INTERNATIONAL JOURNAL OF SPEECH TECHNOLOGY 6, 57–71, 2003


# Rare Events and Closed Domains: Two Delicate Concepts in Speech Synthesis

BERND MÖBIUS
*Institute of Natural Language Processing, University of Stuttgart, Azenbergstraße 12, D-70174 Stuttgart, Germany*
Bernd.Moebius@IMS.Uni-Stuttgart.DE

**Abstract.** One of the most serious challenges for speech synthesis is the systematic treatment of events in language and speech that are known to have low frequencies of occurrence. The problems that extremely unbalanced frequency distributions pose for rule-based or data-driven models are often underestimated or even unrecognized. This paper discusses the problems pertinent to rare events in four components of speech synthesis systems: in linguistic text analysis, where productive word formation processes generate a potentially unbounded lexicon and cause heavily skewed word frequency distributions; in syllabification, where some syllables occur very frequently but most phonotactically possible syllables are very infrequent; in speech timing, where most constellations of factors affecting segmental duration are sparsely or not at all represented in training databases; and in unit selection synthesis, where the uneven distribution of speech unit frequencies poses challenges to speech corpus design. Currently available techniques for coping with the problem of rare or unseen events in each of these components are reviewed. Finally, a distinction is made between a strictly closed domain with a fixed vocabulary and a merely restricted domain with loopholes for unseen words and names, and the consequences of the respective type of domain for appropriate synthesis strategies are discussed.

**Keywords:** speech synthesis, low-frequency events, linguistic analysis, prosody, unit selection

## 1. Introduction

In this paper I intend to point out two common concepts in speech synthesis that I consider delicate, if not misguided and wrong. The first of these concepts is the often nonchalant treatment of phenomena in language and speech that are known or assumed to have low frequencies of occurrence.

In the context of text-to-speech synthesis (TTS), such low-frequency events play an important role in linguistic text analysis, in the form of heavily skewed word frequency distributions, caused to a large extent by productive word formation processes (Section 2.1), as well as in the context of syllabification (Section 2.2). Extremely uneven frequency distributions are also observed in segmental duration modeling, where most factorial constellations are sparsely or not at all represented in training databases (Section 2.3). The fourth area in TTS conversion that is affected by non-uniform frequency distributions is the design of acoustic unit inventories for data-driven speech synthesis (Section 2.4).

Various statistical techniques have been developed to cope with the problem of events that are rare or unseen in training databases and yet must be expected to occur in the text input to an open-domain TTS system. In the context of word frequencies and, more generally, language modeling, models based on Zipf's law and Good-Turing estimates are sometimes applied to predict the frequencies of unseen events, such as unseen words formed by a given morphological process. Another standard technique for assigning probabilities to data unseen in the training corpus is the Expectation Maximization algorithm, which I will discuss in some detail in the context of syllabification. A class of statistical models known as sums-of-products models has been shown to yield very good results for the task of assigning segmental durations in speech synthesis.

Finally, certain techniques that are well-established in speech recognition have been applied to cope with the LNRE problem in corpus-based unit selection synthesis, such as the prediction of the properties of unseen units (e.g., triphones, phone-sized units, subphone units) by interpolation from the known properties of similar units. With the possible exception of duration models, whose current performance appears to leave little room for further improvement at least on the segmental level, the LNRE problem is not removed in any of the contexts discussed below.

The second concept that I consider questionable is the notion of a "restricted" application domain (Section 3). I suggest that, at least in languages with a large number of distinct syllables, such as English or German, word or syllable concatenation schemes are only feasible in strictly closed domains, i.e. those domains that have a fixed and unchanging vocabulary. However, some progress has recently been made in the design and construction of unit selection voices that offer, on the one hand, very good synthesis quality whenever there is a close match between the application domain and the domain in which the speech corpus was collected and, on the other hand, fair synthesis quality for open-domain applications. The obvious impact of the domain in these studies only emphasizes the importance of a careful database design; how to achieve the transfer from one limited domain to the other, let alone to an open domain scenario with unrestricted text input, remains a challenge for TTS research.

## 2. Rare Events

Several phenomena in language and speech can be characterized as belonging to the LNRE class of distributions. LNRE is the acronym for *Large Number of Rare Events*, apparently first introduced by Khmaladze (1987) as a descriptor of distributions for which the law of large numbers does not hold. LNRE classes have the property of extremely uneven frequency distributions: while some members of the class have a high frequency of occurrence, i.e. they are *types* with a large *token* count, most class members are extremely rare. Even logarithmically transformed word frequencies, for instance, are not normally distributed but remain skewed.

A special case of such a distribution is known as *Zipf's Law* (Zipf, 1935, 1949), which gives a fairly good approximation of, for example, word frequency counts. However, Zipfian models suffer from the fact that the model parameters change systematically as a function of the sample size and thus need to be continuously adjusted to accommodate changes in sample size. The same holds for the log-normal model, whose parameters (mean and standard deviation) appear as increasing functions of sample size (Baayen, 2001).

The relationship between Zipf's law and Turing's formula (Good, 1953) was explored in Samuelsson (1996). Both formulas are of interest in natural language processing because they can be used to improve probability estimates from relative frequencies as well as to predict the frequencies of unseen events, such as unseen words formed by a given morphological process. Turing's formula, in particular in the flavor of the back-off method (Katz, 1987), is now a standard technique in speech recognition, where it is used to improve the estimation of parameters in probabilistic language models. Samuelsson shows that in contrast to common belief in the field, the ideal Turing distribution does not have Zipf's law as some special or limiting form but is qualitatively different. More concretely, Turing's formula provides an appropriate probability distribution even for infinite populations, whereas Zipf's law implies a finite total population.

In my work on German and multilingual speech synthesis (Möbius, 1998a, 1999, 2001). I have encountered LNRE distributions in three contexts: in linguistic text analysis, in segmental duration modeling, and in acoustic inventory design. Many TTS systems rely on a full-form pronunciation dictionary in conjunction with generic pronunciation rules. Words in the input text are usually looked up in the pronunciation dictionary or, if not listed there, transcribed by rule. The main problem with this approach is the *productivity of word formation* processes, both derivational and compositional ones, in particular in German but more generally in almost any natural language.

The work of Baayen (2001) reveals that monomorphemic content words, viz. nouns, adjectives and verbs, are outside the LNRE zone, but that frequencies of words formed by productive derivational affixes, for instance, have typical LNRE distributions. The LNRE zone, according to Baayen, can be defined as the range of successively increasing sample sizes taken from a corpus where one keeps finding previously unseen words. Note that samples in this definition are not independent of each other: sample number $n + 1$ is assumed to include sample number $n$. For word frequency estimations, even large corpora (tens of millions of words) are generally within the LNRE zone. This means that in

open-domain TTS, the probability of encountering previously unseen words in the input text is very high. A TTS system therefore needs to be capable of analyzing unknown words (Section 2.1).

Languages with complex *syllable structure*, such as English or German, are known to have a large inventory of distinct syllables, whose frequency distributions also display typical LNRE characteristics. A few hundred distinct syllables account for the majority of realized syllable tokens in speech production, whereas most syllables in the inventory are very rarely used. Preferred approaches to syllabification are therefore those that can assign probabilities to under-represented or even unseen syllables (Section 2.2).

