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

Bernd Möbius

Institute of Natural Language Processing,
University of 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 these problems in the contexts of morphology, syllabification, segmental duration and unit selection, and also suggests possible solutions. The design of databases for restricted application domains, where the distributions of linguistic and phonetic factors are known, is also critically reviewed.

1. Introduction

In this paper we intend to point out two common concepts in speech synthesis that we 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 extremely uneven 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). Heavily skewed frequency distributions are also observed in segmental duration modeling, where the majority of relevant feature vectors is sparsely or not at all represented in training databases (section 2.3). The fourth area in TTS conversion that is affected by imbalanced frequency distributions is the design of acoustic unit inventories for data-driven speech synthesis (section 2.4).

The second concept that we consider questionable is the notion of a "restricted" application domain (section 3). We suggest that word or syllable concatenation schemes are only feasible in strictly closed domains, i.e. those domains that have a fixed and unchanging vocabulary.

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*. 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, the vast majority of the class members is extremely rare. In our work on German and multilingual speech synthesis [1, 2] we 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 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, in particular in German but more generally in almost any natural language.

The work of Harald Baayen [3] reveals that monomorphemic content words, viz. nouns, adjectives and verbs, are outside the LNRE zone, but that word frequencies of affixes, for instance, which are the main means of derivation, have prototypical LNRE distributions. The LNRE zone, according to Baayen, is the range of sample sizes where one keeps finding previously unseen words, no matter how large the sample size is. 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).

Syllable type frequency distributions in languages with complex *syllable structure*, such as English or German, also display typical LNRE characteristics. A few hundred syllable types account for the majority of realized syllable tokens in speech production, whereas the vast majority of syllable types are very rarely used. Preferred approaches to syllabification are therefore those that can assign probabilities to underrepresented or even unseen syllable types (section 2.2).

Similarly unpleasant frequency distributions are observed in *segmental duration* modeling (section 2.3). The factors and features that have an effect on the duration of speech sounds jointly define a large feature space; for English and German tens of thousands of distinct feature vectors exist [4, 5]. Durational feature vectors belong to the LNRE class of distributions: the majority of observed feature vectors has 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. But diphone-based systems using a pre-defined unit set may be affected as well.

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 [6]. 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.

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 [7]. Baayen's approach exploits the observation that the proportion of hapax legomena in a text database is much higher for intuitively productive affixes than for unproductive ones. Hapax legomena are here defined relative to a text corpus. Given a particular morpheme, all word types in the corpus that are formed by this morpheme are listed and their 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 a morpheme can be expressed as the ratio of hapax legomena ($n1$) to the total number of tokens containing that morpheme in the database (N): $P = n1/N$.

An analysis component for morphologically complex unknown words (and names) that incorporates Baayen's approach has been integrated into the linguistic text analysis of the Bell Labs German TTS system [1]. This 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 [8]. Thus, the TTS system has the capability to morphologically decompose unknown words 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 slope of the vocabulary growth curve pertaining to a given morphological process. The vocabulary is defined as the number of types (or lemmata) that the process can generate. A truly productive pattern may be characterized by an infinite vocabulary, whereas an unproductive pattern may be expected to have a finite, and often quite small, vocabulary [9]. Based on a given text corpus, we obtain the vocabulary growth curve of a morphological process by plotting the number of distinct types observed as we increase the number of tokens formed by the process (Figure 1). The vocabulary growth curve of an unproductive process will flatten out and converge to a constant value after enough data has been sampled. The vocabulary of a productive pattern will continue to grow indefinitely.

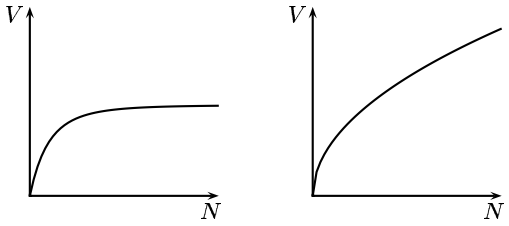

Figure 1: Typical shapes of vocabulary growth curves (V = types, N = tokens): the curve pertaining to an unproductive pattern will flatten out (left panel), whereas the vocabulary of a productive pattern will continue to grow indefinitely (right panel). Adapted from [9].

