological annotation requires the following specifications, which will subsequently be explained and exemplified:

Nouns: word class, paradigm, gender, number, case, lexical stress, morpheme boundaries, allomorphy, origin, semantics.
Adjectives: word class, paradigm, gender, number, case, lexical stress, morpheme boundaries, allomorphy, origin, semantics.
Verbs: word class, paradigm, number, person, tense, mood, voice, lexical stress, morpheme boundaries, allomorphy, origin, prefix type, semantics.

Nouns. The sublexicon for each of the inflected word classes is separated into a word list and a paradigm file, and the entries in these files are represented as regular expressions. Figure 3 shows a segment of the word list file for nouns and the corresponding paradigm file segment. The word list for nouns comprises approximately 11 000 entries. The number of distinct inflectional paradigms for nouns is 116. Paradigms are identified by labels (e.g. N1) in the left column of the word list file.

By convention, grammatical, morphological and phonological annotations are enclosed in curly braces. Most of these labels are self-explanatory. Morpheme boundaries are marked by {++}, both in the expanded word forms, where they typically separate base forms and suffixes, and for morphologically complex entries in the word list itself. Primary and secondary lexical stress are marked by (') and (''), respectively.

The paradigm compiler expands each entry in the word list to all possible word forms according to the inflectional paradigm the word belongs to. As shown in Figure 3, the base word *Masse* belongs to paradigm N6 which states that the suffix *+n* attaches to the base form for all plural forms, disregarding the case, whereas the singular forms take on a zero suffix ({Eps} or ϵ, the empty string), i.e. their phonological substance remains unchanged. Other paradigms are more elaborate. For example, one of the two entries for *Bock* belongs to paradigm N5. Here, three out of four singular cases, nominative, dative and accusative, remain unchanged by default; there exists, however, an alternative dative form *Bock+e*. The genitive singular can be either *Bock+s* or *Bock+es*, which is expressed by the question mark after the *e*. The noun *Bock* is a homograph, its masculine version meaning "buck" and the neuter form referring to a specific type of beer. The two readings are tagged with semantic labels that may be activated once homograph disambiguation techniques (e.g. Yarowsky, 1994) are available for German. *Meter*, even though it can be either masculine or neuter, is not a homograph, because both genders are customarily used and have the same meaning ("meter"), and they also share the same inflectional paradigm.

The label {−sz} in the entry for *Abguß* "cast" indicates that the sharp-s *ß* has to be converted into the digraph <ss> in some of the expanded word forms, whereas the label {+sz} means that this transformation does not occur in the pertinent entry, e.g. in *Anstoß* "inducement". There is a correlation—with unclear causal relation—between the phonological length of the vowel preceding <ß> and the (non-) conversion of <ß> into <ss>: if the vowel is short, then <ß> in the base form turns into <ss> in all pertinent inflected word forms, and if the vowel is long, no conversion takes place. In our implementation, we exploit the label pair {+/−sz} for two purposes. First, the labels determine whether or not to convert sharp-s into a digraph. This is achieved by composing the noun paradigms with a transducer that is derived from an appropriate rewrite rule. The label {−sz} matches the context for that rule while {+sz} does not. Second, the labels are used in the pronunciation rules to decide on the vowel length, which depends on whether the vowel is followed by <ß> or <ss>.

The label {almo}, as in the entry for *Abguß*, indicates that some form of stem allomorphy or "umlaut" occurs in derived forms of the word. Foreign loan words whose pronunciations retain some properties of the phonological system of the source language, are tagged with

```
/{N1}       :               ('Ab{++}guß{noun}{masc}{almo}{−sz})/
/{N2}       :               (Acc''ess'oire{noun}{neut}{fren})/
/{N3}       :               ('Albern{++}heit{noun}{femi})/
/{N1}       :               ('An{++}stoß{noun}{masc}{almo}{+sz})/
/{N4}       :               (B'esen{noun}{masc})/
/{N4}       :               (B'ock{noun}{neut}{sen2})/
/{N5}       :               (B'ock{noun}{masc}{almo}{sen1})/
/{N6}       :               (M'asse{noun}{femi})/

Paradigm                    {N3}
Suffix      {Eps}           {sg}
Suffix      {++}en          {pl}
###
Paradigm                    {N6}
Suffix      {Eps}           {sg}
Suffix      {++}n           {pl}
###
Paradigm                    {N1}
Suffix      {Eps}           {sg}({nom}|{acc}|{dat})
Suffix      {++}es          {sg}{gen}
Suffix      {++}e           {sg}{dat}
Suffix      {++}{+front}e   {pl}({nom}|{gen}|{acc})
Suffix      {++}{+front}en  {pl}{dat}
###
Paradigm                    {N4}
Suffix      {Eps}           {sg}({nom}|{dat}|{acc})
Suffix      {++}s           {sg}{gen}
Suffix      {Eps}           {pl}
###
Paradigm                    {N2}
Suffix      {Eps}           {sg}({nom}|{dat}|{acc})
Suffix      {++}s           {sg}{gen}
Suffix      {++}s           {pl}
###
Paradigm                    {N5}
Suffix      {Eps}           {sg}({nom}|{dat}|{acc})
Suffix      {++}e?s         {sg}{gen}
Suffix      {++}e           {sg}{dat}
Suffix      {++}{+front}e   {pl}({nom}|{gen}|{acc})
Suffix      {++}{+front}en  {pl}{dat}
```

Figure 3. A segment of the word list for nouns (top panel) and pertinent inflectional paradigms.

a label indicating the language of origin. In Figure 3 *Accessoire* "accessory" is labeled as a French loan word, which can be handled by special pronunciation rules (Section 2.3.3). Currently, loan words have to be identified and tagged by hand. Small sets of special pronunciation rules have so far been implemented for English, French and Italian.

```
/{N1}       :              (F′uß{noun}{masc}{+sz})/
/{N1}       :              (Fl′uß{noun}{masc}{almo}{−sz})/

Paradigm                   {N1}
Suffix      {Eps}          {sg}({nom}|{acc}|{dat})
Suffix      {++}es         {sg}{gen}
Suffix      {++}e          {sg}{dat}
Suffix      {++}{+front}e  {pl}({nom}|{gen}|{acc})
Suffix      {++}{+front}en {pl}{dat}

allomorphy rule:
ß → ss/__({noun}|{verb}){Grammatical}*{−sz}{++}{Allomorphy}?
{Grammatical}*{OVowel};

umlaut rule:
u → ü/__{Letter}*{noun}{Grammatical}*{Allomorphy}?{++}{+front};
```

Figure 4. Lexicon entries and inflectional paradigm for *Fuß* and *Fluß*, and allomorphy and umlaut rules applying to some of their word forms.

The noun paradigm transducer is then composed with a set of rules that handle stem allomorphy and umlaut processes. We will now discuss a few concrete examples to illustrate these processes. The examples include <ß> conversion, umlaut, and irregular stem allomorphy.

Umlaut and <ß> conversion are jointly exemplified by the lexicon entries for *Fuß* "foot" and *Fluß* "river", which share the same inflectional paradigm (see Fig. 4).

The allomorphy rule states that the grapheme <ß> is rewritten as the digraph <ss> if it is followed by a noun or verb label, any number of grammatical labels, the label {−sz}, a morpheme boundary, an optional allomorphy label (such as {+front}), more grammatical labels, and a vowel grapheme. Evidently, the entry for *Fluß* matches this context specification, whereas that for *Fuß* does not. Thus, the genitive singular forms of the two words are written as *Fuß+es* and *Fluss+es*, respectively.

The umlaut rule states that the grapheme <u> has to be converted into <ü> if it is followed by any number of orthographical characters, a noun label, any number of grammatical labels, an optional allomorphy label (such as {+sz}), a morpheme boundary, and the label {+front}. This context is matched by all plural forms of both words, as specified in the inflectional paradigm. Therefore, the correct nominative plural forms of the two words are *Füß+e* and *Flüss+e*, respectively.

A number of nouns, the majority of which are of Latin or Greek origin, have irregular plural forms that require the stem to lose one or more final graphemes before the plural suffix is attached. In Figure 5 the paradigm compiler generates (incorrect) plural forms **Dinosaurus+ier* and **Dinosaurus+iern*, and the allomorphy rule deletes the stem ending *−us*, which yields the correct plurals *Dinosaur+ier* and *Dinosaur+iern*.

Adjectives. Inflectional processes of adjectives are much more regular than those for nouns in the sense that there is one common paradigm for all adjectives. Adjectives are annotated for the same list of grammatical and morphological attributes as nouns because of the requirement of grammatical agreement of any adjective-noun pair. However, the adjectival paradigm is larger than nominal ones because adjectives can occur, first, in the degrees

```
/{N133}     :          (Dino{++}s′aurus{noun}{masc})/

Paradigm               {N133}
Suffix      {Eps}      {sg}
Suffix      {++}ier    {pl}({nom}|{gen}|{acc})
Suffix      {++}iern   {pl}{dat}

 allomorphy rule:
 us → {Eps}/__{noun}{Grammatical}*{++}iern?{pl};
```

Figure 5. Lexicon entry and inflectional paradigm for *Dinosaurus*. The allomorphy rule truncates the stem in the plural forms.

of positive, comparative, and superlative, as well as in predicative function, and, second, in the context of a determinate article ("weak" inflection) or an indeterminate article ("strong" inflection).

The subparadigms for comparative and superlative degrees share the set of endings with the positive degree but they insert a degree-specific morpheme: *die klein+e Stadt – die klein+er+e Stadt – die klein+st+e Stadt* "the small (pos. – comp. – superl.) town". The difference between the weak and strong inflection can be exemplified by the following pair: *der alt+e Baum – ein alt+er Baum* "the/an old tree".

Two groups of adjectives are subject to obligatory stem allomorphy processes, and optional allomorphy applies to a third group. First, adjectives ending on unstressed *–e* (pronounced as [ə]) have to be stripped of this final vowel before the regular set of inflectional suffixes is attached: *rigide – rigid+er* "rigid". Second, in adjectives ending on unstressed *–Cel*, where *C* is any consonant, the schwa has to be deleted before inflectional suffixes are attached: *dunkel – dunkl+er* "dark" or *sensibel – sensibl+er* "sensitive". Third, in adjectives ending on unstressed *–Cer* or *–Cen*, the schwa is optionally deleted in originally German stems (*finster – finster+er/finstr+er* "dark"), mostly for stylistic or rhythmic reasons, whereas in stems of foreign origin the schwa deletion rule is obligatory (*makaber – makabr+er* "macabre").

Furthermore, the comparative and superlative degrees cause vowels in the stem of many adjectives to be umlauted: *jung – jüng+er – jüng+st+er* "young (pos. – comp. – superl.)".

The paradigm compiler expands the adjectival word list entries (approximately 10 000) to all possible inflectional forms. All stem allomorphy processes are modeled by a set of appropriate rewrite rules. The adjective paradigms are then composed with the transducer that is derived from the allomorphy and umlaut rules.

Verbs. In some analogy to English, German verbs can be categorized into two classes, regular and irregular verbs. Traditionally, verbs belonging to these classes have also been termed "weak" and "strong", respectively. The vast majority of verbs have regular inflectional paradigms, and novel verbs invariably are conjugated regularly. Irregular verbs can be enumerated because they are lexically and morphologically unproductive. German grammars typically list slightly less than 300 simplex irregular verbs. Note however that many of these verbs are among the most frequently occurring content words in both written and spoken language.

For the regular verbs, two pairs of word lists and paradigm files are needed, one for the past participle and one for all other tenses and moods. The structure of both paradigm files is parallel in that they each comprise six paradigms which in turn depend upon the phonological

structure of the verb stems. This is a different way of handling stem allomorphy compared to the treatment of nouns but, in the case of verbs, this method is more convenient. The set of inflectional suffixes is almost identical across the paradigms. The six paradigms apply to:

(1) standard verb stems (*lieb+en* "to love");
(2) verb stems ending on $-d/-t$ (*red+en* "to talk") or on $-CN$, where C is one or more consonants and N is a nasal consonant (*atm+en* "to breathe");
(3) verb stems ending on $-er$ (*lager+n* "to store");
(4) verb stems ending on $-el$ (*sammel+n* "to collect");
(5) verb stems ending on a vowel or diphthong (*freu+en* "to please"), or on $-h$ (*flieh+en* "to flee");
(6) verb stems ending on *–x/–z/–s/– ß* (*reis+en* "to travel" or *feix+en* "to smirk").

Each of the paradigms includes subparadigms for present and past tenses, both in indicative and subjunctive moods, as well as for infinitive, present participle and imperative forms.

