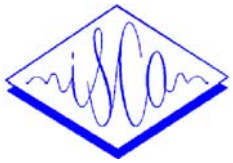

EUROSPEECH '97
5th European Conference on Speech Communication and Technology
Rhodes, Greece, September 22-25, 1997

COMBINATORIAL ISSUES IN TEXT-TO-SPEECH SYNTHESIS

Jan P. H. van Santen

Lucent Technologies – Bell Labs, 600 Mountain Ave., Murray Hill, NJ 07974, U.S.A.
jphvs@research.bell-labs.com

ABSTRACT

Enhanced storage capacities and new learning algorithms have increased the role of text and speech training data bases in the construction of text-to-speech systems. It has become apparent, however, that not always learning algorithms are available that have strong generalization capabilities – the ability to generalize from cases seen in the training data base to new cases encountered during TTS operation. This makes it important to measure and understand the degree of *coverage* of the input domain of a text-to-speech system (usually, the entire language) by a given training data base. The goal of this paper is to investigate the feasibility of coverage in several domains of interest for TTS. It is shown that, as a result of the combinatorics of language, coverage is typically quite disappointing. This puts a premium on the generalization capability of learning algorithms.

10.21437/Eurospeech.1997-651

1. INTRODUCTION

A typical text-to-speech (*TTS*) system represents textual input in terms of various classes of units, such as words, phonetically transcribed words, contextual vectors for duration prediction (e.g., $< Stressed, InCoda, Phrase-Medial, \ldots >$), and acoustic inventory elements. Since the goal of text-to-speech synthesis is to mimic human speech, we can characterize the task of a TTS component as that of *predicting* human output from input units.

Two distinct traditions have developed in TTS construction. In one (*data based approaches*), general purpose statistical techniques extract parameters from a training data base; little knowledge of the content area is used. A typical example is usage of classification and regression trees (*CART*) for prediction of segmental duration from context [5]. Here, the learning algorithm receives as input combinations of segmental durations and vectors for duration prediction, and sets up a tree by branching on vector components such as to minimize the variance at each terminal node. Except for the decision what information to include in the vectors, no content specific knowledge is used.

In the other (*knowledge based approaches*), content-specific models are used, perhaps in conjunction with parameter estimation. An example is the approach to segmental duration modeling used by Klatt in the MITalk system [1], where durations are predicted with rules such as *if vowel V is stressed, subtract m_V ms from the intrinsic duration i_v, multiply by 1.4, and add m_V*. The mathematical structure of this rule is knowledge based, and the values of the parameters m_V and i_v are obtained via statistical estimation. The knowledge involved here is the reasonable idea that all vowels are lengthened by stress, and that this effect is larger (measured in ms) for intrinsically longer vowels.

These two approaches represent opposite poles on a continuum of increasing reliance on training data bases and decreasing reliance on prior knowledge. Increases in data storage, computational speed, and new algorithms seem to make data based approaches increasingly more attractive. At the same time, it has become apparent that these approaches have fundamental shortcomings generalizing from unit types seen in the training data to unseen unit types. (We distinguish between unit *types* and unit *tokens*; the latter are individual instantiations of the former.) A good example is provided in a study by Maghbouleh [4], who found that a knowledge based approach [7] generalized much better than CART to test materials drawn from a different corpus than the training materials. In fact, even when CART was given two orders of magnitude more training data than the knowledge based approach, the difference in performance hardly decreased.

However, if all unit types can be covered in a training corpus, then it may be preferable to use general-purpose statistical techniques. The reason is that we have to pay a price for the better generalization capability of knowledge based techniques, which is that the assumptions these techniques are based on are imperfectly true. Hence, they cannot represent data on a given set of unit types as accurately as (unconstrained) general-purpose statistical techniques. If no generalization is needed, the latter techniques should be superior.

A fundamental issue to be resolved when deciding what type of approach to use for a particular TTS component is whether a practically feasible training data base can be developed that is large enough to cover the unit type space. If such a data base cannot be developed, then one might consider using knowledge-based methods, even when initial results with general-purpose statistical techniques look promising.

The goal of this paper is to investigate the feasibility of coverage in several domains of interest for TTS.

2. CONCATENATIVE INVENTORIES

The first analyses concern concatenative inventories for text-to-speech synthesis. In most systems, a small num-

ber (fewer than 5,000, often fewer than 2,000) of context-independent acoustic inventory elements (corresponding to n-phones, such as diphones) is generated from a speech data base. These units are often diphones, although longer units are also used. A common idea is to take advantage of increased storage capabilities, and drastically increase the size of concatenative inventories. We discuss here two proposals: (1) Context-specific concatenative units, and (2) Obstruent-terminated units.

2.1. Prosodic units

At run time, signal processing methods change timing and pitch to make the unit appropriate for the context in which it occurs. Since these methods often introduce audible distortions, it has been proposed to use far larger concatenative inventories that would cover combinations of n-phones and contexts and thus would require little signal processing at run time.