Similarly unpleasant frequency distributions are observed in *segmental duration* modeling (Section 2.3). There are many factors that affect speech timing, such as the identity of the speech sound and its neighbours as well as positional and prosodic factors. The number of different constellations of these factors is language-dependent; for English and German more than 10,000 distinct constellations exist (van Santen, 1995; Möbius and van Santen, 1996), and their frequency distributions belong to the LNRE class: most observed constellations have a very low frequency of occurrence.

LNRE distributions also pose problems for the design of *acoustic unit inventories* for concatenative speech synthesis (Section 2.4). This observation holds especially for corpus-based synthesis systems that perform an online unit selection from a large annotated speech database.

### *2.1. Morphological Productivity*

Text input to a general-purpose TTS system is likely to contain words that are not listed in the TTS lexicon. All natural languages have productive word formation processes, and the community of speakers of a language creates novel words (and names) as need arises.

It has been suggested that productivity be distinguished from creativity (Schultink, 1961). Productivity is a notion based on linguistic rules. Words formed by means of productive morphological processes are usually not noticed by the listener as new words and not formed by the speaker by any conscious, intentional effort. Creativity, in contrast, is not restricted to morphology but rather a general cognitive ability. Words formed by creative processes are carefully and intentionally produced and often perceived as new words. Forming a new adjective by attaching the negation prefix *un-* to an adjectival base will probably go unnoticed by both speaker and listener in English or German, as long as certain constraints are not violated. In contrast, forming the new adjective *kaputtbar* in German by attaching the (productive) deverbal suffix *-bar* to an adjectival base violates the constraint that this suffix can only attach to transitive verb bases; this word formation product requires a conscious effort by the speaker and will trigger a strong reaction by the listener. There is a gray zone, of course: attaching a denominal suffix to form the new adjective *nerdulent* is morphologically regular and semantically transparent but involves an unproductive pattern.

Productive word formation patterns are unlimited. In German and a number of other languages, derivation and compounding are the most important means of productive word formation, and they can generate an unlimited number of new words. The construction of a finite, exhaustive lexicon that contains all the words in the language is therefore impossible.

In a language like German, where deriving the pronunciation of a word from its spelling is difficult and where pronunciation and syllabic stress rules require access to the morphological structure of the word, a TTS system needs a component that linguistically analyzes words that are unknown to the system. This is where the distinction between productivity and creativity is relevant. Productive processes are morphosyntactically and semantically regular: this is why new words formed by productive processes are not consciously coined and not recognized as new words. It is therefore useful to know which word formation patterns can be modeled by rules and which ones have to be listed, and quantitative studies can provide this knowledge.

A simple statistical estimate of productivity has been suggested, and applied, by Baayen (1993). His approach exploits the observation that the proportion of *hapax legomena* in a text corpus is much higher for intuitively productive affixes than for unproductive ones. Hapax legomena are here defined relative to a text corpus. Given a particular word-forming affix, all distinct word types in the corpus that are formed by this affix are listed and their token frequencies are counted; a hapax legomenon is a—morphologically complex—word type with a token count of 1. Under certain simplifying assumptions the productivity index ($P$) of an affix can then be expressed as the ratio of hapax legomena ($n1$) to the total number of tokens formed by that affix in the corpus ($N$): $P = n1/N$.

This estimate of morphological productivity has been integrated into the linguistic text analysis component of the Bell Labs German TTS system (Möbius, 1999). It is applied to the analysis of morphologically complex words (and names) that are unknown to the system. This analysis component is based on a model of the morphological structure of words and the phonological structure of syllables, building on a quantitative study of the productivity of word forming affixes in German (Möbius, 1998b). Thus, the TTS system has the capability to decompose unknown words morphologically and to provide for these words an annotation whose granularity approaches that of the annotation of words listed in the TTS lexicon.

The productivity index ($P$) corresponds to the rate at which new word types are encountered when more and more tokens generated by a given morphological process are sampled. More precisely, it uses as a measure of productivity the slope of the word type growth curve after the entire corpus has been sampled. If the *vocabulary* is defined as the number of distinct types that a morphological process can generate, then a truly productive word formation pattern may be characterized by an infinite vocabulary, whereas an unproductive pattern is expected to have a finite, and often quite small, vocabulary (Evert and Lüdeling, 2001). Based on a given text corpus, the word type growth curve of a morphological process is obtained by plotting the number of distinct types ($V$) encountered as a function of the number of tokens ($N$) formed by the process in the corpus (Fig. 1). The growth curve of an unproductive process will flatten out and converge to a constant value after enough data have been sampled. The type count of a productive pattern will continue to grow indefinitely.

Given the fact that word frequency distributions tend to have large growth rates even at the full sample and that many more word types are expected to be encountered whenever more word tokens are added, Good-Turing estimates should be more appropriate for measuring morphological productivity than Zipfian models. Good-Turing estimates make a fraction of the total probability mass free for the as yet unseen word types; indeed, this fraction is equal to the word type growth rate at a given sample size. Yet, the Good-Turing method has its limitations too, and Baayen (2001) provides a thorough review of where and why this is the case. For instance, he demonstrates that for sample sizes in the LNRE zone the use of the relative sample frequencies results in a severe underestimation of the vocabulary size and that this is a problem not only for Zipfian models but for the Good-Turing method too, even though the latter adjusts the sample-relative frequencies. In addition, extrapolation to sample sizes significantly larger than the sample size upon which the model has been conditioned will break down for technical reasons. Thus, in a nutshell, the limitations are pertinent to both interpolation and extrapolation from a given sample size.

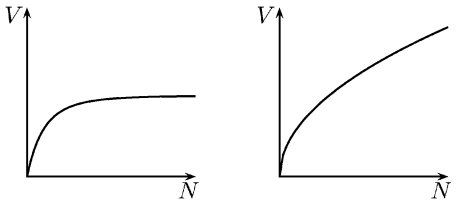


*Figure 1.* Typical, idealized shapes of word type growth curves ($V$ = types, $N$ = tokens): the curve pertaining to an unproductive pattern will flatten out (left panel), whereas the type count of a productive pattern will continue to grow indefinitely (right panel). Adapted from Evert and Lüdeling (2001).

Further elaborate statistical methods exist for estimating word frequency distributions and morphological productivity and, more generally, for coping with the LNRE distributions of word frequencies. A review of these methods and further refinement of some of them are presented in Baayen (2001), along with applications to word frequency distributions, morphological productivity, consonant-vowel pattern distributions, and word co-occurrences (bigrams).

One important conclusion from this work is that the word type growth curve, and therefore also the productivity index ($P$), is a function of the sample size. Even though Baayen proposes various statistical models that address this problem, it is still hard, if not impossible, to compare the productivity of two morphological processes with substantially differing sample sizes. For instance, the slope of the growth curve of the unproductive pattern in Fig. 1 is flat for the full sample, as expected, but in the early part of the unproductive curve the slope may be as steep as, or even steeper than that of the productive pattern for the full sample. If we do not know where exactly in the curve we are, we cannot compare the productivity indices of patterns with substantially differing sample sizes.

Another relevant implication of Baayen's work is that the text corpora used in the earlier studies (Baayen and Lieber, 1991; Möbius, 1998b) were too small for reliable estimates—too small by several orders of magnitude. As it turns out, even large corpora (tens of millions

of words) are generally still within the LNRE zone; that is, as the sample size increases incrementally, one keeps finding previously unseen word types, and it is hard to predict the future growth rate.