More recently, Baayen has developed much more elaborate statistical methods for estimating word frequency distributions and morphological productivity [3] and, more generally, for coping with the extremely uneven LNRE distributions of word frequencies. One important conclusion from this work is that the vocabulary growth curve, and therefore also the productivity index (P), is a function of the sample size; in other words, it is hard, if not impossible, to compare the productivity of two morphological processes with substantially differing sample sizes. Another relevant implication is that the text corpora used in the earlier studies [10, 8] 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, 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 (DeKo, [11]) 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 [12, 9]. It was further shown that only manual clean-up and correction will yield reliable input to the models. Unfortunately, manual preprocessing is not feasible for corpora of the required size, and automatic procedures, while yielding some improvement over the uncorrected data, are not sufficiently reliable [12, 9].

Figure 2 displays raw and manually corrected vocabulary growth curves for the German adjective-forming suffixes *-bar* and *-sam*. Only the corrected curves reflect the expected characteristics: *-bar* is intuitively productive, whereas *-sam* is intuitively unproductive. The raw curves suggest that the two morphological patterns have very similar productivity rates.

To make matters worse, automatic preprocessing is not even reliable as a basis for further manual correction because it has been shown to produce misleading results in many cases [9]. We 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.

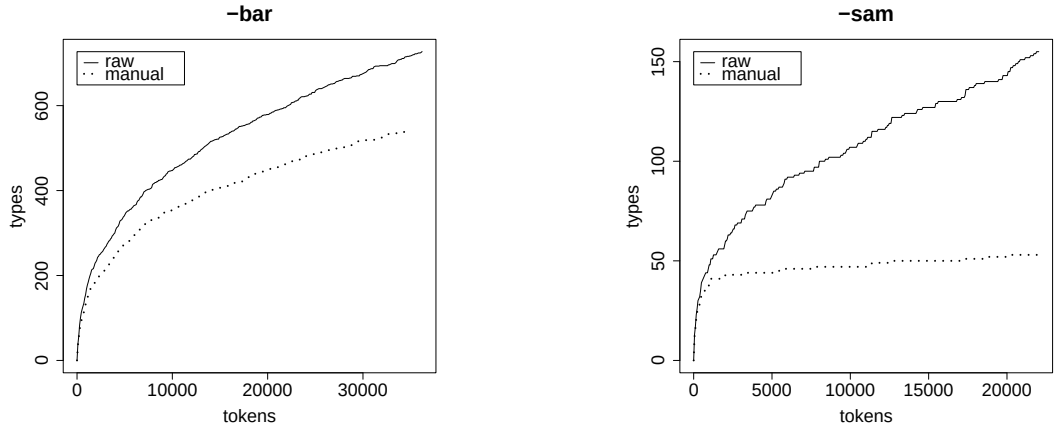

Figure 2: Vocabulary growth curves of the German adjective-forming suffixes *-bar* and *-sam*. The raw curves (continuous lines) suggest that the two morphological patterns have very similar productivity rates. Only after manual correction (dotted lines) do the curves reflect the expected characteristics: *-bar* is intuitively productive, whereas *-sam* is intuitively unproductive. Adapted from [9].

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 [13, 14], 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. In German codas, clusters of up to 5 consonants can be observed, whereas English allows up to 4 coda consonants. Thus, the maximum number of consecutive consonants across syllable boundaries is 9 in German, and 7 in English.

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 more than 12,000 distinct syllable types in either language, only about 500 types are systematically and regularly used in speech production. According to the concept of a mental syllabary [15, 16], these high-frequency syllables are stored as complete gestural programs which are executed during speech production, whereas the vast majority of low-frequency and very rare syllables is assembled online by using the segmental and metrical information provided by the phonological encoder.

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 [1, 17] 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 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 [18]. This approach applies multidimensional EM-based clustering to syllable structure, modeling either three dimensions (onset, nucleus, coda) or five dimensions (stress and position, additionally). The advantage of this probabilistic method is that the induced models assign probabilities even to syllable types that are not covered by the training database, thus offering a reasonable solution to the LNRE problem in the domain of syllabification.

2.3. Duration modeling

The task of the duration component in a TTS system is to predict the temporal structure of synthetic speech from symbolic input. Among the most important factors in many languages are the position of the word in the phrase or utterance, the accent status of the word, syllabic stress, and the segmental context. These factors and their values define a large feature space.

The prevalent type of duration model is a sequential rule system such as the one proposed by Klatt [19]. 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 corpora became available, new approaches were proposed based on, for example, Classification and Regression Trees (CART [20]) [21, 22] and neural networks [23]. It has been shown, however, that even huge amounts of training data cannot exhaustively cover all possible feature vectors [24].

