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

Bernd Möbius

Institute of Natural Language Processing
University of Stuttgart

1 Introduction

In this paper we intend to point out two common concepts in speech synthesis that we consider questionable, if not misguided and wrong. The first of these concepts is the 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. 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. The third area in TTS conversion that is affected by imbalanced frequency distributions is the design of acoustic unit inventories for data-driven speech synthesis.

The second concept that we consider questionable is the notion of a "restricted" application domain. We conclude 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 (Möbius, 1999; Sproat, 1998) 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 (2000) 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).

Similarly unpleasant frequency distributions are observed in *segmental duration* modeling (section 2.2). 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 (van Santen, 1995; Möbius and van Santen, 1996). 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 2.3). 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. In German, derivation and composition are the most important means of productive word formation, and they can generate an unlimited number of new words. The construction of exhaustive word lists 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.

The Bell Labs German TTS system (Möbius, 1999) has therefore been equipped with the capability to morphologically decompose unknown words and provide an annotation whose granularity approaches that of the annotation of words listed in the TTS lexicon. The analysis component for unknown words and names is based on a model of the morphological structure of words and the phonological structure of syllables. We also performed a study of the productivity of word forming affixes (Möbius, 1998), applying the simple statistical estimate of productivity suggested by Baayen (1993).

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 then is a—morphologically complex—word type with a token count of 1. 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): $P = n1/N$.

More recently, Baayen has developed much more elaborate statistical methods for estimating morphological productivity (Baayen, 2000) and, more generally, for coping with the extremely uneven LNRE distributions of word frequencies. One important conclusion from this work is that the text corpora used in the earlier studies (Baayen and Lieber, 1991; Möbius, 1998) 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.

2.2 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 (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 corpora became available, new approaches were proposed based on, for example, Classification and Regression Trees (CART (Breiman et al., 1984)) (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 feature vectors (van Santen, 1994).

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 (van Santen, 1993b).

It is certainly true that 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 (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, Takeda, and Sagisaka, 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 proposed by 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.

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

2.3 Concatenative speech synthesis

Evidently, LNRE distributions also play a crucial role in data-driven concatenative speech synthesis. Beutnagel and Conkie (1999) report that more than 300 diphones out of a complete set of approximately 2000 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 (Tanaka et al., 1999) 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% (Tanaka et al., 1999, Fig. 2). The authors state that for unrestricted text the actually required number of units approaches infinity, and that the majority of the 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.

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 (Wahlster, 1997), 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.

A system based on word and syllable concatenation has also been presented by Lewis and Tatham (1999), for the limited domain of weather forecasting. The system has an inventory of 2000 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, Lewis and Tatham 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, we are led to believe that Lewis and Tatham's 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 frequently used in synthesis" (Donovan and Woodland, 1999, 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 three components of a TTS system: linguistic text analysis, 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 (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.

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 solution to this problem was the application of a class of arithmetic models known as sums-of-products models (van Santen, 1993a).

No 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 its LNRE properties in particular. Yet, the most promising line 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 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, resulting in very uneven speech output quality.

References

Baayen, Harald. 1993. On frequency, transparency and productivity. *Yearbook of Morphology 1992*, pages 181–208.

Baayen, Harald. 2000. *Word Frequency Distributions.* Kluwer, Dordrecht.

Baayen, Harald and Rochelle Lieber. 1991. Productivity and English derivation: a corpus based study. *Linguistics*, 29:801–843.

Beutnagel, Mark and Alistair Conkie. 1999. Interaction of units in a unit selection database. In *Proceedings of the European Conference on Speech Communication and Technology (Budapest, Hungary)*, volume 3, pages 1063–1066.

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

Campbell, W. Nick. 1992. Syllable-based segmental duration. In Gérard Bailly, Christian Benoît, and Thomas R. Sawallis, editors, *Talking Machines: Theories, Models, and Designs*. Elsevier, Amsterdam, pages 211–224.

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

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

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

Lewis, Eric and Mark Tatham. 1999. Word and syllable concatenation in text-to-speech synthesis. In *Proceedings of the European Conference on Speech Communication and Technology (Budapest, Hungary)*, volume 2, pages 615–618.

Maghbouleh, Arman. 1996. 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)*, pages 1–7.

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

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

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

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

Riley, Michael D. 1992. Tree-based modeling for speech synthesis. In Gérard Bailly, Christian Benoît, and Thomas R. Sawallis, editors, *Talking Machines: Theories, Models, and Designs.* Elsevier, Amsterdam, pages 265–273.

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

Sproat, Richard, editor. 1998. *Multilingual Text-to-Speech Synthesis: The Bell Labs Approach.* Kluwer, Dordrecht.

Stöber, Karlheinz, Thomas Portele, Petra Wagner, and Wolfgang Hess. 1999. Synthesis by word concatenation. In *Proceedings of the European Conference on Speech Communication and Technology (Budapest, Hungary)*, volume 2, pages 619–622.

Tanaka, Kimihito, Hideyuki Mizuno, Masanobu Abe, and Shin-ya Nakajima. 1999. 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)*, volume 2, pages 839–842.

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

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

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

van Santen, Jan P. H. 1995. Computation of timing in text-to-speech synthesis. In W. Bastiaan Kleijn and Kuldip K. Paliwal, editors, *Speech Coding and Synthesis*. Elsevier, Amsterdam, pages 663–684.

van Santen, Jan P. H. 1997. Combinatorial issues in text-to-speech synthesis. In *Proceedings of the European Conference on Speech Communication and Technology (Rhodos, Greece)*, volume 5, pages 2511–2514.

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

Wahlster, Wolfgang. 1997. VERBMOBIL: Erkennung, Analyse, Transfer, Generierung und Synthese von Spontansprache. Technical report, DFKI, Saarbrücken. Verbmobil-Report 198.