In a research project on derivational and compositional morphology of German (Schmid et al., 2001) a number of problems pertaining to the application of the productivity measures was encountered. For instance, it was demonstrated that corpus data have to be thoroughly preprocessed before they can be used in the statistical models applied to the quantitative analysis of morphological productivity (Lüdeling et al., 2000; Evert and Lüdeling, 2001). Raw data correction and clean-up is required because of errors in the corpus (e.g., misspellings, or repeated sections that affect frequency distributions) and because automatic annotation is flawed (e.g., part-of-speech tagging errors). In addition, there are linguistic factors that need to be properly addressed: (i) compounding is a major source of hapax legomena, but a complex word should be counted as a new type only if compounding happens before derivation (the complex German noun *Kinderreichtum* 'having many children' is a derivation of the adjective *kinderreich* 'prolific' and therefore counts as a new type; the erroneous reading as a result of compounding *Kinder* 'children' and the already derived *Reichtum* 'wealth' does not give rise to a new type); (ii) the order in which affixes attach matters (*unverzichtbar* 'indispensable' counts as a new type of the *un-* pattern but not of the *-able* pattern because *-able* attaches first, and *unverzicht* is not a stem); (iii) accidental substrings must be discounted (*Balsam* 'balm' is not formed by the suffix *-sam*); (iv) creative word formation products must be discounted.

The nature of the problems described above is a serious obstacle to automatic preprocessing: either the problems are introduced by the automatic tools themselves, as in the case of flawed part-of-speech annotation, or the solution requires careful analysis of linguistic structure, e.g. hierarchical morphological structure, that presently cannot be performed automatically. Indeed, designing tools with the desired capabilities presupposes precisely the kind of linguistic knowledge that we are currently trying to build up—a vicious circle. Therefore, only manual clean-up and correction yields reliable input to the statistical models. Unfortunately, manual preprocessing is not feasible for corpora of the required size, and the available automatic procedures, while yielding some improvement over the uncorrected data, are not sufficiently reliable (Lüdeling et al., 2000; Evert and Lüdeling, 2001).

Figure 2 displays raw and manually corrected word type growth curves for the German adjective-forming suffixes *-bar* and *-sam*. The raw curves suggest that the two morphological patterns have similar productivity rates. Both suffixes have sample sizes of the same order of magnitude ($N$(*-bar*) = 37783, $N$(*-sam*) = 22667); both appear to be productive because they generate many new word types ($P$(*-bar*) = 0.0086, $P$(*-sam*) = 0.0034); and even though the axis scalings are not identical, it is clear that the shapes of the raw curves are quite similar.

However, native-speaker intuition predicts that *-bar* is productive, whereas *-sam* is intuitively unproductive. Only the corrected curves reflect the expected characteristics. In quantitative terms, this becomes manifest in the productivity indices too ($P$(*-bar*) = 0.0053, $P$(*-sam*) = 0.0002).

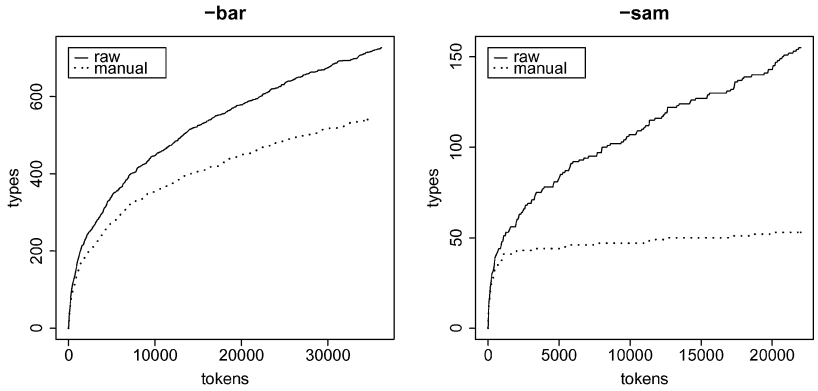


*Figure 2.* Computed word type growth curves of the German adjective-forming suffixes *-bar* and *-sam*. The raw curves (continuous lines) suggest that the two morphological patterns have very similar productivity rates. Only after manual correction (dotted lines) do the curves reflect the expected characteristics: *-bar* is intuitively productive, whereas *-sam* is intuitively unproductive. Adapted from Evert and Lüdeling (2001).

The results of this case study (Evert and Lüdeling, 2001) lead us to conclude that sufficiently reliable correction results can only be achieved by a morphology system that, besides derivation and compounding analysis (and generation) capabilities, also computes the hierarchical structure of complex words, building on a model of the order in which word formation processes operate on a simplex form. In short, such an ideal morphology system would perform like a human corrector. Given the current state of the art in automatic linguistic analysis, there does not seem to be a usable alternative to manual preprocessing and correction.

### 2.2. *Syllabification*

Syllabification is an important component of speech synthesis systems. In many languages the pronunciation of phonemes is a function of their location in the syllable relative to the syllable boundaries. Location in the syllable also has a strong effect on the duration of the phone and on the temporal alignment of the fundamental frequency contour with the segmental chain (House, 1996; van Santen and Möbius, 2000), and is therefore a crucial piece of information for segmental duration and intonation models.

The phonotactics of English and German allow complex consonant clusters in both the onset and the coda of syllables. The maximum number of consonants in the onset is 3 in both languages, e.g. /str/ as in *street* in English and /Str/ as in *Straße* in German. In German codas, clusters of up to 5 consonants can be observed, e.g. /mpfst/ as in *du kämpfst* 'you fight', whereas English allows up to 4 coda consonants, e.g. /ksts/ as in *texts* or /mpst/ as in *glimpsed*. Thus, the maximum number of consecutive consonants across syllable boundaries is 8 in German and 7 in English.

Disk followed

The complexity of syllable onset and coda structure poses serious problems for a syllabification algorithm because—despite restrictions as to which consonants, or classes of consonants, may occur in any given position within the onset or coda of a syllable—ambiguous and multiple alternative syllable boundary locations are usually observed in polysyllabic words, notably in compounds.

Syllable structure in English and German displays typical LNRE characteristics. It has been observed that out of the inventory of more than 12,000 distinct syllables (*syllable types*) in either language, only about 500 syllable types are systematically and regularly used in speech production. Levelt has argued that speakers have access to a mental syllabary (Levelt, 1989, 1999; Levelt and Wheeldon, 1994). A mental syllabary is an inventory of fixed syllable programs, each comprising a set of highly overlearned articulatory gestures. According to this syllabary concept, high-frequency syllables are stored as complete gestural programs that are executed during speech production, whereas low-frequency and very rare syllables are assembled online phone by phone, by using the segmental and metrical information provided by the phonological encoder. The decisive difference between holistic gestural programs and online assembly is that in the latter case the segmental spellout, which is initially underspecified and rather abstract for each segmental unit, needs to be specified to accommodate the particular context in which the segment occurs, whereas in the former case the spellout is already fully specified within the syllable domain.