The past participle is not only marked by a suffix (*+t*) but in the general case also by a prefix (*ge+*), e.g. *lieb+en – ge+lieb+t* "to love – loved". Exceptions are verbs that are not stressed on the first syllable, most notably foreign load words (*kredenz+en – kredenz+t* "to proffer – proffered") and latinate stems (*studier+en – studier+t* "to study – studied").

Irregular verbs are characterized by a type of stem allomorphy known as "ablaut", i.e. an alternation of the stem vowel as the most important marker for present tense, past tense, and past participle forms of irregular verbs, as in *sing+en – sang – ge+sung+en* "to sing – sang – sung".

A widely held view in the relevant literature on ablaut is that it is a vowel alternation similar in status, and closely related, to the process known as "umlaut" (see the discussion in Section 2.3.2). We agree, however, with Wiese (1996) that the internal systematics of the two phenomena are fundamentally different. Umlaut can be shown to be a lexically and morphologically conditioned phonological rule, which is why we discuss it in the Phonology Section (2.3). The place of ablaut in the grammar of German is not in phonology.

Between the almost 300 irregular verbs there are 39 distinct ablaut patterns. For a given verb the appropriate pattern is idiosyncratic and unpredictable. This means that ablauted forms need to be represented in the lexicon and cannot be derived by rule. Moreover, ablaut is entirely unproductive in German, and it can be observed that a number of strong verb forms are increasingly replaced with weak forms, e.g. *buk* → *backte* "to bake (past)". Finally, several strong verbs also display consonantal changes that are as unpredictable as the vowel alternation, e.g. *gehen – ging* "to go (pres. – past)" and *bringen – brachte* "to bring (pres. – past)".

The appropriate way to treat the ablaut vowel alternations is therefore to represent them explicitly in the lexicon. We set up six pairs of word lists and paradigm files, corresponding to present tense indicative, present tense subjunctive, past tense indicative, past tense subjunctive, past participle, and infinitive and imperative, respectively.

In compounded verbs, regular and irregular ones, the past participle prefix *ge+* is inserted between the components if the first component, i.e. not the stem, is stressed, and *ge+* is omitted if the stem is stressed. Among the examples for these two cases is a set of interesting prosodic minimal pairs, e.g. (stressed syllables are underlined):

(1) irregular verb: *um+fahr+en – um+ge+fahr+en* "to run over" vs. *um+fahr+en – um+fahr+en* "to drive around";

(2) regular verb: *über+setz+en – über+ge+setz+t* "to ferry across" vs. *über+setz+en – über+setz+t* "to translate".

A very special case is the verb *sein* "to be". Its inflectional patterns are so irregular that it cannot conveniently be represented in the paradigm format. Therefore, all possible word forms of *sein* are enumerated and compiled into a finite-state transducer.

For all irregular verbs, the most frequent compounded derivations are listed in addition to the simplex verbs, which amounts to more than 2000 entries in the irregular verb lexicon. The number of regular verbs in our system is about 7500. The paradigm compiler expands the regular and irregular verb list entries to all possible inflectional forms.

Other word lists. The lexicon of regular words includes several special word lists, most notably collections of city names (9300 entries), country names (300), and other geographical names (250). These collections are partially based on publically available data on the Web, and partially extracted from a phone and address directory of Germany on CD-ROM (D-Info, 1995).

The inflectional paradigms of these lexical entries are extremely impoverished. As a rule, geographical names have only singular inflections, and only the genitive form takes on a suffix (*Berlin – Berlin+s*). Exceptions are names whose base form is the plural to begin with, such as *die Niederlande* "the Netherlands" or *die Anden* "the Andes". However, the plural of geographical names, if necessary, can be formed by optionally attaching *+s*, e.g. *Amerika – beide Amerika/Amerikas* "America – both Americas" (namely North and South America) or *Frankfurt – beide Frankfurt/Frankfurts* (namely on the Main and on the Oder).

Obviously, other special word lists can easily be added to the lexicon, for instance lists of personal names, company names, trade marks, automobiles, etc.

2.1.2. Uninflected word classes

The uninflected word classes of German are: adverbs, determiners or articles, pronouns, prepositions, conjunctions, simplex number names (except *eins* "one", see Section 2.1.3), and interjections. With the exception of adverbs, which can be treated as special word forms of adjectives, these word classes are essentially closed lists whose members can be enumerated. They are represented in a common word list of indeclinables with about 1350 entries. The tool compwl compiles this list into a finite-state transducer.

The union of the lexical transducers for nouns, adjectives, verbs, cities, countries, other geographical names, and uninflected words is compiled into an FST that represents the lexicon of regular words.

2.1.3. Numeral expansion

Many aspects of the expansion of numerals and the conversion of strings of digits into number names in German are quite similar to English. In fact, the first phase ("factorization"), which involves expanding the numeral sequence into a representation in terms of sums of products of powers of the base 10, is identical for the two languages. English and German also share the structure of the "number lexicon", which constitutes the second phase and maps the factored representation into number names. The communality between the two languages extends to the level of providing simplex lexical items for small numbers up to and including 12 (*zwölf* "twelve"). Obviously, the lexical entries themselves are language-specific.

However, there are a few peculiarities specific to numeral expansion in German (and several other Germanic languages). Let us first consider the phenomenon often referred to as

"decade flop", i.e. the reversal of the order of decades and units in numbers between 13 and 99.[3] To give an example: the correct expansion of *21* is *einundzwanzig*. The pertinent factorization is

(3) $2 \times 10^1 + 1$.

The output of the factorization transducer is modified by a language-specific filter that performs the reversal of decades and units:

(4) $2 \times 10^1 + 1 \rightarrow 1 + 2 \times 10^1$.

The result of the filtered transducer output is *einszwanzig*.

A second process concomitant with the decade flop phenomenon is the obligatory insertion of *und* "and" between the unit and the decade. This process is implemented in the form of a rewrite rule:

(5) {Eps} → und{conju}{clit}/
(s|(ei)|(ier)|(nf)|n|(cht)) {num}{Grammatical}*___(Sigma & ! {##})* ig {num}.

The rule states that the cliticized conjunction *und* is inserted after orthographic substrings that pertain to unit digits (*eins, zwei, drei, vier, fünf, sechs, sieben, acht, neun*) and are tagged in the lexicon with the label {num} and possibly other grammatical labels ({Grammatical}), and before any sequence of symbols, with the exception of a word boundary ({##}), that ends with the substring *−ig* (indicating decades) and again is tagged with the label {num}. After application of this rule the expansion of our example *21* now reads *einsundzwanzig*.

Additional rules are required that model certain allomorphic processes in German numeral expansion. For instance, *eins* has to be rewritten as *ein* if it is followed by *hundert, tausend* or *und*; this rule completes the correct expansion of *21* into *einundzwanzig*. The number words for 10^6 *Million*, 10^9 *Milliarde*, 10^{12} *Billion*, 10^{15} *Billiarde*, etc. are feminine nouns. They require a preceding *eins* to agree in gender (*eine*). Simultaneously, these number words also have to agree in grammatical number with the preceding unit: 2×10^6 *zwei Millionen*, but 1×10^6 *eine Million*.

All these number rules are compiled into an FST. The transducer that represents the composition of the factorization and number lexicon transducers is then composed with these rules.

A rather minor difference between the writing conventions for numbers in English and German is the opposite usage of the decimal point and number comma symbols. The decimal point symbol in German is in fact a comma, and it is customary to enhance the legibility of long numbers by separating triples of digits by means of a period.

Numeral expressions in German can consist of combinations of digits, periods, commas, slashes, percent symbols, and letters. A complex numeral expression like

(6) *4.135.678,749%igem* "4 million... point seven... percent (adj., masc./neut., dat., sg.)"

would be analyzed by the system as follows:

(7) {##} v'ier {##} milli'onen {pl}{femi} {##} 'ein {masc}{sg} {##} h'undert {##} f'ünf {##} und {conju}{clit} {##} dr'ei{++}ßig {##} t'ausend {##} s'echs {##} h'undert {##} 'acht

[3] Interestingly, in its section on the formation of cardinal numbers, the Duden Grammatik (Duden, 1984) does not mention the decade flop process at all. It may not be too far-fetched to interpret this oversight as an indication that the phenomenon is prone to be taken for granted by native speakers of German, and even by German linguists. Coincidentally, it is often ignored that English also displays the decade flop property, if only for the numbers between 13 and 19, at least from the point of view of etymology.

```
/(Fr.)    :  (Fr'au{abbr}{noun}{femi}{sg})/
/(Dr.)    :  (D'oktor{abbr}{noun}{masc}{sg})/
/(St.)    :  (S'ankt{abbr}{noun})/
/(DM)     :  (M'ark{abbr}{noun}{femi}{sg})/
/(z.B.)   :  (zum{clit}{##}B'ei{++}sp''iel{abbr}{noun})/
/(BLZ)    :  (B'ank{++}leit{++}zahl{abbr}{noun}{femi}{sg})/
/(Km)     :  (K''ilo{++}m'eter{abbr}{noun}{masc})/
/(dB)     :  (D'ezi{++}b''el{abbr}{noun}{neut})/
/(Do.)    :  (D'onners{++}tag{abbr}{noun}{masc}{sg})/
/(Nov.)   :  (Nov'ember{abbr}{noun}{masc}{sg})/
```

Figure 6. Some common abbreviations and their expansions.

{##} und {conju}{clit}{##} s'ieb{++}zig {##} Komma {##} s'ieben {##}v'ier {##} n'eun {num} {##} pro{++}z'ent {noun} igem {adj}{masc}{sg}{dat} {##}

and expanded as *vier Millionen einhundertfünfunddreißigtausend sechshundertachtundsiebzig Komma sieben vier neun prozentigem.* This is achieved by concatenating the number model with a set of transducers that handle special number-related symbols (percent signs, slashes) and number-extending suffixes [−*igem* in examples (6) and (7)]. Moreover, special FSTs handle the expansion of ordinal numbers and the appropriate treatment of dates.

It is important to select the correct word form for expanded cardinal and ordinal numbers—correct in terms of agreement with the syntactic context. For example, the phrases *mit 1 Katze* "with one cat" and *mit 1 Hund* "with one dog" have to be expanded such that the word form of the number *eins* agrees with the grammatical gender of the noun (feminine in the case of *Katze*, masculine for *Hund*). The correct word form is determined by applying language model filters (see Section 2.2).

Finally, we build a finite-state machine that represents the union of the generic number transducer with the specialized FSTs for ordinals and dates.

2.1.4. Abbreviations and acronyms

For our German TTS system we currently represent 300 of the most common abbreviations and acronyms and their expansions in a word list file (Fig. 6). This list is compiled into a finite-state transducer by means of the tool compwl. To provide a perspective for the option, or the necessity, to augment this list: a special volume of the Duden book series (Duden, 1987) contains a collection of almost 20 000 abbreviations and acronyms, with twice as many expansions.

Input strings that can be neither matched to any lexical entries nor analyzed as morphologically complex words or names (see the subsequent section) are treated as acronyms. Whether an acronym is pronounced as a word or spelled out character by character cannot be reliably predicted. In fact, conventions in the language community are more important than phonotactics. For instance, the acronym *NATO* is read as a word, whereas *ADAC*, the German equivalent of the AAA, is spelled out letter by letter as [a: de: a: tse:] even though it would be possible from a phonotactic point of view to pronounce it [a:dak].

The alphabet word list file includes all regular letters of German orthography as well as various special symbols, such as <&, @, #> or arithmetic symbols. For each of these characters an orthographic representation of their pronunciation in isolation is indicated (Fig. 7).

```
/(((a|A)    :   ('a< 40.0 >{abbr}{##})) |
((b|B)      :   (b'e< 40.0 >{abbr}{##})) |
((c|C)      :   (z'e< 40.0 >{abbr}{##})) |
((y|Y)      :   ('ypsilon< 40.0 >{abbr}{##})) |
((z|Z)      :   (z'ett< 40.0 >{abbr}{##})) |
((+)        :   (plus< 40.0 >{clit}{##})) |
((−)        :   (minus< 40.0 >{clit}{##})) |
(({*})      :   (mal< 30.0 >{clit}{##})) |
(({*}{*})   :   (hoch< 30.0 >{clit}{##})) |
(({#})      :   (Nummer{clit}< 40.0 >{##}))/
```

Figure 7. Selected letters and symbols and their expansions when spoken in isolation.

Spelling out strings character by character should be the last resort for the TTS system. Therefore the costs of this transduction are higher than in any other linguistic word model.

2.1.5. Novel words

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: first, all natural languages have productive word formation processes, and the community of speakers of a language creates novel words as the need arises. Second, the set of (personal, place, brand, etc.) names is very large and, more importantly, names are subjected to similar productive and innovative processes as regular words. Therefore, *a priori* construction of exhaustive lists of words and names is impossible.

