Keeping Rare Events Rare

Ove Andersen, Charles Hoequist

Department of Communication Technology, Aalborg University,
Fredrik Bajers Vej 7, DK-9220 Aalborg Øst, Denmark
`[oa,ch]@kom.auc.dk`

Abstract

It has been claimed that corpus-based TTS is unworkable because it is not practical to include representative units to cover all or most of the combinations of segments and prosodic characteristics found in general texts, a problem characterized as Large Numbers of Rare Events (LNRE). We argue that part of this problem is in its formulation, and that a closer look, including investigations into corpus-based TTS for Danish, show that LNRE need not be a fatal problem for inventory design in corpus-based TTS.

1. Introduction

As storage has become cheap, concatenative speech synthesis has moved away from small and fixed-size inventories to large corpus-based systems containing multiple instances of most or all tokens in the database. The motivation is to minimize the need for runtime modification of the database units, which is necessary in small-library concatenation both to smooth discontinuities at unit boundaries and to modify prosodic features of the stored signals. No matter how sophisticated the signal processing, the more a token is modified away from its original stored form, the more its naturalness and perhaps intelligibility suffers. In creating expanded libraries of units, systems have run up against a feature of human language mentioned early in every introductory linguistics textbook: the capacity of a language to use a fixed number of distinctive units to create an unbounded number of different outputs. Typically, this capacity is illustrated by morphological and syntactic examples, but it is equally true for phonetics.

It is obviously impossible to store every possible word of a language in a database. It may be less obvious, but it is equally impossible, to store every combination of concatenation unit (syllable, demisyllable, diphone, triphone) and prosodic feature which a system may encode. Tanaka et al. provide a cautionary example: an inventory of the most frequent 50,000 monomoraic and bimoraic syllables in Japanese achieves coverage of only 77% of an average text, even with very limited use of prosodic factors [1].

However, we may be holding synthesis to an impossible and unnecessary standard. In the real world, phonotactic restrictions ensure that we will not have to account for every orthogonal feature combination. This may restrict the unit plus feature combinations to something that can be adequately covered in a corpus of reasonable size.

2. The LNRE principle

Unfortunately, even this restricted search space seems to be difficult to master for corpus-based text-to-speech. Möbius argues that corpus-based synthesis is inherently hampered by a property of frequency distributions called Large Number of Rare Events (LNRE) [2]. The core characteristic of these distributions is a very long tail of very low-frequency events. In his article, these are discussed in three contexts: text analysis, durational modeling, and inventory design, with different events in each context [2]. We will focus on one context, inventory design for a database. For corpus-based synthesis, these events are rarely occurring concatenation units. It is argued that these are distributed so that almost any text, however small, has a few, but no corpus, however large, has them all. This means that every corpus-based system will display gaps in coverage almost every time it is used. The conclusion is that such events are unavoidable, and that they set a low ceiling on the quality of corpus-based TTS for all but the most restricted domains.

We would like to argue that the situation is not as grim as has been presented. Several related questions are being asked simultaneously, and it is not at all clear that they all deserve the same answer.

First, there is the unit being discussed. A diphone label by itself does not adequately specify the input for generation of acoustic output in a TTS system. An input can be viewed as a vector containing additional descriptors both of the target and its context, e.g. diphone identity, F0 contour, diphones on either side, accent category and duration. Even such a small vector, however, can create a huge search space. If, for example, each of a thousand diphones has only ten possible independent left and right contexts, two

possible accent levels and two durations, complete coverage in a database entails four hundred thousand units. Any real example, based on a system capable of unrestricted text input, has a space many orders of magnitude larger. Starting from this basis, the LNRE principle is almost trivial: the overwhelming majority of vectors in a text will be rare because of the size of the vector space. It is always possible to make a vector long enough to make coverage of the search space miniscule, especially since there is no agreed-upon set of factors or number of levels within a factor. Presumably, this is not what is meant by the LNRE principle applied to inventory design.