Typical state-of-the-art syllabification methods can be characterized either as supervised learning of syllable structure from annotated training data or as unsupervised learning from unannotated training data. For instance, the finite-state syllabification method used in some versions of the Bell Labs TTS system (Kiraz and Möbius, 1998; Möbius, 1998b, 1999) was constructed by obtaining syllables as well as their internal structures and their frequencies of occurrence from a lexical database. Weights on the transitions between states of the transducer were derived directly from the frequencies of onset, nucleus and coda types in the database. The weights reflect the plausibility of onset, nucleus and coda types. This approach relies on the coverage of distinct syllable types by the training data. A *post hoc* hand-tuning procedure has been provided to cope with syllable types whose numbers of observations are extremely low or which do not occur in the training data at all.

An unsupervised training method on unannotated data which induces probabilistic syllable classes by means of multivariate clustering has also recently been proposed (Müller et al., 2000). This approach defines the clustering task as induction of hidden parameters of a probability model. The induction is achieved by maximum-likelihood estimation from incomplete data via the *Expectation Maximization* (EM) algorithm (Baum et al., 1970; Dempster et al., 1977). The EM algorithm is the stochastic basis of many machine learning algorithms for natural language processing. The mathematical principles of EM theory were presented

in Dempster et al. (1977), where several, non-linguistic applications were illustrated too. A linguistic application, namely error modeling in speech recognition, was demonstrated by Baker (1979), who introduced the *inside-outside* algorithm for context-free grammars and compared it to the forward-backward algorithm for hidden Markov models (Baum, 1972). Very recently a formal proof has been given that the inside-outside algorithm can be regarded as a dynamic-programming variant of the EM algorithm (Prescher, 2002)—which means that most of the probabilistic models used by computational linguists (*n*-gram models, Markov models, hidden Markov models, tree bank grammars, probabilistic context-free grammars) are in fact trained by a version of the EM algorithm.

EM-based clustering has been shown to be applicable to dyadic (two-dimensional) linguistic data, for instance to the tasks of inducing semantic labels for subcategorization slots of English and German lexical verbs as well as selecting among candidate English translations of German nouns by using clustering models on English verb-noun combinations (Rooth et al., 1998). Multidimensional data are observed in various applications of natural language processing ranging from phonology to pragmatics. In EM-based clustering for multivariate data, classes are defined as hidden data which are learned from a training corpus of data without class annotations ("incomplete data"). In such an application the main task of EM-based clustering is the automatic detection of the hidden class structure in a given corpus (Müller et al., 2000; Prescher, 2002).

The new multidimensional EM-based clustering method (Müller et al., 2000) was applied to syllable structure, modeling either three dimensions (onset, nucleus, coda) or five dimensions (position of the syllable in the word and syllabic stress, additionally). *Soft* clustering is performed in the syllabification task: a given syllable type consisting of onset, nucleus and coda may be a member of more than one class; a probability is assigned to each individual class membership.

A very important property of probabilistic models is their ability to cope with unknown data or, in other words, their ability to assign probabilities to data unseen in the training corpus. In the syllabification task the EM-based clustering will assign a positive probability to every possible syllable, even if it does not actually occur in the training data. What constitutes a possible syllable in a given language can be described by, e.g., a probabilistic context-free grammar (Müller et al., 2000). Another useful property of the EM algorithm is that estimated parameters are available for inspection after each iteration; it is therefore both reasonable and easy to perform a systematic *linguistic* evaluation of estimated parameter values externally, in addition to the internal quantitative evaluation in terms of log-likelihood-based stopping criteria (Prescher, 2002). The qualitative assessment and interpretation of obtained syllable classes in Müller et al. (2000) is an example of such an external linguistic evaluation. For instance, certain syllable types with high probabilities were found that systematically occur in high-frequency function words. High-frequency syllable types are candidates for entries in the syllabary, i.e. in the set of highly overlearned articulatory programs executed during speech production.

### 2.3. *Duration Modeling*

Among the most important factors that have an effect on the duration of speech sounds are, in many languages, the identity of the speech sound; its immediate segmental context; its position in the syllable, word and phrase; and prosodic factors such as syllabic stress and the accent status of the word. Some of these factors have only two values; for instance, the factor "syllabic stress" may be either "stressed" or "unstressed". Other factors can assume a much larger number of values; for example, the immediate segmental context has as many values as there are phones that can occur adjacent to the speech sound in question. The cross-product of all factor values gives the total number of combinatorially possible constellations in a given language.

The task of the duration component in a TTS system is to predict the temporal structure of synthetic speech from symbolic input. This is usually achieved by assigning a duration to each speech sound in the utterance at synthesis runtime, based on the particular factorial constellation in which the speech sound occurs. The symbolic descriptor of the factorial constellation, i.e. the current values of all factors, is often called a *feature vector*. The number of combinatorially possible feature vectors is usually in the tens of thousands, as has been shown for English and German (van Santen, 1995; Möbius and van Santen, 1996), and at least 17,500 distinct feature vectors have been actually observed in American English (van Santen, 1993b).

Duration models differ in terms of how they use the information expressed in the feature vector to assign the segmental duration. A widely used type of duration

model is a sequential rule system such as the one proposed by Klatt (1973). Starting from some intrinsic value, the duration of a segment is modified by successively applied rules, which are intended to reflect contextual, positional and prosodic factors that have a lengthening or shortening effect. When large speech databases and the computational means for analyzing these data became available, new approaches were proposed based on, for example, Classification and Regression Trees (CART) (Pitrelli and 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 factorial constellations or feature vectors (van Santen, 1994). Manual database construction, on the other hand, is not practical because of the size of the factorial space.

Most observed feature vectors have a very low frequency of occurrence. Durational feature vectors thus belong to the LNRE class of distributions. It would be misguided, however, to accept poor modeling of the rare vectors or to ignore them altogether. The reason is that the cumulative frequency of rare vectors all but guarantees the occurrence of at least one unseen vector in any given sentence. In an analysis for English, van Santen (1995) computed a probability of more than 95% that a randomly selected 50-phoneme sentence contains a vector that occurs at most once in a million segments.

Therefore, the duration model has to be capable of predicting, by some form of extrapolation from observed feature vectors, durations for vectors that are insufficiently represented in the training material. CART-based methods and other general-purpose prediction systems are known for coping poorly with sparse training data and, most seriously, with missing feature vector types because they lack this extrapolation capability. Extrapolation is further complicated by interactions between the factors.

Factor interactions also 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 sums-of-products method (van Santen, 1993a, 1994) 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 that of 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.

Building a sums-of-products duration model requires large annotated speech corpora, sophisticated statistical tools, and the type of linguistic and phonetic knowledge that is incorporated in traditional rule systems. The approach uses statistical techniques that can cope with the problem of confounding factors and, most importantly, with data sparsity caused by the LNRE frequency distributions of durational feature vectors.

van Santen's method has been applied to a number of languages including American English (van Santen, 1993b, 1994), German (Möbius and van Santen, 1996), Mandarin Chinese (Shih and Ao, 1997), and Japanese (Venditti and van Santen, 1998).

### *2.4. Concatenative Speech Synthesis*

Corpus-based approaches to speech synthesis have been advocated to overcome the limitations of concatenative synthesis from a fixed acoustic unit inventory. The frequency of unit concatenations in diphone synthesis, viz. one concatenation point per phone, has been argued to contribute to the perceived lack of naturalness of synthetic speech. The key idea of corpus-based synthesis is to use an entire speech corpus as the acoustic inventory and to select at run-time from this corpus the longest available strings of phonetic segments that match a sequence of target speech sounds in the utterance to be synthesized, thereby minimizing the number of concatenations and reducing the need for signal processing.