In our analysis, 250,000 sentences from the Associated Press Newswire were automatically transcribed by the text analysis components of the Bell Labs Text-to-Speech System. Next, we constructed a contextual vector for each diphone in the transcribed sentences, containing information about key factors such as accent status (accented vs. deaccented) and within-utterance position (initial, final, medial). The factors were kept deliberately coarse. The type count of diphone-context combinations was 222,678. Since most of these combinations are extremely rare, one cannot conclude that all these combinations must be covered in the training set (acoustic inventory) for adequate coverage of the input domain: if a system misses a unit once every thousand sentences, this should not be considered a problem. However, when we analyzed the data using a statistic we call the *coverage index* of a training set, the situation turned out to be more problematic than that.

The coverage index of a given training set with respect to a given domain is defined as the probability that all combinations occurring in a randomly selected test sentence are represented in the training set. Thus, 0.75 coverage means that the probability is 0.75 that all combinations in a randomly selected test sentence also occurred in the training set.

We found that a training set of 25,000 combinations had an index of only 0.03, and that an index of 0.75 required a training set of more than 150,000 combinations (see Figure 1). Given that both training and test sentences were from the same text "genre" (Associated Press Newswire), the values of the coverage index will be worse when there are large differences between the text genres used for training vs. test. Hence, context-sensitive approaches to acoustic inventory construction require astronomical databases, which may pose problems not so much for computer storage as for speakers.

The likely cause for these results is that the type count of rare combinations is quite large, so that even when the probability of a *particular* rare combination occurring in a given sentence is by definition small, the probability of *some* rare combination occurring is surprisingly large.

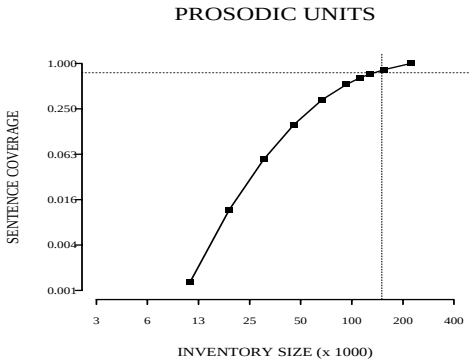

Figure 1: **Coverage index for prosodic units as a function of inventory size. The horizontal dotted line indicates a coverage index of 0.75, for which a 150,000 unit training set is required**

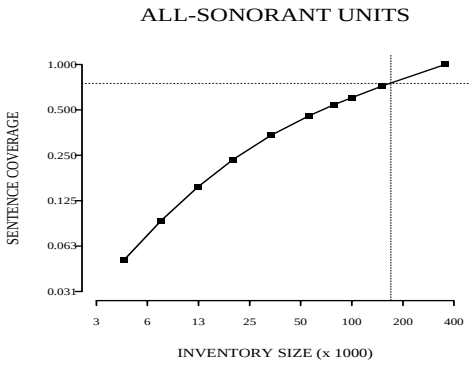

Figure 2: **Coverage index for all-sonorant units as a function of inventory size. The horizontal dotted line indicates a coverage index of 0.75.**

2.2. Obstruent-terminated units

It it obvious that some acoustic unit cutpoints are preferable over other cutpoints. For example, voiceless stop closures are a good place to cut, because these regions are hardly affected by coarticulatory processes and have very low energy. On the other hand, the schwa is short and is heavily affected by surrounding phones. Acoustic units that end on or start with schwa are likely to have spectral discrepancies. Hence, many systems embed schwa's in triphones, not diphones. A proposal flowing from these considerations is to only use units that are cut in obstruent regions.

We conducted the same type of analysis as in the previous subsection, with similar results (see Figure 2).

2.3. Conclusions

These two analyses show that the frequency distributions of the proposed units make it impractical to construct con-

catenative inventories from them. The fact that both training and test materials were from the same genre makes these results all the more powerful.

One practical implication is that while it certainly does not hurt speech quality to include a certain number of all-sonorant or prosodic units, the expected additional coverage from those units may be disappointingly small. For the overall quality of the system, it may be much more important to focus resources on (1) optimizing a smaller set of units that is known to have complete coverage, and on (2) development of signal processing techniques that are robust with respect to the concatenation operation and to fundamental frequency and timing alterations.

3. DURATION MODELING

The next analysis is related to the preceding analysis, but concerns construction of the component that computes segmental duration. Here, we used the same contextual annotation scheme as before, but applied it to individual phones (not diphones). Thus, the basic unit here consisted of combinations of segment identity and contextual vector. We analyzed 797,524 sentences, names, and addresses (total word token count: 5,868,172; total phonetic segment count 22,249,882). The total combination type count was 17,547. Of these 17,547 types, about 10 percent occurred only once in the entire data base and 40 percent occurred less than once in a million phone occurrences. We found that for samples in excess of 5000 units, the type count increased linearly with the logarithm of the number of units, with no sign of deceleration. Hence, it is uncertain whether the true type count in the language is 20,000, 30,000, or significantly larger than that. We also found that even in samples as small as 320 units (the equivalent of a small paragraph), the probability of encountering a unit occurring only once in a million cases is near unity. Not surprisingly, the coverage index of randomly selected training sets was quite low even for large sets.