As a consequence, 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 German language is notorious for productive compounding as a means to coin neologisms. What makes this a challenge for linguistic analysis is the fact that compounding is extraordinarily productive, and speakers of German can, and in fact do, coin new compounds on the fly any time. The famous *Donaudampfschiffahrtsgesellschaftskapitän* "captain of the steam boat shipping company on the Danube river" is actually less typical than a spontaneously created word, such as *Unerfindlichkeitsunterstellung* "allegation of incomprehensibility" (for the purpose of this paper, by its author) or *Oberweserdampfschiffahrtsgesellschaftskapitänsmützenberatungsteekränzchen* "tea klatsch for the advice on yachting caps worn by captains of the steam boat shipping company on the upper Weser river" (submitted by an anonymous user of our interactive TTS web site). Therefore, linguistic analysis has to provide a mechanism to appropriately decompose compounds and, more generally, to handle unknown words.

Compound analysis. The word analysis component is based on the morphological structure of words and the phonological structure of syllables. The core of the module is a list of approximately 5000 nominal, verbal, and adjectival stems that were extracted from the

morphologically annotated lexicon files of the TTS system. To this collection we added about 250 prefixes and 220 suffixes that were found to be productive or marginally productive. We also included eight infixes (*Fugen*) which 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". A detailed description of the productivity study and the word and syllable models has been presented elsewhere (Möbius, 1998).

The module was implemented in arclist format. Figure 8 shows segments of the arclist source file for unknown word decomposition. In each line, the first and second column are labels of two states, the state of origin and the state of destination, respectively. The third column contains a string of symbols on the transitions, or *arcs*, between the two states. These strings consist of regular orthography, annotated with lexical and morphological labels including morpheme boundaries (+), symbols for primary (') and secondary (") lexical stress, and an optional cost for the transition.

The two most important aspects of this linguistic description 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 arc between states.

The transition from PREFIX to ROOT that is labeled "SyllableModel" in Figure 8 is a place holder for a phonetic syllable model which reflects the phonotactics and the segmental structure of syllables in German, or rather their correlates on the orthographic surface. This allows the module to analyze substrings of words that are unaccounted for by the explicitly listed stems and affixes in arbitrary locations in a morphologically complex word. Applying the syllable model is expensive because we want to cover the orthographic string with as many known components as possible. The costs actually vary depending upon the number of syllables in the residual string and the number of graphemes in each syllable. For the sake of simplicity we assign a flat cost of 10.0 in our example.

On termination the machine labels the word as a noun by default. In a more sophisticated word model it may be possible to assign the part-of-speech category of the novel word on the basis of the types of stems and affixes involved, by distinguishing between noun, verb, and adjective-forming affixes. However, as of now we are lacking the capability to disambiguate concurrent analyses, which are very likely to occur because many stems and affixes are ambiguous in terms of their part-of-speech status. Thus, in the current implementation it is sufficient for the prosodic components of the TTS system to know that the word is a content word, which is a safe assumption for novel words. By default, content words receive a word accent status of "accented" (see Section 2.2.3).

Most arc labels are weighted by a cost. Weights in the unknown word analysis module are for the most part based on linguistic intuition. In the case of affixes they also reflect some results of the productivity study (Möbius, 1998). The weights are assigned such that direct matches of input strings to entries in the lexicon will be less expensive than unknown word analysis. There is no legal path through the unknown word graph that comes at zero cost. The minimal cost of 1.0 would be for a simplex stem that is explicitly listed and does not have any affixes.

The declarative grammar for unknown word analysis is compiled into a finite-state transducer. Unfortunately, this transducer is far too complex to be usefully diagrammed. For the sake of exemplification, Figure 9 shows a graphical representation of the transducer corresponding to a small grammar that analyzes and decomposes the morphologically complex

```
START     PREFIX    {Eps}
START     PREFIX    ”a {pref}+<0.3>
START     PREFIX    ’ab {pref}+<0.3>
START     PREFIX    ”aber {pref}+<0.3>
START     PREFIX    er {pref}+<0.3>
START     PREFIX    ”un {pref}+<0.3>
START     PREFIX    ”unter {pref}+<0.3>
START     PREFIX    zw”ei {pref}+<0.3>
START     PREFIX    zw”ischen {pref}+<0.3>
          ...
PREFIX    START     {Eps}<1.0>
PREFIX    ROOT      SyllableModel {root}<10.0>
PREFIX    ROOT      ’aal {root}
PREFIX    ROOT      ’aar {root}
PREFIX    ROOT      f’ind {root}
PREFIX    ROOT      st’ell {root}
PREFIX    ROOT      zw’eig {root}
PREFIX    ROOT      z’yste {root}
          ...
ROOT      PREFIX    {comp}+<0.1>
ROOT      SUFFIX    {Eps}<0.2>
ROOT      SUFFIX    +”abel {suff}<0.2>
ROOT      SUFFIX    +ab”il {suff}<0.2>
ROOT      SUFFIX    +lich {suff}<0.2>
ROOT      SUFFIX    +keit {suff}<0.2>
ROOT      SUFFIX    +ung {suff}<0.2>
ROOT      SUFFIX    +zig {suff}<0.2>
ROOT      END       {noun}<1.0>
          ...
SUFFIX    ROOT      {Eps}<0.2>
SUFFIX    END       {noun}<1.0>
SUFFIX    FUGE      +n<0.2>
SUFFIX    FUGE      +s<0.2>
SUFFIX    FUGE      +en<0.2>
          ...
FUGE      START     {comp}+<0.5>
END
```

Figure 8. Segments of a declarative arclist grammar for unknown word decomposition in German. Column 1: state of origin; column 2: state of destination; column 3: string on arc between states, with optional cost.

word *Unerfindlichkeitsunterstellung* (“allegation of incomprehensibility”); this compound is novel in the sense that it is unknown to the system.

Name analysis. The approach to unknown word decomposition described above has also been applied to the analysis of names in our German TTS system. Arguably, names, and in particular proper names, are not equally amenable to morphological processes, such as word

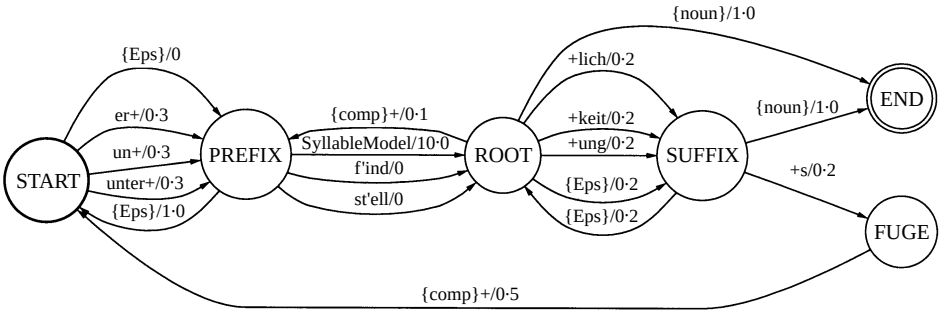

Figure 9. The transducer compiled from a subgrammar that decomposes the morphologically complex word *Unerfindlichkeitsunterstellung*.

formation and derivation, or to morphological decomposition, as regular words are. That does not render such an approach unfeasible, though, as was shown in a recent evaluation of the system's performance on street names. Street names are an interesting category because they encompass aspects of geographical and personal names. In our study (Jannedy & Möbius, 1997), we reported a pronunciation error rate *by word* of 11–13% for unknown names. In other words, roughly one out of eight names is pronounced incorrectly.

This performance compares rather favorably with results reported in the literature, for instance from the German branch of the European Onomastica project (Onomastica, 1995). Onomastica was funded by the European Community from 1993 to 1995 and aimed to produce pronunciation dictionaries of proper names and place names in 11 languages. The final report describes the performance of grapheme-to-phoneme rule sets developed for each language. For German, the accuracy rate by word for quality band III—names that were transcribed by rule only—was 71%, yielding an error rate of 29%. The grapheme-to-phoneme conversion rules in Onomastica were written by experts, based on tens of thousands of the most frequent names that were manually transcribed by an expert phonetician. In our TTS system, the phonological or pronunciation rules capitalize on the extensive morphological information provided by annotated lexica *and* the unknown word analysis component.

One obvious area for improvement is to add a name-specific set of pronunciation rules to the general-purpose one. Using this approach, Belhoula (1993) reports error rates of 4.3% for German place names and 10% for last names. These results are obtained in recall tests on a manually transcribed training corpus. 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 (Belhoula, 1993), it presupposes an unrealistically reliable detection of names in arbitrary text (see Thielen, 1995 for an approach to German name tagging). Proper name pronunciation remains one of the most problematic tasks for TTS systems.

2.1.6. Summary: from text to lexical/morphological analysis

In the preceding sections, linguistic models for various types of textual "words" were presented: nouns, adjectives, verbs, uninflected words, and specialized sublexica; abbreviations and acronyms; the alphabet; complex numeral expressions; unknown and morphologically complex words and names.

What is needed in addition to the analysis of lexical items is a model of the material that

can occur between any two such items: an inter-word model. Punctuation marks are currently used to determine sentence mode (e.g. <.>= declarative, <?> = interrogative) and phrase boundaries (e.g. <,> = minor phrase boundary). Colons and semicolons are treated as end-of-sentence markers. Other symbols, such as parentheses, brackets and quotes, are deleted. Note that no assumption is made for these latter symbols to be either preceded or followed by white space.

The union of the lexical models enumerated above is concatenated with the inter-word model transducer. The Kleene closure is then formed to allow for the analysis of sequences of words and inter-word material, i.e. sentences. The resulting WFST represents the lexical and morphological analysis of orthographic input sentences.

In the following section we will discuss the disambiguation of multiple possible analyses by means of local context models as well as the assignment of word accent and syllabic stress.

2.2. Language models and prosodic models

Under the generic term "language models" we subsume cross-word local context models, mostly in the form of syntactic agreement constraints. These models perform the disambiguation of multiple lexical and morphological analyses with equal cost for a given word by analyzing the local context. We also include here a collection of prosodic models, in particular rules that aim at determining sentence mode, syntactic and prosodic phrasing, accenting, and syllabic stress.

2.2.1. Agreement constraints

In the current implementation, most cross-word language models deal with the proper agreement of number words with their context. Certain aspects of agreement rules for numbers have already been discussed in Section 2.1.3, but there we were only concerned with agreement within numeral expressions. The following agreement rules have been implemented; note that in most cases, the appropriate word form selected by enforcing the agreement differs from the default word form in terms of its segmental material, and therefore in its pronunciation.

- Gender and case agreement of numeral expressions ending on *1* according to the gender of the content word (noun, adjective) to the right; e.g.: *mit 1 Katze* "with one cat" → *mit*{*dat*} *einer*{*num*}{*dat*}{*sg*} *Katze*{*noun*}{*dat*}{*sg*}.
- Case and number agreement of ordinal numbers according to the case and number of the word to the left; e.g.: *am 13. Januar* "on January 13" → *am*{*prep*}{*dat*}{*sg*} *dreizehnten*{*num*}{*dat*}{*sg*} *Januar*{*noun*}{*dat*}{*sg*}.
- Number agreement of abbreviations according to the number of the word to the left; e.g.: *1 t* → {*num*}{*sg*} *Tonne*{*abbr*}{*noun*}{*sg*}, as opposed to *10 t* → *zehn*{*num*}{*pl*} *Tonnen*{*abbr*}{*noun*}{*pl*}.
- Interpretation of special symbols; e.g.: <–> is read as *minus* if surrounded, preceded, or followed by numbers, and as hyphen or dash otherwise.

These agreements are formulated as rewrite rules, which are compiled into a series of finite-state transducers. The rules tag non-agreeing contexts with the label {wrong}. The transducers are then composed with a filter that disallows all sequences containing the label {wrong}. Obviously, many more cross-word context models might be useful, especially

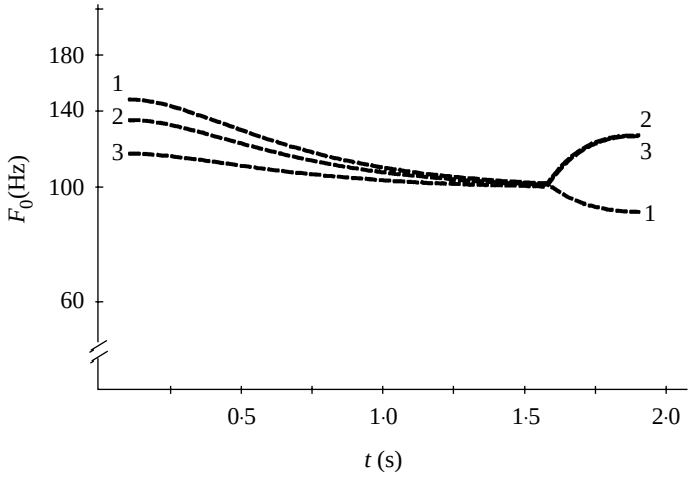

Figure 10. Stylized intonation contours for the interrogative sentence modes wh-question (1), yes/no question (2), and echo question (3) in German.

in the domain of syntactic agreement; but in fact the most appropriate solution would be to use a syntactic parser in conjunction with a part-of-speech tagger for this purpose.