In an ideal world, the target utterance would be found in its entirety in the speech database and simply played back by the system without any concatenations and without any signal processing applied, effectively rendering natural speech. Given the complexity and combinatorics of language and speech, this ideal case is extremely unlikely to occur in unrestricted application domains. However, given a speech database of several hours worth of recordings, chances are that a target utterance may be produced by a small number of units each of which is considerably longer than a classical diphone.

Defining the optimal speech database for unit selection has become one of the most important research issues in speech synthesis. A well-designed speech

corpus has a huge impact on the quality of the synthesized speech, no matter what the basic unit is defined to be, a phone, a demiphone, a diphone, or even a triphone. It is now generally accepted that to be able to benefit from long acoustic units, a meticulous design of the text materials to be recorded is required. The database should be designed or constructed such as to include all relevant acoustic realizations of phonemes, a point made already by Iwahashi and Sagisaka (1995).

There are hardly any systematic studies of coverage in the area of speech synthesis, with the exception of van Santen (1997), and the results from this study are quite discouraging. For example, van Santen constructed a contextual feature vector for diphone units that included key prosodic factors such as word accent status and position in the utterance. He then computed the *coverage index* of training sets, which is defined as the probability that all diphone-vector combinations occurring in a randomly selected test sentence are also represented in the training set. It turned out that a training set of 25,000 combinations had a coverage index of 0.03, which means that the probability is 0.03 that the training set covers all combinations occurring in the test sentence. To reach a coverage index of 0.75 a training set of more than 150,000 combinations is required. Given that the factors used for the feature vector were coarse and few, unit selection approaches based on diphone units would require absurdly large speech databases to achieve reasonable coverage. These findings shed an unfavorable light on corpus-based speech synthesis approaches that attempt to cover an unrestricted domain—typically, the whole language—by simply re-sequencing recorded speech.

If it is practically impossible to construct an optimal speech database, what are the requirements of a corpus if approximate coverage is the goal? The answer, again, is tentative, and pessimistic. Evidently, LNRE distributions also play a crucial role in data-driven concatenative speech synthesis. To illustrate this point, let us consider two case studies.

*Case 1*. Beutnagel and Conkie (1999) report that more than 300 diphones out of a complete set of approximately 2,000 diphones, which serve as the core acoustic unit inventory in the demiphone-based AT&T TTS system, occur only once in a two-hour database recorded for unit selection. These rare diphones were actually included in the database only by way of embedding them in carefully constructed sentences; evidently, they were not expected to occur naturally in the recorded speech at all. The authors observe that the unit selection algorithm prefers these rare diphones for target sentences, instead of concatenating them from the smaller demiphone units. The interpretation offered by the authors is that the preferred selection of these diphones by the selection algorithm will likely generate superior synthesis quality compared to the demiphone solution (Beutnagel and Conkie, 1999).

*Case 2*. For the construction of the database for a new Japanese synthesis system (Tanaka et al., 1999) "multi-form units" were collected that were intended to cover all Japanese consonant-vowel (CV) syllables and all possible CVk chains. CVk chains are defined by the authors as sequences consisting of a consonant followed by any number ($k$) of vowels, semivowels and (tauto-)syllabic nasals. The units were realized by the speaker in a variety of prosodic contexts. About 41 million multi-form units were collected this way, yielding 100,000 distinct CVk units. Experiments showed that the 50,000 most frequent multi-form units cover approximately 75% of Japanese text. Given the relatively simple syllable structure of Japanese, the emphasis should probably be on *only* 75% coverage. On the other hand, Japanese allows very long sequences of vocalic speech sounds, which makes a complete coverage of such sequences virtually impossible. Note that in conjunction with another set of 10,000 diphone units, the multi-form unit database accounts for as much as 6.3 hours of speech. Increasing the unit inventory to 80,000 does not result in a significantly higher coverage, and the growth curve appears to converge to about 80% (Tanaka et al., 1999, Fig. 2). The authors state that for unrestricted text the actually required number of units approaches infinity, and that most units are rarely used—a characteristic of LNRE distributions. The question of how to get to near 100% coverage remains unanswered, in fact even unasked.

In the Laureate system (Breen and Jackson, 1998) an attempt is made to optimize the speech database based on linguistic criteria. The result is a database that contains at least one instance of each diphone in the language. This baseline inventory is augmented by embedding the diphones not in carrier phrases but in phonetically rich text passages. This self-restrained optimization attempt is a consequence of the fact that annotation and quality control are considered to be too unreliable for larger databases. The authors argue that it is also difficult to ensure a consistent speaking style in a large set of recordings and that speech segments from very different styles will result in a patchwork of

concatenated speech. Speaking style itself is currently not considered to be a useful selection criterion.

Established techniques from speech recognition have been applied to cope with the LNRE problem. For instance, in the now classical unit selection algorithm presented in Hunt and Black (1996), each unit in the database is represented by a state in a state transition network, where the state occupancy costs are given by the measure of unit distortion and the state transition costs are given by the measure of continuity distortion. This design is somewhat reminiscent of hidden Markov model (HMM) based speech recognition systems. The key difference is in the use of cost functions in the unit selection framework as opposed to the probabilistic models used in speech recognition. Using a similar framework, Holzapfel and Campbell (1999) attempt to enhance generalization to unseen cases in runtime unit selection. They train a set of triphone HMM's on the speech database to assess the similarity of segmental contexts. All contexts of each phone are first pooled; the pools are then iteratively split according to phonetically motivated criteria, with a maximum likelihood criterion ensuring optimal improvement of the models with every split of a cluster. By classifying the contexts according to the criteria learned by the clustering tree, triphone contexts that do not occur in the database and were unseen during training can be reconstructed and mapped appropriately, a standard procedure in speech recognition (Jelinek and Mercer, 1980; Young, 1992). A similar approach was implemented in Microsoft's TTS system (Huang et al., 1996; Hon et al., 1998).

The key idea of the *context clustering* method in speech synthesis (Nakajima and Hamada, 1988; Nakajima, 1994) is to cluster into equivalence classes all realizations of phonemes that are found in a single-speaker database. Equivalence classes are defined by segmental phonetic context. Clustering is performed by decision trees that are constructed automatically such that they maximize the acoustic similarity within each equivalence class. Each leaf in the tree is represented by a segment ("allophone") and its features, as extracted from the database. One advantage of this method is that it automatically determines the relative importance of different contextual and coarticulatory effects. Through interpolation even context specifications that were not seen during training can be met. A modified version of the clustering method has been implemented in the English speech synthesizer developed at Cambridge University (Donovan and Woodland, 1999) and in the IBM speech synthesizer (Donovan and Eide, 1998).

Some unit selection systems apply a threshold to either continuity or target costs or both, the reasoning being that units exceeding the thresholds either are not good representatives of the unit target or will not concatenate smoothly with adjacent units. In some cases there will be no unit candidates below the cost threshold. In the case of continuity distortions, a *backing-off* strategy is sometimes applied (Donovan and Eide, 1998): if there is no unit in the current cluster that concatenates smoothly with any unit in the subsequent cluster, then a new continuity cost is computed for all units available at the parent nodes of the two clusters in the decision tree. This process is applied iteratively until a pair of units is found that concatenates sufficiently smoothly. It is still possible that no appropriate pair of units is found that connects smoothly with the rest of the unit sequence; in this case, a discontinuity is unavoidable.