Manual database construction, on the other hand, is only feasible if the factorial space is not too large. Unfortunately, at least 17,500 distinct feature vectors have been observed in

American English [25].

The majority of observed feature vectors has 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 [4] 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 [26], 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 proposed by van Santen [27, 24] has been shown to be superior to CART-based approaches, for several reasons [28]. 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 factor levels 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 [25, 24], Mandarin Chinese [29], Japanese [30], and German [5].

2.4. Concatenative speech synthesis

Evidently, LNRE distributions also play a crucial role in data-driven concatenative speech synthesis. Beutnagel and Conkie [31] 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; 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, which means that they also generate superior synthesis quality compared to the demiphone solution.

For the construction of the database for a new Japanese synthesis system [32] 50,000 multi-form units were collected that cover approximately 75% of Japanese text. Multi-form units are designed to cover all Japanese syllables and all possible vowel sequences, realized in a variety of prosodic contexts. In conjunction with another set of 10,000 diphone units this database accounts for 6.3 hours of speech. Given the relatively simple syllable structure of Japanese, the emphasis should be on *only* 75% coverage.

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% [32, Fig. 2]. The authors state that for unrestricted text the actually required number of units approaches infinity, and that the majority of the units is rarely used—a characteristic of LNRE distributions. The question of how to get to near 100% coverage remains unanswered, in fact even unasked.

3. Closed domains

It has often been suggested that for restricted domains a version of the unit selection synthesis strategy might be feasible that exploits units larger than demiphones, phones, or diphones. In the most recent version of the synthesis component developed in the Verbmobil project [33], a word concatenation approach has been implemented [34].

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. A feasible alternative might be to generate syllables from phoneme realizations and words from syllables [34].

A system based on word and syllable concatenation has also been presented for the limited domain of weather forecasting [35]. 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 [36] study on the coverage index of training databases for unit selection synthesis, we are led to believe that their optimism is unwarranted.

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 fre-

quently used in synthesis" [37, page 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, such as in a sentence to be synthesized, approaches certainty.

In this paper we 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 we 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 [1] 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, building on a model of the order in which word formation processes operate on a simplex form. Such a system would apply sophisticated statistical models that are capable of dealing with LNRE properties, to the quantitative analysis of morphological productivity [12, 9].

A probabilistic approach to syllabification 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 [18].

In the context of modeling segmental durations we concluded 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 suggested solution to this problem was the application of a class of arithmetic models known as sums-of-products models [27]. These models have been shown to cope 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 partially attributed to complexity and combinatorics of language and speech in general, and to LNRE properties in particular. Yet, we suggest 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.

The design of databases for restricted application domains, where the distributions of linguistic and phonetic factors are known, might be a feasible step in this direction. Two caveats were discussed in this context. First, we 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.

5. References

[1] Bernd Möbius, "The Bell Labs German text-to-speech system," *Computer Speech and Language*, vol. 13, pp. 319–358, 1999.

[2] Richard Sproat, Ed., *Multilingual Text-to-Speech Synthesis: The Bell Labs Approach*, Kluwer, Dordrecht, 1998.

[3] Harald Baayen, *Word Frequency Distributions*, Kluwer, Dordrecht, 2000.

[4] Jan P. H. van Santen, "Computation of timing in text-to-speech synthesis," in *Speech Coding and Synthesis*, W. Bastiaan Kleijn and Kuldip K. Paliwal, Eds., pp. 663–684. Elsevier, Amsterdam, 1995.

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

[6] H. Schultink, "Produktiviteit als morfologisch fenomeen," *Forum der Letteren*, vol. 2, pp. 110–125, 1961.

[7] Harald Baayen, "On frequency, transparency and productivity," *Yearbook of Morphology 1992*, pp. 181–208, 1993.

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

[9] Stefan Evert and Anke Lüdeling, "Measuring morphological productivity: Is automatic preprocessing sufficient?," in *Proceedings of Corpus Linguistics 2001 (Lancaster, UK)*, 2001.

[10] Harald Baayen and Rochelle Lieber, "Productivity and English derivation: a corpus based study," *Linguistics*, vol. 29, pp. 801–843, 1991.

[11] Tanja Schmid, Anke Lüdeling, Bettina Säuberlich, Ulrich Heid, and Bernd Möbius, "DeKo: Ein System zur Analyse komplexer Wörter," in *Proceedings of GLDV-2001 (Gießen, Germany)*, Henning Lobin, Ed., 2001, pp. 49–57.