2.2.2. Sentence mode and phrasing

For lack of a more sophisticated model, syntactic and prosodic phrases are currently assumed to be co-extensive. Phrase boundaries are derived from punctuation. While this is admittedly the most coarse and straightforward approach, and also a frequently inappropriate one, the situation is helped by the fact that German punctuation is rather explicit: subordinate clauses, for instance, are always separated from the superordinate clause, and from each other, by commas. As explained in Section 2.1.6, commas are interpreted as minor phrase boundary markers.

Sentence mode is determined based on the type of end-of-sentence punctuation. Obviously, a question mark triggers the sentence mode to be set to "interrogative". There are, however, at least three intonationally distinct interrogative sentence types in German (Möbius, 1993), as illustrated in Figure 10. Yes/no questions and echo questions have a steep utterance-final rise, while echo questions display an almost flat declination. Wh-question contours are remarkably similar to those of declaratives, with a clear declination and an utterance-final fall. This is possible because a wh-question is marked by both syntactic structure and, most importantly, by the sentence-initial wh-word, such as *wo, wer, warum, wann* "where, who, why, when". In the case of an echo question, whose syntactic structure is identical to that of a declarative sentence (e.g. *Du hast das Auto schon verkauft?* "You have already sold the car?"), the functional load of marking sentence mode is exclusively on the intonation contour.

Yes/no and echo questions are currently not distinguished by our TTS system. Assigning a declarative intonation contour to a wh-question is achieved by changing the sentence mode marker from interrogative to declarative by means of a simple rule (8):

(8) {??} → {..}/{whword} ({Sigma} & !{Boundary})* ___;

Both a part-of-speech tagger and a syntactic parser are ultimately required to provide the prosodic components of the TTS system with sufficient and reliable information to generate

natural-sounding prosody. These functionalities have not yet been implemented in our TTS system.

2.2.3. Word accent and syllabic stress

The text analysis module assigns to each word in the sentence a word accent status. Possible values are "accented", "deaccented", and "cliticized". Accent status is exploited by the prosodic components for segmental duration and intonation. For example, the intonation module generates pitch accents only for accented words, and the duration module distinguishes between all three accent statuses. Nouns, adjectives, verbs, names, number words, and abbreviations are treated as accented words. Auxiliary verbs and wh-words are deaccented, and articles, pronouns, conjunctions, and short prepositions are cliticized. The pertinent rewrite rules are compiled into a finite-state machine.

In general, syllabic stress is marked in the lexicon. In the case of morphologically complex words, such as unknown compounds and names, the procedure of determining the stressed syllables is as follows. Analogous to the regular lexicon, the explicitly listed stems in the word and name models are tagged for syllabic stress. However, stress shift can be caused by certain affixes that attract primary stress. To determine the stressed syllable of a derived or compounded word, a small set of rewrite rules is applied. First, the stress mark of stems is removed if they are preceded by a stress-attracting prefix (9). If there is more than one stressed prefix, only the first of them retains the stress (10). Second, any stress-attracting suffixes override the prefix rules (11). In the case of multiple suffixes, only the last stressed suffix keeps the stress mark (12).

(9) *magn'et*{*root*} — *el'ektro*{*pref*}+*magnet*{*root*}
(10) *m'ikro*{*pref*}+*elektro*{*pref*}+*magnet*{*root*}
(11) *mikro*{*pref*}+*elektro*{*pref*}+*magnet*{*root*}+*osk'op*{*suff*}
(12) *mikro*{*pref*}+*elektro*{*pref*}+*magnet*{*root*}+*oskop*{*suff*}+*'ie*{*suff*}

The scope of the rules is confined to within one major component of a compound. Arguably, even within such a component distinctions between primary and secondary stress could be made, especially if the affixes are polysyllabic as in our example. This feature has not been implemented yet. The stress assignment rules are compiled into an FST.

2.3. Phonology and pronunciation

The output of lexical analysis, filtered by the language models, serves as input to the phonology component. In this section we will first discuss several issues concerning the design of the phonological component and then present the pronunciation rule system.

2.3.1. Design issues

In the early stages of the German TTS system we temporarily used a preliminary front end that had been developed at the University of Bonn, Germany. The software consisted of a pronunciation dictionary, pronunciation rules, and a rule interpreter. The number of pronunciation rules was close to 1200. About 80% of those rules were concerned with handling exception strings, which resulted from the fact that no morphological information was available.

In our current system, the number of general-purpose pronunciation rules is 259. The main reason for this drastically reduced number of rules is that the phonological component now

draws upon the rich morphological information provided by the annotation of the lexicon and by the analysis of unknown words. The number of rules handling exceptions is small, at least as far as regular words of German are concerned.

Archigraphemes. Many words have idiosyncratic pronunciations that cannot be derived by general-purpose pronunciation rules. Conventionally, for example in the Bonn system described above, these cases are often treated as exceptions and handled by very specific rules. We argue that the proper level of description and modeling of idiosyncratic pronunciations is not in the phonology component but in the lexicon.

For instance, one general rule for determining vowel length is that a vowel is short if it is followed by the character string <sch>, e.g. in the nouns *Tusche, Muschel* [t'ʊʃə, m'ʊʃəl] "ink, shell". However, <u> is long in the noun *Dusche* [d'u:ʃə] "shower". Instead of writing a pronunciation rule that applies to just this one word and some derivations, we indicate the idiosyncratic pronunciation of *Dusche* in the lexicon by means of a special archigrapheme symbol:

(13) /{N6} : (D'{˜U}sche{noun}{femi})/

Archigraphemes can be viewed as abstract representations of graphemic symbols. An archigrapheme is not actually used in orthography, but it has a defined mapping onto one or more orthographic characters and on exactly one allophone symbol.

Intermediate representation. In word pronunciation systems that rely on large sets of ordered rules, interactions between rules can be complex and unpredictable. This is of particular concern when the graphemic and allophonic symbol sets overlap to a significant extent. It can be shown that, in our system, there is no way of ordering the pronunciation rules in such a way that no unwanted side effects occur. The solution is to define an additional symbol set that does not overlap with either graphemes or allophones. The rules convert graphemic symbols into this intermediate representation, and after all pronunciation rules have been applied, the intermediate representation is mapped on the allophonic set.

In example (14), a context-free rule rewrites the grapheme <ß> as the allophone [s]. Several subsequent rules operate on the symbol <s> and output different allophones ([s, ʃ, z]) depending on the context specifications. Thus, the output of the first rule will be further modified, which is of course not what is intended. This unwanted recursive operation can be prevented by converting any left-hand symbols into intermediate symbols, e.g. {ss}, {zz} or {SS}, which are not affected by any subsequent rules operating on graphemic symbols (15).

(14) ß → s;
s → z/({Vowel}|n|l|m|r) {++}___{&}l {Grammatical}*({##}|{++}|{Boundary});
s → s/{++}___{&}l {Grammatical}*({##}|{++}|{Boundary});
s → S/({##}|{++}|{Boundary}) {Accent}?___(t|p)(({Stress}?{Vowel})|r|l);

(15) ß → {ss};
s → {zz}/({Vowel}|n|l|m|r) {++}___{&}l {Grammatical}*({##}|{++}|{Boundary});
s → {ss}/{++}___{&}l {Grammatical}*({##}|{++}|{Boundary});
s → {SS}/({##}|{++}|{Boundary}) {Accent}?___(t|p) (({Stress}?{Vowel})|r|l);

2.3.2. Umlaut

Umlaut is a vowel alternation in German that occurs in a variety of diverse morphological conditions. In pairs of morphologically related words, non-front vowels in the base words

TABLE II. Umlaut vowel alternations in German

Vowel alternation	Base word	Derived word
/u:/ → /y:/	*Zug* "train (noun, sg.)"	*Züg+e* "train (noun, pl.)"
	Ruhm "glory (noun)"	*rühm+en* "praise (verb)"
/U/ → /Y/	*dumm* "silly (adj.)"	*dümm+lich* "silly (adj., grad.)"
	Bund "bond (noun, masc.)"	*Bünd+nis* "alliance (noun, neut.)"
/o:/ → /ø:/	*grob* "coarse (adj., pos.)"	*gröb+st* "coarse (adj., superl.)"
	Ton "tone (noun)"	*Tön+chen* "tone (noun, dimin.)"
/ɔ/ → /œ/	*Tochter* "daughter (noun, sg.)"	*Töchter* "daughter (noun, pl.)"
	Korn "grain (noun)"	*körn+ig* "granular (adj.)"
/a:/ → /ɛ:/	*Europa* "Europe (noun)"	*europä+isch* "Europe (adj.)"
	kam "came (verb, ind.)"	*käm+e* "came (verb, subj.)"
/a/ → /ɛ/	*Sorg+falt* "care (noun)"	[*sorg+fält*]+*ig* "careful (adj.)"
	tanz+en "dance (verb)"	*tänz+el+n* "dance (verb, dimin.)"
/aU/ → /ɔY/	*sauf+e* "drink (verb, 1per.)"	*säuf+t* "drinks (verb, 3per.)"
	lauf+en "run (verb)"	*Läuf+er* "runner (noun)"

alternate with their front counterparts in the derived word forms. Examples for all the vowels affected are given in Table II. Umlaut occurs on the final full (i.e. non-schwa) vowel of the stem. It is independent of syllabic stress, and it is not confined to the Germanic stock in the lexicon of German. Note that the non-alternating umlauts in base words, such as *Käse* "cheese", *schön* "beautiful", or *Tür* "door", are treated as constant lexical forms and will therefore not be discussed here.

In the relevant literature there are two competing interpretations of the umlaut phenomenon. One school of thought, represented by Lieber (1987), starts from a classification of contexts for umlauting. More specifically, certain derivational suffixes are considered as regularly umlaut-conditioning, while others are called umlaut-variable because sometimes they cause umlauting and sometimes they do not. Furthermore, certain inflectional categories (noun plurals, comparative and superlative adjectival forms, subjunctive verb tense) trigger umlauting. In the framework of autosegmental phonology, Lieber accounts for these processes by defining a floating feature {−back} that is associated with the whole suffix, not with a particular segment of the suffix, and is therefore placed on a tier of its own ("autosegmentalized"). The floating feature then docks onto a segment of the stem that has an open slot for this feature. For a more detailed account of the phonological processes involved (delinking and automatic reassociation) see Lieber (1987) or Wiese (1996).

There are several problems with this approach. First, umlaut-variable suffixes have to be represented by two allomorphs, one with and the other without the floating {−back} feature. In addition, for each suffix allomorph a list of stems has to be provided that it attaches to. For example, the adjective forming suffix *+lich* derives *ärzt+lich* "medical" from *Arzt* "doctor", but *amt+lich* "official" from *Amt* "office", even though umlauted forms of the stem *Amt* exist in other contexts (e.g. *Ämt+er* "office (pl.)". Second, many stems also have to be represented by two allomorphs, one umlauted and the other plain, in order to account for inflectional variations, such as *Bach+es* "brook (gen., sg.)" vs. *Bäch+e* "brook (pl.)", both from the stem *Bach*, or as in the previous *Amt* example. Third, the number of potentially umlaut-conditioning suffixes is small. In fact, on close inspection the only unambiguously

umlaut-conditioning suffix is the diminutive *+lein*, for which no non-umlauted noun base is known (Wiese, 1996).

The fact that the vast majority of potentially umlaut-conditioning suffixes are umlaut variable implies that a massive amount of allomorphy for both stems and suffixes has to be acknowledged. Besides being challengeable on theoretical grounds, this approach also appears to be highly impractical from an implementational point of view.

To summarize the most important aspects of umlauting:

- umlaut is a morphologically (e.g. umlaut is possible with plural suffix *+er*: *Männ+er* "men", but impossible with plural suffix *+en*: *Frau+en* "women") and lexically (e.g. *Fahr+er* "driver" vs. *Bäck+er* "baker") conditioned phonological rule.
- umlaut applies across morphological structure (*mut+ig* "courageous" vs. *[groß+müt]+ig* "generous"), but not across overt suffixes (*Fahr+er* "driver" and **[Fähr+er]+in* "driver (fem.)"); i.e. "umlaut is sensitive not to abstract morphological structure but to the phonological material instantiating the structure." (Wiese, 1996, p.124).
- umlaut can be triggered by almost any morphological distinction in the German language.
- the analysis of umlaut involves the whole vowel system and most aspects of morphology.
- depending on the stem involved, the same morphological operation may occur with or without umlauting.
- umlaut is a property of the stem and not the suffix.