Note that in this and related work (Donovan and Woodland, 1999), the basic synthesis units are of a size that corresponds to an HMM state, i.e. they are subphone units, whereas in most other unit selection systems the basic units are phone-sized (or sometimes demiphones). HMM-based clustering has the advantage, inherited from speech recognition, that the most appropriate clustered state can be reached in any new context encountered at synthesis time, even in contexts that were not seen during training. Still, as the authors concede, less frequently seen contexts will be modeled less well, and if the discrepancy between contexts required during synthesis and anything seen in the training database is too large, serious continuity problems will occur (and have indeed been observed). The authors conclude that some degree of database construction is required to ensure a reasonable coverage of possible contexts.

## 3. Closed Domains

The coverage problem encountered in the context of corpus-based speech synthesis for unrestricted application domains is evidently due to the complexity and combinatorics of language and speech. In contrast, the distributions of linguistic and phonetic factors in restricted domains are known in advance. It has often been suggested therefore that in such a scenario a version of the unit selection synthesis strategy might be feasible that exploits basic units larger than demiphones, phones, or diphones.

For instance, in the most recent version of the synthesis component developed in the Verbmobil project (Wahlster, 2000), a word concatenation approach has been implemented (Stöber et al., 1999). The Verbmobil domain comprises a fixed vocabulary of about 10,000 words from the travel planning domain. Each word in the domain's lexicon was recorded in a variety of prosodic and positional contexts. The only signal processing step applied was a simple amplitude smoothing on all adjacent words that do not co-occur in the database.

Unfortunately, the Verbmobil domain is not entirely closed. Its lexicon has a loophole that allows proper names to sneak into the domain. To synthesize these names, and novel words in general, the system resorts to diphone synthesis. This strategy is not altogether satisfactory because the quality difference between phrases generated by word concatenation and the high-entropy novel words synthesized from diphones is too striking. To extend and generalize their approach to unrestricted domains, the authors propose to develop rules that enable the system to compose missing syllabic units from phoneme realizations and whole words from syllables (Stöber et al., 1999). For this approach to be feasible, however, phonemes and syllables will need to be available whose contexts are not restricted by a fixed domain-specific lexicon. Thus, we are back at square one: the need to design a speech database with optimal coverage for open-domain synthesis.

A system based on word and syllable concatenation has also been presented for the limited domain of weather forecasting (Lewis and Tatham, 1999). The system has an inventory of 2,000 recorded monosyllabic and polysyllabic words. There are numerous problems with this approach. For instance, monosyllables are embedded in a fixed-context carrier phrase during recordings, making them almost automatically inappropriate for recombination. Also, some of the recombination rules appear to be of an *ad hoc* nature, such as to cut three periods from the start or end of syllables whose onsets or codas are periodic. The authors admit that such rules will probably have to be modified for other voices or recording rates. These problems notwithstanding, the authors are confident that their synthesis strategy can be extended to much larger databases and to unrestricted TTS scenarios. In the light of the depressing results of van Santen's (1997) study on the coverage index of training databases for unit selection synthesis, I am led to believe that their optimism is unwarranted.

In an attempt to narrow the discrepancy between the speech database and the sentences to be synthesized, Black and Lenzo (2000) opted for designing corpora specifically for each target application domain. Considering that such applications often involve a dialog system that generates the spoken language output, the authors emphasize the need for a set of prompt-style sentences that occur frequently and cover the domain adequately. A backup method, which is basically a phone-based standard unit selection, is provided for the less frequent out-of-domain cases. When evaluated in the context of the CMU DARPA Communicator flight information system, only 2.5% of all synthesized phrases turned out to contain out-of-domain words, comprising only 75 distinct out-of-vocabulary words, all of them in fact place names. Even though the authors do not recommend this system for general-purpose synthesis, they have demonstrated that reliable high-quality synthetic voices can be built for limited application domains.

In an extension of this work the question was asked how to find the optimal set of utterances to be recorded to cover both a relatively restricted domain and completely open domains (Black and Lenzo, 2001). In a first step, an acoustic distance measure that is both database-specific and speaker-specific was established to optimize unit clustering. Each resulting cluster represents an acoustically distinct unit. This information was then exploited to establish unit type frequency and contexts in a large database and thus the coverage of units by the database. In a computationally very expensive, iterative, greedy-like procedure a minimal set of sentences was extracted from a text corpus such that it has the same coverage as the entire corpus. Starting from a corpus of 1.2 million words, this method yielded a list of 241 sentences, which were pruned to 221 by removing some unacceptably weird sentences. These sentences were then recorded; in informal listening tests the synthetic voice built from these recordings received the best scores for test sentences taken from the original training set. Somewhat lower scores were obtained for test sentences from different domains, which underlines the importance of the adequacy of a database for the target domain.

Two important decisions were made in these experiments: (1) finding an acoustic distance measure that helps decide which acoustic distinctions need to be made and which ones can be ignored; (2) the use of frequency information to prune rare cases.

The first of these decisions is very valuable, albeit somewhat ambiguous with respect to an interpretation of its consequences. The distance measure is speaker and domain dependent. This means that even within a given domain, the ranking or weighting of the criteria that define the distance measure will have to be re-estimated for every new synthetic voice. Thus, while this approach optimizes the set of recordings for a given speaker, it also significantly impedes the process of building voices because different textual materials will have to be constructed for each speaker.

Moreover, establishing the relationship between computed ("objective") distances and perceptual differences is a difficult task, and the body of research on this topic is quite small and mainly focused on speech coding (Quackenbush et al., 1988). In early unit selection experiments (Black and Campbell, 1995) the mean Euclidean cepstral distance between the feature vectors of the target unit and those of the candidate units in the database was used as a score for the set of weights. However, the cepstral distance measure appeared to give higher priority to unit distortion, often at the expense of continuity distortion, whereas human listeners seem to prefer smoother transitions at the concatenation points.

Some insight into the usability of objective distance measures as predictors of perceptual differences in unit selection was provided by Wouters and Macon (1998) and Macon et al. (1998). They attempted to find measures that best predict phonetic variations in the realizations of phonemes. These measures are intended to reflect specific phonetic changes instead of overall quality of distorted (coded) speech and to quantify the distance between two candidate units. Some of the best-known measures such as mel-based cepstral distance and the Itakura-Saito distance were found to be quite useful, yielding a moderate correlation ($r = 0.66$) with perceptual distances. The authors feel, however, that this strength of correlation is still not sufficient for objective distance measures to be reliable predictors of perceptual differences.

The second decision in the Black and Lenzo (2001) study, viz. the pruning of units occurring in rare constellations, may be promising to some extent in limited domains whose coverage and distributional properties are well-known; but it is much less likely to work in open-domain synthesis for the very reasons that are discussed throughout the present paper: in short, the large number of rare events.