Though there is no agreed-upon principle, we would like to make a suggestion. It is implicit in the construction of databases for corpus-driven TTS that there is a hierarchy of importance for coverage. For example, obtaining an unstressed and a stressed version of a vowel is far more important than a phrase-medial and a phrase-final version, and both are more important than getting two examples differing by only 10 Hz. Though the particular ordering in the middle of the hierarchy is variable, and is certainly language-dependent to a degree, at one extreme is what we will call target coverage, that is, getting at least one example of a diphone. Failure of target coverage is a serious gap in the database. To the degree that gaps in the inventory of diphone labels are a part of LNRE, corpus-based TTS is going to have problems. F0 or duration can be modified or interpolated; modifying one consonant into another is, at best, impractical.

It does not seem to be the case that target coverage is unachievable. Much of van Santen and Buchsbaum, in fact, is devoted to showing precisely the opposite; namely that with careful planning, target coverage is not only possible but also feasible with a surprisingly small set of properly crafted sentences [3]. This is where our current work comes in.

3. LNRE and databases in Danish

The authors are involved in the development of a TTS system for Danish [4]. The system currently uses a set of some 3000 diphones and triphones excised from nonsense utterances for its database, and we are investigating the feasibility of instead using recorded texts as diphone sources. To test the feasibility of using arbitrary texts, we have compared coverage among arbitrary texts and word lists, to see how much coverage can be obtained from unrelated corpuses, without any crafting of texts. To the degree that this is feasible, it would mean that new corpuses for TTS could be created from existing recorded texts, rather than having to make new recordings each time.

The following sources were used:
1. the text of the book *De Kompetente Forældre (The Capable Parents)* [5] (DKF)
2. the text of the book *Vinsmagning (Wine Tasting)* [6] (VIN)
3. the unique entries of the Danish Language Commision's Orthographic Dictionary [7] (DIC), discounting multiple senses or grammatical categories of words.

Statistics on text sizes are given in Table 1.

Table 1 : Source text sizes

	DKF	VIN	DIC
#words	48606	64107	61387
Audio Length	5h30m	9h42m	n/a
#different words	5361	9246	61387
#phones	38	38	38
#diphones	318226	416073	618067
#different diphones	994	1083	1159

4. Analysis of type and token coverage

In all cases, words were transcribed and turned into diphones by the analysis and database-search code of the TTS system described in [4]. The book texts additionally yielded transitional diphones across words, since entire sentences were processed at once. DIC was used as a baseline for checking coverage of Danish words and assimilated loanwords. While there is no such thing as an exhaustive list of words, especially given productive affixes, DICs coverage of lexical items is likely to be the best of any corpus available in Danish.

Fig. 1 shows type coverage results for the diphone labels of DKF as covered by those found in VIN. That is, we took VIN as the source for diphone types and asked how successful this inventory would be in synthesizing the text of DKF. The coverage is given in percent of the DKF text as a funtion of the number of words included from the VIN text for generating a diphone inventory. The five curves show various ranges of tokens for target type. For example, if we want at least 25 instances of a given target diphone, and if we use the first 25,000 words in VIN, we will cover around 60% of the diphone types required to synthesize DKF. This is not strictly relevant to the question of label

coverage, but will affect prosodic modeling, which we will touch on briefly later. For multiple tokens, the type coverage is comparable to the low coverage found in Möbius [8]. If one wants to do modeling of prosodic features for all database segments, there will be a problem, in that low-frequency types will be inadequately covered. However, the target coverage—at least one token for a type—is very nearly complete, which is pleasantly surprising given the considerably different text domains. In addition, coverage keeps rising with increasing corpus size, and does not flatten out early.