In our TTS system, we implemented an alternative account of the umlauting process, which by and large corresponds to the analysis proposed by Wiese (1996) but was independently arrived at.

We adopt Lieber's suggestion of using a floating feature, but we associate it with the stems, not with suffixes. Also, we opt for a feature that is independently motivated by selecting it from the set of distinctive features of the German vowel system. We select the feature {+front} because the vowel alternation under discussion is one of fronting. In our implementation, we therefore associate the floating feature {+front} with the pertinent stems, i.e. we tag every potentially umlauted stem in the lexicon with the label {+front}, but we do so only for those forms in the inflectional paradigm of the stem that actually undergo the vowel alternation (see the example in Fig. 4). This implementational design reflects the finding that, as stated above, umlaut is both lexically and morphologically conditioned. The umlauting rule (see Fig. 4) is then applied in the appropriate inflectional contexts.

2.3.3. Pronunciation rules

The phonological component is implemented in the form of a collection of ordered phonological rewrite rules. The rules are compiled into FSTs by means of the tool rulecomp, which is based on an efficient rule compilation algorithm (Mohri & Sproat, 1996). The pronunciation rules can be grouped into four categories: (1) rules for handling affixes as well as substrings from foreign loan words; (2) rules for the pronunciation of vowels; (3) rules for the pronunciation of consonants; (4) some cross-word pronunciation rules.

Affixes and other substrings. The pronunciation of several productive prefixes and suffixes deviates from the default pronunciation of the substring. For instance, the sequence <ab> is generally pronounced [a:b], as in *aber* "but", but as a prefix it is pronounced [ap].

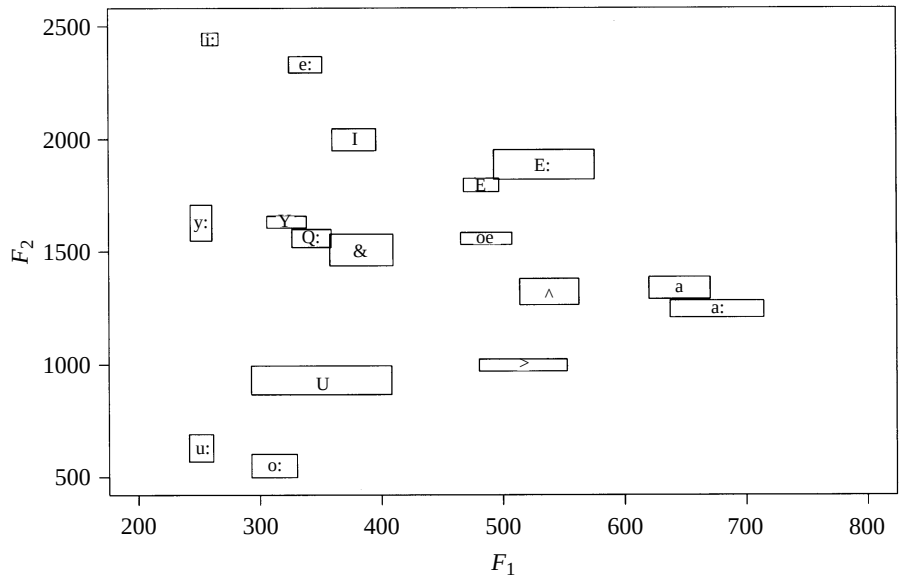

Figure 11. Vowel space of our German speaker. Vowel symbols mark median values (in Hz) in F_1/F_2 space, boxes delimit standard deviations. Non-IPA symbols map on IPA symbols as follows: & =ə, ʌ= ɐ, E = ɛ, Q: = ø:, oe = œ, >=ɔ.

Such cases are best handled by a prefix-specific rule. Another set of rules handles substrings of words that are tagged as foreign loan words in the lexicon. Currently, such pronunciation rules exist for loan words from English, French and Italian. Obviously, some mapping of the phonological system of the original language onto the phoneme inventory of German is necessary.

Vowels. German has a rich and densely populated vowel space (Fig. 11). It is customary to group the monophthongs into pairs whose members differ in terms of both phonological quantity and phonetic quality, e.g. /u:, ʊ or /y:, ʏ. Therefore, for the sake of elegance and economy of phonological analysis, some researchers prefer a count of eight monophthong phonemes /i y e ø ɛ u o a/, plus a quantity phoneme /:/, which distinguishes the long from the short allophone of a given phoneme.

However, from the point of view of speech synthesis, spectral quality is the key criterion here. The vowel [I] is not simply the short allophone of the phoneme /i/ ([i:] being the long variant). Its typical formant values differ quite significantly from [i:], and its location in the F_1/F_2 space is rather distinct from [i:] (see Fig. 11). The count of distinct vowels in German TTS is therefore 20, which includes the central vowels [ə] and [ɐ] as well as the diphthongs [aɪ], [aʊ], and [ɔɪ].

A total of 130 rules is required to determine vowel length and quality. For instance, four rules are needed for the grapheme sequence <ie> (16). The rules state that (a) <ie> is pronounced [ɪ] in a few special cases, such as *vierzig*, *vielleicht* [f'ɪɐtsɪç, fɪl'aɪçt] "forty, perhaps"; (b) <ie> becomes [ɪə] if it is unstressed and is followed by either the grapheme <n> or the allophone [n] or a morpheme, word or phrase boundary, e.g. in *Albanien* [alb'a:nɪən] "Albania"; (c) <ie> becomes [i:ə] in a stressed syllable if it is followed by a morpheme boundary and either the grapheme <n> or the allophone [n] (both representing the plural morpheme),

followed by optional grammatical labels and a morpheme, word or phrase boundary, e.g. in *Kolonie+n* [kolon'i:ən] "colonies"; (d) <ie> is pronounced [i:] in all other cases.

(16) ie → {iE}/(v|{ff})___(r(t|z)|(ll));
ie → {iE}{&}/(!{Stress} & {Sigma})___(n|{nn}|({##}|{++}|{Boundary}));
ie → {EE}{&}/{Stress}___{++}(n|{nn}){Grammatical}*({##}|{++}|{Boundary});
ie → {EE};

Consonants. Another set of 129 rules is required to determine the pronunciation of consonants. Almost one third of these rules are used to model the neutralized voicing opposition in morpheme- and word-final position, a phenomenon known as *Auslautverhärtung*: phonologically voiced obstruents and clusters of obstruents in this position turn into their voiceless counterparts. However, this neutralization does not apply in certain contexts (17). The rules, here simplified for the sake of exemplification, state that the grapheme <d> is pronounced [d] if it is followed by a morpheme boundary and various derivational suffixes, as in *Rund+ung* [r'ʊndʊŋ] "rounding (noun)", but it is pronounced [t] if followed by a morpheme, word or phrase boundary in other contexts, as in *rund* [r'ʊnt] "round (adj.)".

(17) d → {dd}/___{++}(l|r|n)? ((er?)|(ung)|(ig)(ich)|(isch));
d → {tt}/___({++}|{##}|{Boundary});

Cross-word rules. A set of rules prevents geminate allophones across morpheme or word boundaries by deleting the first of two identical allophone symbols.

2.4. Final steps

The output of the lexical analysis, filtered by the language models, is sequentially composed with the transducers that are generated from the phonological rules. First, however, all upper case grapheme symbols are mapped onto their lower case correlates. Then all archigraphemes coming from the lexicon are converted into the intermediate phonemic representation. Next, all rules dealing with foreign loan words are applied. At this point, any domain-specific rules, e.g. rules for name pronunciation, could also be applied if implemented. After that we delete all grammatical labels.

After applying the pronunciation rules, a few clean-up steps remain to be taken. First of all, we delete all stress symbols from cliticized words as well as remaining secondary stress symbols. Second, we map the intermediate phonemic representation onto the set of allophone symbols that is actually used in the TTS system. Third, we remove all remaining grammatical and morphological annotations, including boundary symbols. The result of these operations is a representation of the pronunciation of words in the form expected by other components of the TTS system, i.e. just sequences of allophones and primary stress symbols.

Information resulting from various stages of linguistic analysis is written into the TTS structures by the software module gentex, which represents the generalized text analysis component in the multilingual TTS pipeline, at run-time. Those structures are accessible to the other components of the TTS system, especially duration and intonation.

3. Segmental duration

In natural speech, segmental duration is strongly context dependent. For instance, in our German speech database we observed instantiations of the vowel [ɛ] that were as short as

35 ms in the word *jetzt* [jɛtst] “now” and as long as 252 ms in the word *Herren* [hɛrən] “gentlemen”. Among the most important factors in many languages are the position of the word in the utterance, the accent status of the word, syllabic stress, and the segmental context. These factors and the levels on them jointly define a large feature space. The task of the duration component of a TTS system is to reliably predict the duration of every speech sound depending on its feature vector. An additional requirement is that the feature vector be computable from text.

The prevalent type of duration model is a sequential rule system as proposed by Klatt (1973, 1976). Starting from some intrinsic value, the duration of a segment is modified by successively applied rules. Models of this type have been developed for several languages including American English (Olive & Liberman, 1985; Allen *et al.*, 1987), Swedish (Carlson & Granström, 1986), German (Kohler, 1988), French (Bartkova & Sorin, 1987), and Brazilian Portuguese (Simoes, 1990). When large speech databases and the computational means for analyzing these corpora became available, new approaches were proposed based on, for example, Classification and Regression Trees (CART) (Pitrelli & Zue, 1989; Riley, 1992) and neural networks (Campbell, 1992). It has been shown, however, that even huge amounts of training data cannot exhaustively cover all possible feature vectors (van Santen, 1994). An alternative method, manual database construction, is only feasible if the factorial space is not too large. However, in the duration analysis for our English TTS system a minimum of 17 500 distinct feature vectors were observed (van Santen, 1993*c*). Since the factorial scheme for German bears some resemblance to the one for English, the number of distinct feature vectors can be assumed to be in the same order of magnitude, making a manual database construction impractical. Rare vectors cannot simply be ignored because the combined frequency of rare vectors almost guarantees the occurrence of at least one unseen vector in any given sentence. In an analysis for English, van Santen (reported in Sproat, 1998, p. 128) found a probability of more than 95% that a randomly selected 50-phoneme sentence contains a vector that occurs in less than one in a million segments. Such quantitative data is not available for German, but there is little reason to assume that a similar study for German would yield significantly different results.

Thus, the duration model has to be capable of predicting—by some form of extrapolation from observed feature vectors—durations for vectors insufficiently represented in the training material. CART-based methods are known for poorly coping with data sparsity, because they lack this extrapolation capability. Extrapolation is complicated by interactions between the factors. Factor interactions prevent simple additive regression models (Kaiki *et al.*, 1990), which have good extrapolation properties, from being an efficient solution. This assertion holds even though the interactions are often regular in the sense that the effects of one factor do not reverse the effect of another factor.

The solution proposed by van Santen (1992, 1993*a*, 1994) is the application of a broad class of arithmetic models, *sums-of-products models*. This approach takes advantage of the fact that most interactions 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. This method has been shown to be superior to CART-based approaches, for several reasons (Maghbouleh, 1996). First, it needs far fewer training data to reach asymptotic performance. Second, this asymptotic performance is better than for CART. Third, the difference in performance grows with the discrepancy between training and test data. Fourth, adding more training data does not improve the performance of CART-based approaches. Van Santen’s method has been applied to American English (van

Santen, 1993*c*, 1994) and Mandarin Chinese (Shih & Ao, 1997), and we used it in the present study.

Following this general method, we constructed a model for segmental duration in German (Möbius & van Santen, 1996). 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.

3.1. Speech database

Our analysis of segmental durations in natural speech is based on the Kiel Corpus of Read Speech, recorded and manually segmented at the Kiel phonetics institute and published on CD-ROM (Kiel Corpus, 1994). The disk contains speech and label files; the latter provide: orthography; canonical transcription according to standard German pronunciation; transcription of actually realized speech (see Kohler, 1994 for details). Two speakers, one female and one male, produced the entire text material. We selected the renditions of the male speaker "k61", given that the voice of our TTS system is male too. Some sanity and consistency checks were performed on the labeled data; for instance, all utterance-initial stop closure data were excluded from the analysis. The database ultimately yielded a total of 23 490 segments: 6991 vowels and 16 499 consonants. We computed feature vectors for all the segments in the database. The following factors were included in the annotation:

- segment identity
- segment type; levels: front, mid, and back vowels, voiced and unvoiced stops and fricatives, nasals, liquids, glides, silence
- word class; levels: function word, content word, compound
- position of phrase in utterance
- phrase length (in number of words)
- position of word in phrase; levels: initial, medial, final
- word length (in number of syllables)
- position of syllable in word; levels: initial, medial, final
- stress; levels: primary, secondary, unstressed
- segment position in syllable; levels: onset, nucleus, coda
- segmental context; levels: identities of first, second, and third segments to left and right
- segment type context; levels: type of first, second, and third segments to left and right
- boundary type; levels: phrase, word, syllable, no boundary to left and right
- context segment cluster; levels: (e.g.) voiceless obstruents in coda, empty onset, diphthong nucleus, etc. to left and right

It is important to note that this database was not optimal for the purpose of duration system construction, because no attempt was made to cover the greatest number of distinct feature vectors. By contrast, in their study of Mandarin Chinese duration, Shih and Ao (1997) used greedy methods (van Santen, 1993*b*) to select a few hundred sentences that covered the same set of feature vector *types* as the much larger set of 15 000 sentences from which the sentences were drawn.