## 4. Conclusion

The LNRE characteristics of language and speech are often unrecognized and the pertinent problems underestimated. For example, it is a common attitude to accept poor modeling of less frequently seen or unseen contexts because "they are less frequently used in synthesis" (Donovan and Woodland, 1999, p. 228). The perverse nature of LNRE distributions is the following: the number of rare events is so large that the probability of encountering at least one of these events in a particular sample approaches certainty. It is the rare vectors and combinations that are poorly modeled, and one or the other of these rare events will show up when utterances are synthesized, just as predicted by the LNRE distribution models.

In this paper I have discussed challenges by LNRE properties to four components of a TTS system: morphological analysis, syllabification, segmental duration modeling, and acoustic inventory design. In the context of lexical and morphological analysis I have argued that a TTS system should be equipped with a component that performs an adequate analysis of unknown words, yielding an annotation of the internal structure of such words that is sufficient to drive general-purpose pronunciation rules. The unknown word analysis component implemented in the Bell Labs German TTS system (Möbius, 1999) relies on a grammar of the structure of morphologically complex words and incorporates results from a study on the productivity of word formation processes. Further improvements may be expected from a morphology system that, besides derivation and compounding analysis (and generation) capabilities, also computes the hierarchical structure of complex words. Such a system would apply sophisticated statistical models, capable of dealing with LNRE properties, to the quantitative analysis of morphological productivity (Lüdeling et al., 2000; Evert and Lüdeling, 2001).

A probabilistic approach to syllabification (Müller et al., 2000) has been discussed that offers a reasonable solution to the LNRE properties of syllable type frequency distributions. The advantage of this multidimensional EM-based clustering method is that the induced models assign probabilities even to syllable types that are not covered by the training database. Moreover, ranking the obtained syllable types by their frequencies may even provide a quantitative basis for deciding which syllables can be considered as potential entries in the mental syllabary, a concept discussed in current

work on speech production (Levelt and Wheeldon, 1994; Levelt, 1999).

In the context of modeling segmental durations I concluded, following van Santen (1995), that rare feature vectors cannot be ignored, because the cumulative frequency of rare vectors all but guarantees the occurrence of at least one unseen vector in any given sentence. The duration model therefore has to be able to predict durations for vectors that are insufficiently, or not at all, represented in the training material. The well-established solution to this problem is the application of a class of arithmetic models known as sums-of-products models (van Santen, 1993a). These models have been shown by van Santen and his colleagues to cope well with the problem of confounding factors and with data sparsity caused by the LNRE frequency distributions of durational feature vectors.

No concrete solution has been offered for the coverage problems encountered in the context of corpus-based speech synthesis. The uneven performance that characterizes unit selection based speech synthesis systems can be attributed, to a large extent, to the complexity and combinatorics of language and speech in general, and to LNRE properties in particular. Methods well-known from speech recognition, such as context clustering for covered units and context reconstruction for missing ones, have been adopted to cope with the LNRE problem in concatenative speech synthesis, but the problem itself is not removed. I believe that the most promising avenue of research is to increase the coverage of speech databases by carefully defining the linguistic and phonetic criteria that the database should meet, and to complement this line of research by further systematic studies of the correlations between objective distance measures and perceived differences.

The design of databases for restricted application domains, where the distributions of linguistic and phonetic factors are known, is a reasonable step in this direction. The relative success of Black and Lenzo's (2001) approach, which yields appropriate coverage for limited domains and fair quality for open-domain synthesis, seems to support this conclusion. But there is a caveat: I have tried to point out the difference between, on the one hand, a strictly closed domain with a fixed vocabulary and, on the other hand, a merely restricted domain with loopholes that may require a mix of synthesis strategies, possibly resulting in very uneven speech output quality.

## Acknowledgments

I am grateful to the people who I had the pleasure to work with on various research issues pertinent to this paper; in particular, Ulrich Heid, George Kiraz, Anke Lüdeling, Karin Müller, Detlef Prescher, Bettina Säuberlich, Tanja Schmid, and Jan van Santen. I also wish to thank Stefan Evert for his advice on statistical matters, and the three anonymous reviewers for their very constructive and thoughtful comments and suggestions.

## References

Baayen, H. (1993). On frequency, transparency and productivity. In G. Booij and J. van Marle (Eds.), *Yearbook of Morphology 1992*. Dordrecht: Kluwer, pp. 181–208.

Baayen, H. (2001). *Word Frequency Distributions*. Dordrecht: Kluwer.

Baayen, H. and Lieber, R. (1991). Productivity and English derivation: A corpus based study. *Linguistics*, *29*:801–843.

Baker, J.K. (1979). Trainable grammars for speech recognition. In D. Klatt and J. Wolf (Eds.), *Speech Communication Papers for ASA'79*, pp. 547–550.

Baum, L.E. (1972). An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. *Inequalities*, *3*:1–8.

Baum, L., Petrie, T., Soules, G., and Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics*, *41*(1):164–171.

Beutnagel, M. and Conkie, A. (1999). Interaction of units in a unit selection database. *Proceedings of the European Conference on Speech Communication and Technology*. Budapest, Hungary, vol. 3, pp. 1063–1066.

Black, A.W. and Campbell, W.N. (1995). Optimising selection of units from speech databases for concatenative synthesis. *Proceedings of the European Conference on Speech Communication and Technology*. Madrid, Spain, vol. 1, pp. 581–584.

Black, A.W. and Lenzo, K.A. (2000). Limited domain synthesis. *Proceedings of the International Conference on Spoken Language Processing*. Beijing, vol. 2, pp. 411–414.

Black, A.W. and Lenzo, K.A. (2001). Optimal data selection for unit selection synthesis. *Proceedings of the 4th ISCA Tutorial and Research Workshop on Speech Synthesis*. Pitlochry, UK, pp. 63–68.

Breen, A.P. and Jackson, P. (1998). Non-uniform unit selection and the similarity metric within BT's Laureate TTS system. *Proceedings of the Third International Workshop on Speech Synthesis*. Jenolan Caves, Australia, pp. 373–376.

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

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, *39*:1–38.

Donovan, R.E. and Eide, E.M. (1998). The IBM trainable speech synthesis system. *Proceedings of the International Conference on Spoken Language Processing*. Sydney, Australia, vol. 5, pp. 1703–1706.

Donovan, R.E. and Woodland, P.C. (1999). A hidden Markov-model-based trainable speech synthesizer. *Computer Speech and Language*, *13*:223–241.

Evert, S. and Lüdeling, A. (2001). Measuring morphological productivity: Is automatic preprocessing sufficient? In P. Rayson, A. Wilson, T. McEnery, A. Hardie, and S. Khoja (Eds.), *Proceedings of the Corpus Linguistics 2001 Conference*. Lancaster, UK, pp. 167–175.

Good, I.J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika*, *40*(3/4):237–264.

Holzapfel, M. and Campbell, N. (1998). A nonlinear unit selection strategy for concatenative speech synthesis based on syllable level features. *Proceedings of the International Conference on Spoken Language Processing*. Sydney, Australia, vol. 6, pp. 2755–2758.

Hon, H.W., Acero, A., Huang, X., Liu, J., and Plumpe, M. (1998). Automatic generation of synthesis units for trainable text-to-speech systems. *Proceedings of the IEEE International Conference on Acoustics and Speech Signal Processing*. Seattle, WA, vol. 1, pp. 293–296.

House, D. (1996). Differential perception of tonal contours through the syllable. *Proceedings of the International Conference on Spoken Language Processing*. Philadelphia, PA, vol. 1, pp. 2048–2051.