Figure 2 shows coverage in terms of token numbers, thus giving an idea of the weighting to assign to the types. Not surprisingly, the types that are covered are the most frequently occurring ones. What is encouraging is that this weighted coverage shows near-total token coverage very early. We believe that this difference illustrates the possibility of target coverage, which is vital to corpus-based TTS. Further, the fact that the weighted results give good coverage is encouraging for both quality and for prosodic modeling, as it indicates that the most frequent diphones will have the largest number of alternatives to choose from, thus reducing the need for signal modification at synthesis time.
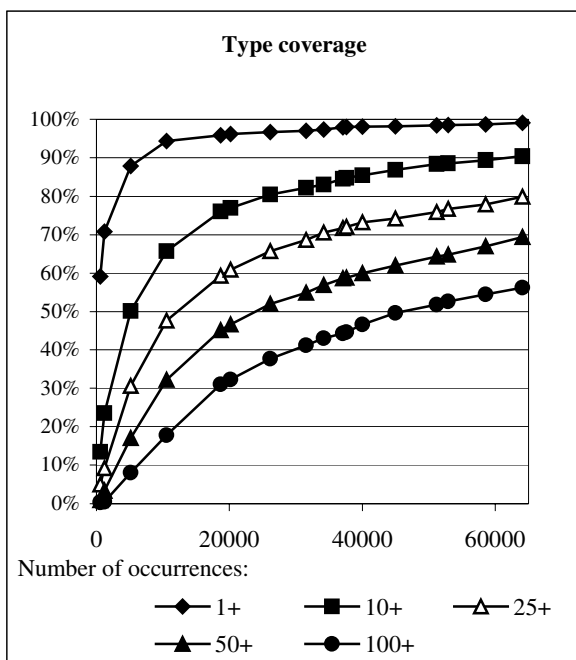

Figure 1: Percent type coverage as a function of the number of words included from the source text. The five curves illustrate coverage as a function of different levels of minimum token numbers.Symbols on the curves show chapter boundaries as a function of cumulative word count.

Fig. 2 provides an illustration of the token coverage that is possible when VIN is used for synthesizing the DKF text.
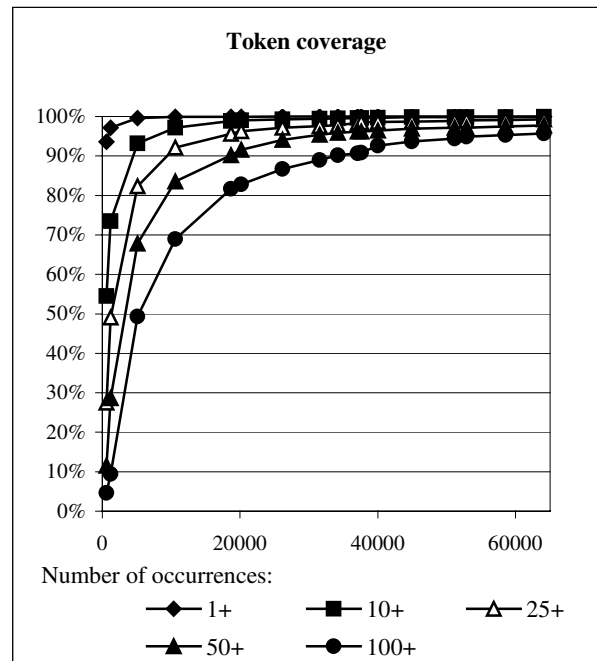

Figure 2: Percent token coverage as a function of the number of words included from the source text. The five curves illustrate coverage as a function of different levels of minimum token numbers. .Symbols on the curves show chapter boundaries as a function of cumulative word count.

A closer look at the results revealed an interesting phenomenon. In all, there were seven diphones in DKF that had no corresponding token in VIN, by no means a large number. When the sources of the missing diphones were investigated, it was found that five of them were not part of any lexical item. Instead, they were diphones connecting words in the text. This leads to an interesting possible interpretation of LNRE, that a large part of the events are not hidden corners of a lexicon, but the result of a system's combinatorics. This preponderance of interword culprits holds when we reversed the comparison, that is, looked at diphones in VIN which are not covered in the smaller DKF text: of a total of 91, 44 are across words. The high total number of missing targets is due to the large number of unassimilated loan words in the text, not surprising in a book for oenophiles. These loans accounted for a further 32 of the non-covered targets.