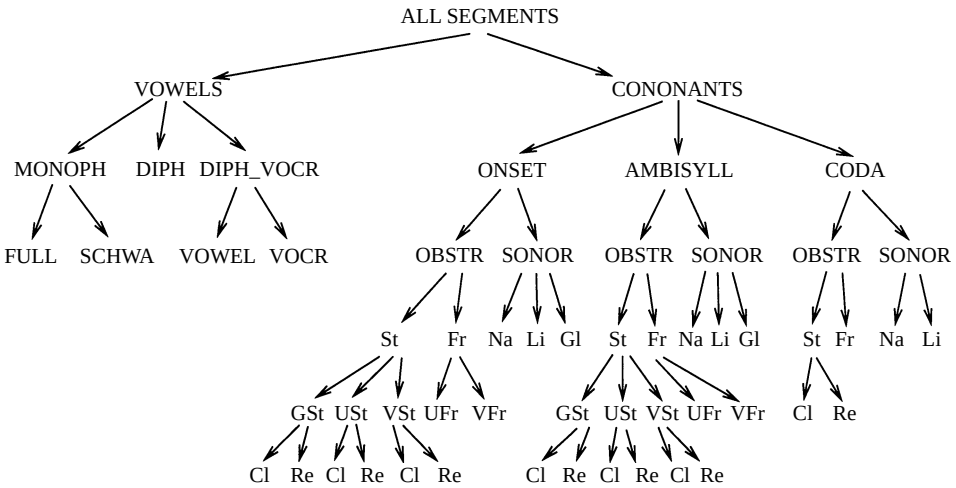

Figure 12. Category tree of the German duration system; MONOPH=monophthongs, DIPH = diphthongs, DIPH_VOCR = diphthongs involving [ɐ], VOCR = [ɐ], AMBISYLL = ambisyllabic, OBSTR = obstruents, SONOR = sonorants, St = stops, Fr = fricatives, Na = nasals, Li = liquids, Gl = glides, GSt = glottal stops, USt/VSt/UFr/VFr = unvoiced/voiced stops/fricatives, Cl = stop closure, Re = stop release.

3.2. Category tree

Constructing the duration system requires two major steps: setting up a category tree that splits up the factorial space, typically in terms of broad phoneme classes and intra-syllabic location, and selecting a particular sums-of-products model for each leaf of the tree.

Figure 12 shows the category tree of the German duration system. The tree represents a factorial scheme, i.e. the set of factors and distinctions on these factors that are known or expected to have a significant impact on segmental durations. Knowledge-based distinctions in the factorial scheme rely on three types of empirical information.

(1) **Conventional distinctions** based on phonetic and phonological features assigned to the segments, e.g. between vowels and consonants or between continuant and abrupt consonants. The underlying criteria for these distinctions are language independent.
(2) **Qualitative observations** as reported in the (sparse) research literature on segmental duration in German, such as: "Utterance-final lengthening affects the final two syllables only if the penultimate syllable is stressed, otherwise only the final syllable is affected" (Kohler, 1988).
(3) **Exploratory studies.** In a pilot experiment we found that the single most important segmental context factor for vowel duration was whether or not the syllable coda was empty, in other words, whether the vowel was the nucleus of an open or a closed syllable. The segmental composition of the coda was significantly less important.

Since our main goal was to develop a duration module as part of a TTS system, an important additional requirement in setting up the category tree was that the factors can be computed from text by the text analysis components of the system. The tree structure reflects a compromise between the attempt to obtain homogeneous classes by fine subcategorization and retaining a reasonable number of observations at each leaf of the tree. Note the we use *homogeneity* not in the sense of the cases at a leaf having similar durations (minimal vari-

TABLE III. Corrected means (in ms) for all segments in the database

Nucleus	ə	I	Y	U	ɐ	ɔ	e	a	œ							
	64	89	99	101	103	105	113	116	125							
Nucleus	i:	u:	y:	e:	o:	ø:	ɛ:	a:	aI	aU	ɔY					
	106	115	132	141	147	171	175	178	150	150	153					
Onset	dCl	gCl	bCl	ʔCl	tCl	kCl	pCl	bRe	dRe	gRe	ʔRe	tRe	pRe	kRe		
	36	38	52	53	71	91	119	11	11	18	10	11	12	22		
Onset	v	z	ʒ	h	ç	f	s	ʃ	n	m	l	r	j			
	66	92	93	51	80	90	91	92	52	57	63	71	71			
Ambisyl1	dCl	gCl	bCl	ʔCl	tCl	kCl	pCl	bRe	dRe	gRe	ʔRe	tRe	pRe	kRe		
	48	50	60	30	55	60	74	11	11	13	10	24	25	36		
Ambisyl1	ʒ	v	z	h	x	ç	f	s	ʃ	n	m	ŋ	r	l	j	
	55	59	71	52	72	90	96	98	105	56	71	75	46	51	84	
Coda	kCl	tCl	pCl	pRe	tRe	kRe	ç	x	f	s	ʃ	n	ŋ	m	r	l
	47	47	59	13	15	16	84	92	96	116	132	77	80	85	51	64

ance, as in CART), but in the sense that the same factors have the same effects on these cases, so that their behavior can be captured by one and the same sums-of-products model. The following categorical distinctions were made:

Vowels vs. consonants. This distinction is rather obvious and based on well-established phonetic and phonological knowledge, e.g. the observation that some factors like stress and speaking rate have, quantitatively speaking, very different effects on vowels than on consonants.

Vocalic distinctions. Vowels were subcategorized into central vowels (schwa), diphthongs, and full (non-central) monophthongs. An additional distinction was made for diphthongs that involve the low central vowel [ɐ] as a result of [r] vocalization. Whereas "regular" diphthongs [aI, aU, ɔY] are each treated as one segment in the acoustic inventory of the TTS system, diphthongs involving [ɐ] are generated by concatenating two segments; thus, durations have to be assigned to both components of these diphthongs.

Consonantal distinctions. The top level distinction among the consonants was based on the location in the syllable. Consonants are classified as being located in the onset or coda of the syllable, or as being ambisyllabic. All single intervocalic consonants are considered ambisyllabic. The next level of distinction was based on manner of articulation: stops, fricatives, nasals, liquids, and glides. Stops are subdivided into a closure and a release phase. In the onset and ambisyllabic locations, obstruents are further classified according to their voicing status. The voicing opposition is not applicable to obstruents in the syllable coda in German; exceptions to this rule (as for the [d] in *Redner* "speaker") are too small in number to justify a separate leaf in the tree.

3.3. Parameter estimation

In this analysis, we did not explore the full space of sums-of-products models; for practical reasons, we only fitted the additive and the multiplicative model. Since the multiplicative model had a uniformly better fit, we only report results on the latter. By fitting the multiplicative model, the resulting parameter estimates can be considered as approximations of the marginal means in a hypothetical database where each factorial combination occurs equally

TABLE IV. Results of model parameter estimation: syllable part, segment type (for legend see Fig. 12), number of observations, correlation and root-mean-squared deviation of observed and predicted data

Position	Segment class	Observ.	Corr.	RMS
Nucleus	full	4552	0.80	25
	schwa	906	0.71	18
	diph	693	0.74	30
	vow (bef. [ɐ]	840	0.75	22
	[ɐ] (aft. vow)	840	0.73	15
Onset	USt-Cl	1123	0.61	17
	VSt-Cl	795	0.74	15
	USt-Re	868	0.86	9
	VSt-Re	820	0.61	5
	UFr	1096	0.66	20
	VFr	368	0.72	16
	Na	451	0.46	18
	Li	487	0.42	17
	Gl	31	0.75	14
Ambisyl1	USt-Cl	458	0.55	17
	VSt-Cl	699	0.64	13
	USt-Re	410	0.63	14
	VSt-Re	792	0.46	4
	UFr	558	0.80	16
	VFr	251	0.63	14
	Na	681	0.59	15
	Li	293	0.36	15
	Gl	29	0.95	14
Coda	St-Cl	1187	0.67	19
	St-Re	1187	0.88	12
	Fr	1369	0.86	25
	Na	1917	0.66	22
	Li	306	0.67	20

often (a balanced design). For this reason, we call these parameter estimates *corrected means*. Table III shows the best estimates of corrected means for the entire database. Table IV gives correlations and root-mean-squared deviations of observed and predicted data. Because of differences in the numbers of observations and in the ranges of durations, these statistics are not strictly comparable with each other. The overall correlation between observed and predicted segmental durations for the entire database is 0.896. In comparison, correlations of 0.872 and 0.847 have been reported for Mandarin Chinese and French, respectively (Sproat, 1998, p. 138) as well as 0.880 for vowels and 0.940 for consonants in Japanese (Venditti & van Santen, 1998); these three languages were analyzed using the same methods and software.

3.4. Summary

We constructed a quantitative model of segmental duration in German by estimating the parameters of the model based on a segmented speech database. This approach uses statistical techniques that can cope with the problem of confounding factors and factor levels, and with data sparsity. The results show rather homogeneous patterns in that speech sounds within a given segment class generally exhibit similar durational trends under the influence of the same combination of factors. Among the most important factors are: (a) syllable stress (for nuclei, and to some extent for stops and fricatives in the onset); (b) word class (for nuclei); and (c) presence of phrase and word boundaries (for coda consonants, and to some extent for nuclei). The analysis yields a comprehensive picture of durational characteristics of one particular speaker. Note that work is under way to efficiently adapt the duration model, estimated on speech data from one speaker, to new speakers (Shih *et al.*, 1998).

4. Intonation

The task of the intonation module in a TTS system is to compute a fundamental frequency (F_0) contour from phonological representations consisting of a string of phoneme and syllabic stress symbols, and symbols related to phrasing and accenting information. The intonation component currently used in the Bell Labs multilingual TTS system has been thoroughly described in recent publications (van Santen & Möbius, 1997, 1999). We will therefore only highlight the most important aspects of the module.

The Bell Labs intonation model computes an 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; and segmental perturbation curves. We define an accent group as an entity that consists of an accented syllable followed by zero or more unaccented syllables.

The model also incorporates results from earlier work (van Santen & Hirschberg, 1994) that 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 (van Santen & Möbius, 1997, 1999). For any given segmental structure, the set of weights is called an alignment parameter matrix, and for each given pitch accent type the alignment parameter matrix characterizes how accent curves are aligned with accent groups.

The model describes and predicts in considerable detail the effects of segments (speech sounds) and their durations on the time course of the F_0 contour. Local pitch excursions associated with pitch accents (*accent curves*) are tied to the accented syllable in a complicated, yet tight manner. Two assumptions are central. First, following Möbius (1993) and Möbius *et al.* (1993), the phonological unit of a pitch accent is not the accented syllable, but the accent group. Second, the time course of an accent curve depends on the entire segmental and temporal structure of the accent group, not only on the properties of the accented syllable. While complicated, this dependency is deeply regular; all else being equal, the pitch peak, as measured from the start of the accented syllable, is shifted rightward as any part of the accent group is lengthened. It has been shown that this regularity can be captured by a simple linear alignment model. A key perceptual advantage of this detailed alignment model is that

intonation contours remain properly aligned with the segment and syllable boundaries, even in extreme cases such as early nuclear pitch accents that are followed by a large number of unaccented syllables, or very strong pitch excursions, or unusually long exclamation words.

Building on the well-known superpositional model proposed by Fujisaki (1983, 1988) and applied to a number of languages, including German (Möbius, 1993, 1995), we broaden the concept of superposition. In Fujisaki's model, an F_0 contour is generated by addition (in the log domain) of two different types of curves, viz. accent curves, generated by smoothing rectangular accent commands, and phrase curves, generated by smoothing impulse-like commands at the start and end of prosodic phrases. Smoothing is performed by filters that have specific mathematical forms. We have proposed several modifications to this model (van Santen & Möbius, 1999). First, we explicitly tie accent curves to accent groups. Second, we include segmental perturbation curves to capture the very large, but short-lived effects of obstruents on F_0 in the initial 50–100 ms of post-consonantal vowels. Third, we loosen the restrictions on the shape of phrase curves, which enables us to adequately model not only descending curves, which are observed in many languages, but also rise–fall patterns, as required for Japanese or Russian. Finally, accent curves are generated by means of a linear alignment model, not by a smoothed rectangular accent command.