Huang, X., Acero, A., Adcock, J., Hon, H.W., Goldsmith, J., Liu, J., and Plumpe, M. (1996). Whistler: A trainable text-to-speech system. *Proceedings of the International Conference on Spoken Language Processing*. Philadelphia, PA, vol. 4, pp. 2387–2390.

Hunt, A.J. and Black, A.W. (1996). Unit selection in a concatenative speech synthesis system using a large speech database. *Proceedings of the IEEE International Conference on Acoustics and Speech Signal Processing*. München, Germany, vol. 1, pp. 373–376.

Iwahashi, N. and Sagisaka, Y. (1995). Speech segment network approach for an optimal synthesis unit set. *Computer Speech and Language*, *9*:335–352.

Jelinek, F. and Mercer, R.L. (1980). Interpolated estimation of Markov source parameters from sparse data. *Proceedings of the Workshop on Pattern Recognition in Practice*. Amsterdam, pp. 381–397.

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

Katz, S.M. (1987). Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Transactions on Acoustics, Speech and Signal Processing*, *33*(6):400–401.

Khmaladze, E. (1987). The statistical analysis of large number of rare events (Tech. Report MS-R8804). Department of Mathematical Statistics, CWI. Amsterdam: Center for Mathematics and Computer Science.

Kiraz, G.A. and Möbius, B. (1998). Multilingual syllabification using weighted finite-state transducers. *Proceedings of the Third International Workshop on Speech Synthesis*. Jenolan Caves, Australia, pp. 71–76.

Klatt, D.H. (1973). Interaction between two factors that influence vowel duration. *Journal of the Acoustical Society of America*, *54*(4):1102–1104.

Levelt, W.J.M. (1989). *Speaking: From Intention to Articulation*. Cambridge, MA: MIT Press.

Levelt, W.J.M. (1999). Producing spoken language: A blueprint of the speaker. In C.M. Brown and P. Hagoort (Eds.), *The Neurocognition of Language*. Oxford, UK: Oxford University Press, pp. 83–122.

Levelt, W.J.M. and Wheeldon, L. (1994). Do speakers have access to a mental syllabary? *Cognition*, *50*:239–269.

Lewis, E. and Tatham, M. (1999). Word and syllable concatenation in text-to-speech synthesis. *Proceedings of the European Conference on Speech Communication and Technology*. Budapest, Hungary, vol. 2, pp. 615–618.

Lüdeling, A., Evert, S., and Heid, U. (2000). On measuring morphological productivity. *Proceedings of KONVENS 2000*. Ilmenau, Germany, pp. 57–61.

Macon, M.W., Cronk, A.E., and Wouters, J. (1998). Generalization and discrimination in tree-structured unit selection. *Proceedings of the Third International Workshop on Speech Synthesis*. Jenolan Caves, Australia, pp. 195–200.

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. (1998a). In R. Sproat (Ed.), *Multilingual Text-to-Speech Synthesis: The Bell Labs Approach*. Dordrecht: Kluwer, Chs. 3, 6, and 7.

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

Möbius, B. (1999). The Bell Labs German text-to-speech system. *Computer Speech and Language*, *13*:319–358.

Möbius, B. (2001). German and multilingual speech synthesis. *Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung (Univ. Stuttgart), AIMS*, *7*(4):1–300.

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

Müller, K., Möbius, B., and Prescher, D. (2000). Inducing probabilistic syllable classes using multivariate clustering. *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*. Hong Kong, pp. 225–232.

Nakajima, S. (1994). Automatic synthesis unit generation for English speech synthesis based on multi-layered context oriented clustering. *Speech Communication*, *14*:313–324.

Nakajima, S. and Hamada, H. (1988). Automatic generation of synthesis units based on context oriented clustering. *Proceedings of the IEEE International Conference on Acoustics and Speech Signal Processing*. New York, NY, pp. 659–662.

Pitrelli, J.F. and 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.

Prescher, D. (2002). EM-basierte maschinelle Lernverfahren für natürliche Sprachen. *Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung (Univ. Stuttgart), AIMS*, *8*(2):1–366.

Quackenbush, S.R., Barnwell, T.P., and Clements, M.A. (1988). *Objective Measures of Speech Quality*. Englewood Cliffs, NJ: Prentice Hall.

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

Rooth, M., Riezler, S., Prescher, D., Carroll, G., and Beil, F. (1998). EM-based clustering for NLP applications. *Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung (Univ. Stuttgart), AIMS*, *4*(3):97–128.

Samuelsson, C. (1996). Relating Turing's formula and Zipf's law. *Proceedings of the 4th Workshop on Very Large Corpora*. Copenhagen, Denmark.

Schmid, T., Lüdeling, A., Säuberlich, B., Heid, U., and Möbius, B. (2001). DeKo: Ein System zur Analyse komplexer Wörter. In H. Lobin (Ed.), *Proceedings of GLDV-2001*. Gießen, Germany, pp. 49–57.

Schultink, H. (1961). Produktiviteit als morfologisch fenomeen. *Forum der Letteren*, *2*:110–125.

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

Stöber, K., Portele, T., Wagner, P., and Hess, W. (1999). Synthesis by word concatenation. *Proceedings of the European Conference on Speech Communication and Technology*. Budapest, Hungary, vol. 2, pp. 619–622.

Tanaka, K., Mizuno, H., Abe, M., and Nakajima, S. (1999). A Japanese text-to-speech system based on multi-form units with consideration of frequency distribution in Japanese. *Proceedings of the European Conference on Speech Communication and Technology*. Budapest, Hungary, vol. 2, pp. 839–842.

van Santen, J.P.H. (1993a). Exploring *N*-way tables with sums-of-products models. *Journal of Mathematical Psychology*, *37*(3):327–371.

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

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. (1995). Computation of timing in text-to-speech synthesis. In W.B. Kleijn and K.K. Paliwal (Eds.), *Speech Coding and Synthesis*. Amsterdam: Elsevier, pp. 663–684.

van Santen, J.P.H. (1997). Combinatorial issues in text-to-speech synthesis. *Proceedings of the European Conference on Speech Communication and Technology*. Rhodes, Greece, vol. 5, pp. 2511–2514.

van Santen, J.P.H. and Möbius, B. (2000). A quantitative model of F0 generation and alignment. In A. Botinis (Ed.), *Intonation—Analysis, Modelling and Technology*. Dordrecht: Kluwer, pp. 269–288.

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

Wahlster, W. (Ed.) (2000). *Verbmobil: Foundations of Speech-to-Speech Translation*. Berlin: Springer.

Wouters, J. and Macon, M.W. (1998). A perceptual evaluation of distance measures for concatenative speech synthesis. *Proceedings of the International Conference on Spoken Language Processing*. Sydney, Australia, vol. 6, pp. 2747–2750.

Young, S. (1992). The general use of tying in phoneme-based HMM speech recognisers. *Proceedings of the IEEE International Conference on Acoustics and Speech Signal Processing*. San Francisco, CA, vol. 1, pp. 569–572.

Zipf, G.K. (1935). *The Psycho-Biology of Language*. Boston, MA: Houghton Mifflin.

Zipf, G.K. (1949). *Human Behavior and the Principle of Least Effort—An Introduction to Human Ecology*. New York: Hafner.