As a further test of lexical coverage, we looked at the number of missing targets using the orthographic dictionary as the source for the database inventory. Initially, we had regarded this as a sanity check, since it

was clear that all the words in DKF were also present in DIC. To our surprise, coverage was *worse* than when using VIN: 24 missing targets. Inspection of the particular targets revealed that all of them were interword diphones, illustrating again that the LNRE problem is mainly one of combinatorics: coverage is not just a question of units, but of unit sequences.

Finally, we looked at number of occurrences of the seven noncovered diphones in VIN's coverage of DKF. We found that 37 words in DKF were affected by a missing target, and of these seven were lacking an internal target, which would block proper synthesis of that particular lexical item. This does not bear out the contention that LNRE leads to frequent gaps: 37 total occurrences in a text of 48,000 words means that it is possible to output at least 1300 words without a gap in target coverage.

5. Conclusion

Notwithstanding the existence of LNRE, corpus-based TTS is not dead, nor necessarily even sick. It is possible to achieve target coverage with a small, unselective text, and even to get multiple-unit coverage for a large part of an arbitrary text to be synthesized. This does not, however, lead us to deny that there is such a thing as LNRE. On the contrary, we can ascribe it to the productivity of language at all levels. For TTS, this means that simply covering a vocabulary is not adequate for a system that wishes to handle text. There will always be interword targets not covered by any mere list of words, regardless of length.

The existence of targets that exist purely to handle interword transitions leads us in two directions for further investigation. First, there is the possibility of creating a quick 'hit list' of target units by comparing intra- and intersyllabic diphones, and focusing on the latter, since they are most likely to be missed, even with a large text. Second, there is a possibility connected to the coverage hierarchy mentioned above. In such a hierarchy, units connecting words may be less important than word-internal units, because of reduced coarticulation or reduced attention by the listener. This in turn suggests the increased possibility of introducing some kind of patch for missing units across words, such as interpolating values for factors not found in the database. We are presently designing experiments addressing this issue.

6. References

1. Tanaka, Kimihito; Mizuno, Hideyuki; Abe, Masanobu; and Shinya Hakajima. "A Japanese Text-To-Speech System Based on Multi-form Units with Consideration of Frequency Distribution in Japanese", *EuroSpeech '99: 6th European Conference on Speech Communication and Technology,* vol. 2, pp.839-842, 1999.
2. Möbius, Bernd. "Corpus-Based Speech Synthesis: Methods and Challenges", *Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung (AIMS) 6* (4), Universität Stuttgart, pp. 87-116, 2000.
3. van Santen, Jan P.H.; Buchsbaum, Adam L.. "Methods for Optimal Text Selection", *EuroSpeech '97: 5th European Conference on Speech Communication and Technology,* vol. 2, pp.553-556, 1997.
4. Andersen, Ove; Hoequist, Charles; and Claus Nielsen. "The Danish Research Ministry's Initiative on Text-to- Speech Synthesis", *NORSIG2000: Proceedings of the Nordic Signal Processing Symposium,* June 13-15, 2000, pp. 327-330, 2000.
5. Brun Hansen, Margrethe, "De Kompetente Forældre", Aschehoug, 2001.
6. Broadbent, Michael, "Vinsmagning", Nordiske Forlag A/S, Copenhagen, 1999.
7. Dansk Sprognævn (1996) "Retskrivningsordbogen", 2. udgave, Aschehoug
8. Möbius, Bernd (2001). "Rare Events and Closed Domains: Two Delicate Concepts in Speech Synthesis.", paper delivered at the 4th ISCA Tutorial and Research Workshop on Speech Synthesis, Perthshire, Scotland, August 29th - September 1st, 2001.