For English and German, we found that phrase curves can be adequately modeled by splitting them into two consecutive parts. The two partial curves are obtained by non-linear interpolation between three points, viz. the start of the phrase, the start of the last accent group in the phrase, and the end of the phrase. In other words, the start of the final accent group serves as a pivot at which the two subcurves connect.

This model is currently used in the Bell Labs TTS system for English, French, German, Italian, Spanish, Russian, Romanian, and Japanese. In keeping with the multilingual characterization of the TTS system, the implementation of the intonation model makes no assumptions that are tied to any particular language. All language-specific data, represented by the sets of alignment parameters, are kept in external data files, which are read by the intonation model at run-time and can easily be altered for experimental purposes.

The weights that express the different degrees of influence on F_0 alignment exerted by the onset and rhyme durations and the duration of the remainder of the accent group, were originally determined by way of an analysis of American English speech data. To apply the model to German, these weights had to be re-estimated on speech data that is representative of segmental and accent group structure in German. Further adjustments involve the steepness and magnitude of rises and falls as well as the slope of the phrase curve. These adjustments were based on informal listening experiments.

5. Acoustic inventory and synthesis

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.

5.1. Inventory design

Speech is highly variable, and concatenative speech synthesis systems, such as Bell Labs TTS, are vulnerable to this variability. Different renditions of the same short phrases of

speech may sound identical to the listeners and yet, when two segments are excised from two different renditions and concatenated, spectral discrepances become audible. As a consequence, more than one rendition is needed for each target inventory unit, and speaking and recording conditions have to be carefully controlled.

The acoustic inventory for a given language is a set of stored speech segments that in its totality covers all legal phone sequences of that language. Further requirements are that, first, all the necessary phonemic and allophonic distinctions are reflected in the inventory structure and, second, the concatenation of two or more inventory units does not produce audible discontinuities in the resulting synthetic speech (see Section 5.3). Also, the final inventory should have a manageable size. In typical concatenative synthesis systems, the number of acoustic inventory units is kept relatively small by storing diphonic units. Prosodic modification is applied at run-time to cover the large combinatorial space that is spanned by the combinations of phoneme sequences and prosodic contexts; this modification is achieved by digital signal processing techniques. The challenge in inventory design is to capture key coarticulatory phenomena while at the same time keeping the number of units small.

The majority of units in the acoustic inventory of our German TTS system are diphones. Given a set of 43 allophones, the number of required diphones is 1849 on purely combinatorial grounds. However, as we have explained in more detail elsewhere (Olive *et al.*, 1998), a drastic reduction of inventory size from 1849 to slightly more than 1100 is possible by, first, excluding consonant pairs with minimal coarticulation and, second, applying phonotactic constraints. On the other hand, acoustic analyses of the speaker's vowel space and various coarticulatory effects that require the selection of context-sensitive units (Olive, 1990) increase the number of units.

For the sake of exemplification, let us first consider the vowel space of German (Fig. 11). We have already shown (Section 2.3.3) that for most pairs of long vs. short vowels, both allophones have to be represented in the acoustic inventories because the respective spectral properties are significantly distinct. The least clear-cut case is the question of how many phonetic qualities there are of the German vowel phoneme /a/. This issue has been addressed in the phonological and phonetic literature for many years and is largely undecided. Recently, Kohler concluded that the key difference between phonologically long and short /a/ is in the quantity domain (Kohler, 1995, p. 170). Figure 11 suggests, however, that for the definition of our TTS inventory a distinction between long /a:/ and short /a/ is preferable because the ranges delimited by the standard deviation boxes for the two vowels do not overlap. A post hoc analysis, i.e. after running the acoustic unit selection procedure described in the following section, clearly indicates that distinguishing between two /a/ qualities is the right choice to make.

The voicing status of sonorants and voiced fricatives depends on the identity of the adjacent segments (Shih & Möbius, 1998). These sounds are therefore candidates for context-sensitive units. As an example of such coarticulatory effects on voicing, Figure 13 illustrates the impact of the immediate segmental context on the acoustic realization of the phonologically voiced fricative /v/. The upper panel shows the waveform and spectrogram of the phone sequence [təvɛs], excised from the utterance *Er wollte Weste testen* [eːɐ vɔltə vɛstə tɛstən]. The fricative /v/ occurs in intervocalic position and is voiced throughout. The lower panel displays the waveform and spectrogram of the phone sequence [kvɛl], taken from the utterance *Er hatte Quelle gesagt* [eːɐ hatə kvɛlə gəzaːkt]. Here, /v/ is preceded by the voiceless stop /k/, and it is realized voiceless throughout. It is therefore necessary to include in the acoustic inventory one /v − ɛ/ unit for intervocalic position, where the /v/ realization is fully voiced, and a second /v − ɛ/ unit for the context of preceding voiceless obstruents, where the /v/

realization is entirely or partially devoiced. As a consequence, consonant-vowel diphones involving sonorants and voiced fricatives are represented in the inventory by two units, one for the voiced and the other for the unvoiced condition.

In its baseline version the acoustic inventory of our German TTS system 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. An augmented version further includes units that represent speech sounds that occur in common foreign words or names, such as the voiced and voiceless interdental fricatives and the glide /w/ for English, and nasalized vowels for French.

5.2. Recordings

The German speech database was constructed by systematically varying the contextual place of articulation for each diphone unit. In the case of the targeted /i − m/ diphone, for instance, recordings were made for the labial /p − i − m/, dental /t − i − m/, velar /k − i − m/, and rhotic /r − i − m/ contexts, thus spanning the whole range of possible formant trajectories in the diphone /i − m/. Stops were chosen as context phones because they have a minimal contaminating effect on the vowel spectrum.

The resulting triphonic segments were embedded in a carrier phrase, thereby keeping the prosodic context as constant as possible. The key words were placed in focal position in the carrier phrase. The syllable containing the target unit, however, did not carry primary stress; it carried secondary stress in order to avoid over-articulation (cf. Portele *et al.*, 1991). An example sentence to be recorded for the diphone /i − m/ in the preceding dental context thus reads:

(18) *Er hatte Timmerei gesagt.*

with primary stress on *−rei*, secondary stress on the key syllable *Tim−*, and a dental context [t] for the target unit /i − m/. Obviously, the carrier phrase needs to be varied slightly to accommodate utterance initial and final units but the construction principles are the same.

While using real words and sentences for the recordings as opposed to constructed carrier words and phrases may appear appealing at first glance, the latter approach has significant advantages: (a) controlled prosodic context and position in the phrase; (b) uniform stress level on all target inventory units; (c) systematically varied phonemic context; and (d) no need for the syntactic constraints desirable for sentences using real words.

This approach, in combination with the exclusion of consonant pairs with minimal coarticulation and the application of phonotactic constraints, results in a fairly small number of sentences to be recorded—even though one sentence is needed for each target inventory unit. For the baseline system we recorded approximately 3800 sentences, yielding on average three candidates for each target unit.

5.3. Unit selection

The best candidate for each target inventory unit was selected based on various criteria including spectral discrepancy and energy measures. We used a procedure that performs an automated optimal unit selection and cut point determination (Olive *et al.*, 1998). The algorithm selects units such that spectral discrepancies between units as well as the distance of each sound from its "spectral ideal" are simultaneously minimized, and the coverage by good candidates of required units is maximized.

This procedure is a hard optimization problem that until recently has only been resolved

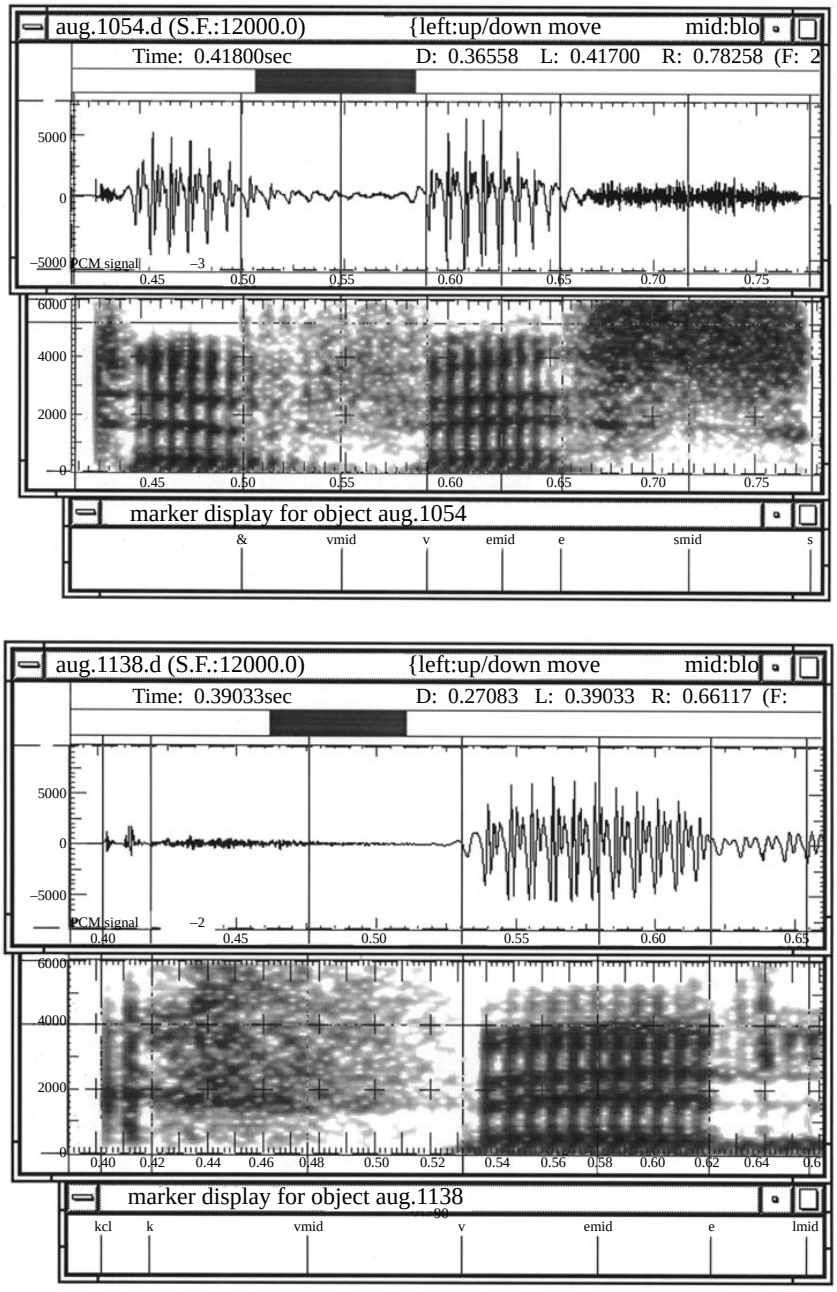

Figure 13. Coarticulatory effects: acoustic realization of the phonologically voiced fricative /v/ in different segmental contexts. Intervocalic position (upper panel): waveform and spectrogram of the phone sequences [təvɛs] from the utterance "Er wollte Weste testen." [eːʁ vɔltə vɛstə tɛstən]; /v/ is voiced throughout. Preceding voiceless stop /k/ (lower panel): waveform and spectrogram of the phone sequence [kvɛl] from the utterance "Er hatte Quelle gesagt" [eːʁ hatə kvɛlə gəzaːkt]; /v/ is realized voiceless throughout.

by means of approximative methods (e.g. Iwahashi & Sagisaka, 1992). We have reduced the complexity of the problem by introducing the concept of *ideal point*. For each vowel, the ideal point is a point in the three-dimensional $F_1/F_2/F_3$ space that has the following property: for each target inventory unit there is at least one candidate in the speech database whose formant trajectory passes the ideal point within a pre-determined distance d. This method guarantees that if all best candidates are cut at their points of smallest distance to the ideal point, the spectral discrepancy between any two such units at the time of concatenation will be at most $2d$. If d is sufficiently small, the discrepancies will be imperceptible.

In the three-dimensional formant space, only three parameters have to be optimized. Note that the algorithm is applicable to speech representations other than formants as well.