[12] Anke Lüdeling, Stefan Evert, and Ulrich Heid, "On measuring morphological productivity," in *Proceedings of KONVENS 2000 (Ilmenau, Germany)*, 2000, pp. 57–61.

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

[14] Jan P. H. van Santen and Bernd Möbius, "A quantitative model of F0 generation and alignment," in *Intonation—Analysis, Modelling and Technology*, Antonis Botinis, Ed., pp. 269–288. Kluwer, Dordrecht, 2000.

[15] Willem J. M. Levelt and L. Wheeldon, "Do speakers have access to a mental syllabary?," *Cognition*, vol. 50, pp. 239–269, 1994.

[16] Willem J. M. Levelt, "Producing spoken language: a blueprint of the speaker," in *The Neurocognition of Language*, Colin M. Brown and Peter Hagoort, Eds., pp. 83–122. Oxford University Press, Oxford, UK, 1999.

[17] Bernd Möbius and George A. Kiraz, "Word models and syllabification in multilingual TTS," in *Progress in Speech*

Synthesis II: Proceedings of the Third International International Workshop on Speech Synthesis, Andrew Breen, W. Nick Campbell, Jan van Santen, and Julie Vonwiller, Eds. Springer, Berlin, 2000, In press.

[18] Karin Müller, Bernd Möbius, and Detlef Prescher, "Inducing probabilistic syllable classes using multivariate clustering," in *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, Hong Kong, 2000, ACL, pp. 225–232.

[19] Dennis H. Klatt, "Interaction between two factors that influence vowel duration," *Journal of the Acoustical Society of America*, vol. 54, pp. 1102–1104, 1973.

[20] Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone, *Classification and Regression Trees*, Wadsworth & Brooks, Pacific Grove, CA, 1984.

[21] John F. Pitrelli and Victor W. Zue, "A hierarchical model for phoneme duration in American English," in *Proceedings of the European Conference on Speech Communication and Technology (Paris)*, 1989, pp. 324–327.

[22] Michael D. Riley, "Tree-based modeling for speech synthesis," in *Talking Machines: Theories, Models, and Designs*, Gérard Bailly, Christian Benoît, and Thomas R. Sawallis, Eds., pp. 265–273. Elsevier, Amsterdam, 1992.

[23] W. Nick Campbell, "Syllable-based segmental duration," in *Talking Machines: Theories, Models, and Designs*, Gérard Bailly, Christian Benoît, and Thomas R. Sawallis, Eds., pp. 211–224. Elsevier, Amsterdam, 1992.

[24] Jan P. H. van Santen, "Assignment of segmental duration in text-to-speech synthesis," *Computer Speech and Language*, vol. 8, pp. 95–128, 1994.

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

[26] Nobuyoshi Kaiki, Kazuya Takeda, and Yoshinori Sagisaka, "Statistical analysis for segmental duration rules in Japanese speech synthesis," in *Proceedings of the International Conference on Spoken Language Processing (Kobe, Japan)*, 1990, pp. 17–20.

[27] Jan P. H. van Santen, "Exploring N-way tables with sums-of-products models," *Journal of Mathematical Psychology*, vol. 37, no. 3, pp. 327–371, 1993.

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

[29] Chilin Shih and Benjamin Ao, "Duration study for the Bell Laboratories Mandarin text-to-speech system," in *Progress in Speech Synthesis*, Jan van Santen, Richard W. Sproat, Joseph P. Olive, and Julia Hirschberg, Eds., pp. 383–399. Springer, New York, 1997.

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

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

[32] Kimihito Tanaka, Hideyuki Mizuno, Masanobu Abe, and Shin-ya Nakajima, "A Japanese text-to-speech system based on multi-form units with consideration of frequency distribution in Japanese," in *Proceedings of the European Conference on Speech Communication and Technology (Budapest, Hungary)*, 1999, vol. 2, pp. 839–842.

[33] Wolfgang Wahlster, Ed., *Verbmobil: Foundations of Speech-to-Speech Translation*, Springer, Berlin, 2000.

[34] Karlheinz Stöber, Thomas Portele, Petra Wagner, and Wolfgang Hess, "Synthesis by word concatenation," in *Proceedings of the European Conference on Speech Communication and Technology (Budapest, Hungary)*, 1999, vol. 2, pp. 619–622.

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

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

[37] Robert E. Donovan and P. C. Woodland, "A hidden Markov-model-based trainable speech synthesizer," *Computer Speech and Language*, vol. 13, pp. 223–241, 1999.