5.4. Synthesis

At run time, the required units are retrieved from the acoustic inventory and concatenated. Further steps involve appropriate interpolation and smoothing operations, assigning new durations, F_0 contours and amplitude profiles, and finally passing 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 (Oliveira, 1993) 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, in considerable detail, a description of the German version of the Bell Labs multilingual TTS system. Mirroring the overall modular structure of the system, the text analysis component itself consists of a multitude of modules which operate on different levels of linguistic description and analysis. This inherently heterogeneous component has been implemented in a unified framework, viz. weighted finite-state transducer technology. Despite the differences in the formalisms applied to different subtasks of text analysis, the linguistic descriptions can all be compiled into WFSTs. The automata serve as precompiled language-specific data input to the generalized text analysis software for multilingual TTS.

The duration component has been realized as a quantitative model of the durations of speech sounds in German whose parameters were estimated from a segmented speech database. This approach uses statistical techniques that can cope with the problem of confounding factors and factor levels, and with data sparsity. Similarly, the intonation component implements a quantitative model that describes and predicts the effects of speech sounds and their durations on the time course of the fundamental frequency contour. This approach shares some concepts with other superpositional intonation models; the key novelty in our approach is that we account for details (“anchor points”) of pitch accent curves that depend on the composition and duration of accent groups.

We then discussed design criteria for the construction of acoustic inventories for concatenative synthesis. We have shown how the size of the inventory can be reduced by applying phonotactic constraints and by excluding units that cover sequences of phones with minimal coarticulation, and we have proposed a carrier phrase approach for the reading materials as a means of keeping the segmental and prosodic context for the target unit constant. The best

candidate for each target inventory unit was then selected based mostly on spectral discrepancy measures. The proposed algorithm achieves a globally minimized inter-unit spectral discrepancy and a minimized distance of each sound from its "spectral ideal".

An interactive demonstration of the system is accessible on the World Wide Web at `http://www.bell-labs.com/project/tts/german.html`.

I am grateful to my colleagues in the TTS group at Bell Labs, for their continuous support and encouragement. In particular, I wish to thank Joe Olive, Jan van Santen, Chilin Shih, Richard Sproat and Michael Tanenblatt. Stefanie Jannedy (The Ohio State University) worked with me on the analysis and pronunciation of names. I also acknowledge the constructive suggestions by two anonymous reviewers.

References

Allen, J., Hunnicutt, M. S., Klatt, D., Armstrong, R. C. & Pisoni, D. B. (1987). *From Text to Speech: The MITalk System*, Cambridge University Press, Cambridge.

Bartkova, K. & Sorin, C. (1987). A model of segmental duration for speech synthesis in French. *Speech Communication*, **6**, 245–260.

Belhoula, K. (1993). A concept for the synthesis of names. *ESCA Workshop on Applications of Speech Technology*, Lautrach, Germany.

Campbell, W. N. (1992). Syllable-based segmental duration. In *Talking Machines: Theories, Models, and Designs*, (Bailly, G., Benoît, C. and Sawallis, T. R., eds), pp. 211–224. Elsevier Science, Amsterdam.

Carlson, R. & Granström, B. (1986). A search for durational rules in a real-speech database. *Phonetica*, **43**, 140–154.

D-Info, (1995). D-Info—Adress- und Telefonauskunft Deutschland. CD-ROM. TopWare, Mannheim.

Duden, (1984). *Duden Grammatik der deutschen Gegenwartssprache*, Dudenverlag, Mannheim.

Duden, (1987). *Wörterbuch der Abkürzungen*, Dudenverlag, Mannheim.

Fujisaki, H. (1983). Dynamic characteristics of voice fundamental frequency in speech and singing. In *The Production of Speech*, (MacNeilage, P. F., ed), pp. 39–55. Springer, New York.

Fujisaki, H. (1988). A note on the physiological and physical basis for the phrase and accent components in the voice fundamental frequency contour. In *Vocal Physiology: Voice Production, Mechanisms and Functions*, (Fujimura, O., ed), pp. 347–355. Raven, New York.

Iwahashi, N. & Sagisaka, Y. (1992). Speech segment network approach for an optimal synthesis unit set. *Proceedings of the International Conference on Spoken Language Processing*, Banff, Alberta, volume 1, pp. 479–482.

Jannedy, S. & Möbius, B. (1997). Name pronunciation in German text-to-speech synthesis. *Proceedings of the Fifth Conference on Applied Natural Language Processing*, Washington, DC, pp. 49–56. ACL.

Kaiki, N., Takeda, K. & Sagisaka, Y. (1990). Statistical analysis for segmental duration rules in Japanese speech synthesis. *Proceedings of the International Conference on Spoken Language Processing*, Kobe, pp. 17–20.

Kiel Corpus, The Kiel corpus of read speech, volume 1. CDROM.

Klatt, D. H. (1973). Interaction between two factors that influence vowel duration. *Journal of the Acoustical Society of America*, **54**, 1102–1104.

Klatt, D. H. (1976). Linguistic uses of segmental duration in English: acoustic and perceptual evidence. *Journal of the Acoustical Society of America*, **59**, 1209–1221.

Kohler, K. J. (1988). Zeitstrukturierung in der Sprachsynthese. *ITG-Fachbericht*, **105**, 165–170.

Kohler, K. J. (1994). Lexica of the Kiel PHONDAT corpus, volume 1/2. *Arbeitsberichte des Instituts für Phonetik und digitale Sprachverarbeitung, Univ. Kiel, AIPUK, 27/28*.

Kohler, K. J. (1995). *Einführung in die Phonetik des Deutschen*, 2nd edition, Erich Schmidt Verlag, Berlin.

Lieber, R. (1987). *An Integrated Theory of Autosegmental Processes*, State University of New York Press, Albany.

Maghbouleh, A. (1996). An empirical comparison of automatic decision tree and hand-configured linear models for vowel durations. *Computational Phonology in Speech Technology: Proceedings of the Second Meeting of ACL SIGPHON*, Santa Cruz, CA, pp. 1–7.

Möbius, B. (1993). *Ein quantitatives Modell der deutschen Intonation—Analyse und Synthese von Grundfrequenzverläufen*, Max Niemeyer Verlag, Tübingen.

Möbius, B. (1995). Components of a quantitative model of German intonation. *Proceedings of the 13th International Congress of Phonetic Sciences*, Stockholm, volume 2, pp. 108–115.

Möbius, B. (1998). Word and syllable models for German text-to-speech synthesis. *Proceedings of the Third ESCA Workshop on Speech Synthesis*, Jenolan Caves, Australia, pp. 59–64.

Möbius, B., Pätzold, M. & Hess, W. (1993). Analysis and synthesis of German F0 contours by means of Fujisaki's model. *Speech Communication*, **13**, 53–61.

Möbius, B. & van Santen, J. (1996). Modeling segmental duration in German text-to-speech synthesis. *Proceedings of the International Conference on Spoken Language Processing*, Philadelphia, PA, volume 4, pp. 2395–2398.

Mohri, M. & Sproat, R. (1996). An efficient compiler for weighted rewrite rules. *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, Santa Cruz, CA, pp. 231–238.

Olive, J., van Santen, J., Möbius, B. & Shih, C. (1998). Synthesis. In *Multilingual Text-to-Speech Synthesis: The Bell Labs Approach*, chapter 7, (Sproat, R., ed), pp. 191–228. Kluwer Academic, Dordrecht, Boston, London.

Olive, J. P. (1990). A new algorithm for a concatenative speech synthesis system using an augmented acoustic inventory of speech sounds. *Proceedings of the ESCA Workshop on Speech Synthesis*, Autrans, (Bailly, G. and Benoît, C., eds), pp. 25–29.

Olive, J. P. & Liberman, M. Y. (1985). Text to speech—an overview. *Journal of the Acoustical Society of America*, **78** (Suppl. 1), S6.

Oliveira, L. C. (1993). Estimation of source parameters by frequency analysis. *Proceedings of the European Conference on Speech Communication and Technology*, Berlin, volume 1, pp. 99–102. ESCA.

Onomastica, (1995). Multi-language pronunciation dictionary of proper names and place names. Technical report, European Community, Linguistic Research and Engineering Programme, European Commission, DG XIII. Project No. LRE-61004, Final Report, 30 May 1995.

Pitrelli, J. F. & Zue, V. W. (1989). A hierarchical model for phoneme duration in American English. *Proceedings of the European Conference on Speech Communication and Technology*, Paris, pp. 324–327. ESCA.

Portele, T., Steffan, B., Preuss, R. & Hess, W. (1991). German speech synthesis by concatenation of non-parametric units. *Proceedings of the European Conference on Speech Communication and Technology*, Genoa, volume 1, pp. 317–320. ESCA.

Riley, M. D. (1992). Tree-based modeling for speech synthesis. In *Talking Machines: Theories, Models, and Designs*, (Bailly, G., Benoît, C. and Sawallis, T. R., eds), pp. 265–273. Elsevier Science, Amsterdam.

Roche, E. & Schabes, Y. (eds) (1997). *Finite-State Language Processing*, MIT Press, Cambridge.

Shih, C. & Ao, B. (1997). Duration study for the Bell Laboratories Mandarin text-to-speech system. In *Progress in Speech Synthesis*, (van Santen, J., Sproat, R. W., Olive, J. P. and Hirschberg, J., eds), pp. 383–399. Springer, New York.

Shih, C., Gu, W. & van Santen, J. P. H. (1998). Efficient adaptation of TTS duration model to new speakers. *Proceedings of the International Conference on Spoken Language Processing*, Sydney, volume 2, pp. 25–28.

Shih, C. & Möbius, B. (1998). Contextual effects on voicing profiles of German and Mandarin consonants. *Proceedings of the Third ESCA Workshop on Speech Synthesis*, Jenolan Caves, Australia, pp. 81–86.

Simoes, A. R. M. (1990). Predicting sound segment duration in connected speech: an acoustical study of Brazilian Portuguese. *Proceedings of the ESCA Workshop on Speech Synthesis*, (Bailly, G. and Benoît, C., eds), pp. 173–176. Autrans.

Sproat, R. (ed.) (1996). Multilingual text analysis for text-to-speech synthesis. *Proceedings of the International Conference on Spoken Language Processing*, Philadelphia, PA, volume 3, pp. 1365–1368.

Sproat, R. (ed.) (1998). *Multilingual Text-to-Speech Synthesis: The Bell Labs Approach*, Kluwer Academic, Dordrecht; Boston, London.

Thielen, C. (1995). An approach to proper name tagging in German. *From Text to Tags—Issues in Multilingual Language Analysis. Proceedings of the ACL SIGDAT Workshop*, University College, Belfield, Dublin, Ireland, pp. 35–40.

Traber, C. (1995). *SVOX: The Implementation of a Text-to-Speech System for German*, VDF Hochschulverlag, Zürich.

van Leeuwen, J. (ed.) (1990). *Handbook of Theoretical Computer Science*, volume B, Elsevier; MIT Press, Amsterdam; Cambridge.

van Santen, J. P. H. (1992). Contextual effects on vowel durations. *Speech Communication*, **11**, 513–546.

van Santen, J. P. H. (1993*a*). Analyzing N-way tables with sums-of-products models. *Journal of Mathematical Psychology*, **37**, 327–371.

van Santen, J. P. H. (1993*b*). Perceptual experiments for diagnostic testing of text-to-speech systems. *Computer Speech and Language*, **7**, 49–100.

van Santen, J. P. H. (1993*c*). Timing in text-to-speech systems. *Proceedings of the European Conference on Speech Communication and Technology*, Berlin, volume 2, pp. 1397–1404. ESCA.

van Santen, J. P. H. (1994). Assignment of segmental duration in text-to-speech synthesis. *Computer Speech and Language*, **8**, 95–128.

van Santen, J. P. H. & Hirschberg, J. (1994). Segmental effects on timing and height of pitch contours. *Proceedings*

of the International Conference on Spoken Language Processing, Yokohama, pp. 719–722.

van Santen, J. P. H. & Möbius, B. (1997). Modeling pitch accent curves. *Intonation: Theory, Models and Applications—Proceedings of an ESCA Workshop*, Athens, (Botinis, A., Kouroupetroglou, G. and Carayiannis, G., eds), pp. 321–324.

van Santen, J. P. H. & Möbius, B. (1999). A model of fundamental frequency contour alignment. In *Intonation: Analysis, Models and Speech Technology*, (Botinis, A., ed), Cambridge University Press, Cambridge, Forthcoming

Venditti, J. J. & van Santen, J. P. H. (1998). Modeling segmental durations for Japanese text-to-speech synthesis. *Proceedings of the Third ESCA Workshop on Speech Synthesis,* Jenolan Caves, Australia, pp. 31–36.

Wiese, R. (1996). Phonological versus morphological rules: on German Umlaut and Ablaut. *Journal of Linguistics*, **32**, 113–135.

Yarowsky, D. (1994). Homograph disambiguation in speech synthesis. *Proceedings of the Second ESCA/IEEE Workshop on Speech Synthesis*, New Paltz, NY, pp. 244–247.

(Received 1 December 1998 and accepted for publication 18 June 1999)