4. DISTRIBUTIONAL DIFFERENCES BETWEEN CORPORA

Often, training data are drawn from one particular corpus or class of corpora. For example, some components of the Bell Labs text-to-speech system are trained on the Associated Press Newswire. This raises the issue of how much overlap there is between different corpora. After all, we intend our text-to-speech system to work equally well on, say, hotel reservation applications, newspaper headlines, email, traffic directions, stock quotes, and wholesale groceries. In this section, we discuss distributional differences (i.e., differences in the frequency distribution of a given unit class) between corpora.

4.1. Triphone distributions

How large are the differences in triphone distribution between two distinct text genres? We selected a set of 169,328 personal names, and a set of 347,857 sentences from the Associated Press Newswire not containing any proper names. We found that the triphone distributions were quite different indeed. For example, of the 26,972 triphone types occurring in either text genre, 12,837 (47.5%) occurred in one sets but not in the other. If one takes pairs of smaller sub-samples of the two sets, the overlap further decreases.

	0	1	10	100	1,000	10,000
0	0	0	1,366	888	33	0
1	0	177	309	327	26	0
10	6,417	466	1,043	1,353	114	3
100	3,024	390	1,378	2,936	672	21
1,000	982	106	381	1,947	1,480	119
10,000	124	8	16	222	479	148
100,000	3	0	0	2	3	9

Table 1: Triphone frequencies cross-tabulated for name corpus (rows) and sentence corpus (columns). For example, 888 triphone types occur not at all in the name corpus, but 11-100 times in the sentence corpus.

Table 1 presents a cross-tabulation of the triphone frequencies. The first row and column show that many triphones absent in one corpus are actually quite frequent in the other corpus. For example, 888 of the 2,287 triphones absent in the name corpus occur at least 100 times in the sentences corpus; similarly for 6,417 of the 10,550 triphones absent in the sentences corpus.

These data show that, unless the application domain of a TTS system is well-defined and severely restricted, one cannot expect a system trained on one text genre to be prepared for different text genres.

4.2. Vocabulary distributions

We compared a number of corpora in terms of vocabulary distributions. The corpora used were: Associated Newswire (1988-1992), the Bible, the Quran, the collected works by Shakespeare, proceedings of the U.S. Department of Energy, proceedings of the Canadian Parliament (Hansard), a collection of fiction and non-fiction literature published by Harper & Row, Grolier's Encyclopedia, the Brown Corpus [3], the Comprehensive Textbook of Psychiatry [2], a sample of quotes, and a list of Home Box Office movie descriptions.

For each pair of corpora, we obtained the distributional overlap by computing the product-moment correlation between the log frequencies; words occurring in only one of the two corpora were assigned frequencies of 0.1; words not occurring in either corpus in a pair were omitted from the analysis. We then performed a multidimensional scaling procedure described by Torgerson [6] which yields a multidimensional representation where inter-point distances are maximally (inversely) correlated with the correlations. Thus, this representation provides insight in the pattern of similarities and dissimilarities between the corpora. Overall, the correlation varied between -0.31 (be-

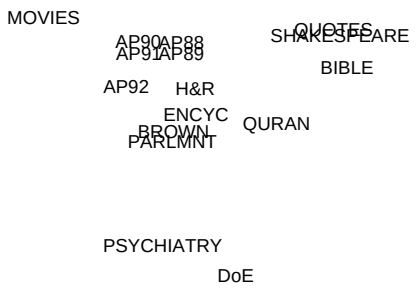

Figure 3: **Multidimensional representation of text genres based on correlations between vocabulary frequency distributions.**

tween the Bible and movie descriptions) and 0.71 (between Associated Press Newswire 1990 and 1991), with a median of 0.22. Thus, even very similar corpora such as the Associated Newswire in two consecutive years did not correlate highly (0.71 represents 50% of the variance).

Figure 3 shows the results of the multidimensional scaling procedure. There is a fair amount of structure in this picture. The Bible, Quran, and Shakespeare share usage of archaic English; the Department of Energy and the Psychiatry Textbook share usage of technical jargon; and the Associated Newswire 1992 saw the arrival of a new cast of characters due to Presidential elections in the US.

These result replicate the results on triphone distributions, and add that the amount of overlap shows systematic patterns where historically related genres have more overlap.

4.3. New tokens vs. new types

A standard method for system evaluation (whether TTS or ASR) is to draw both test and training samples from the same corpus. Although there may exist considerable distributional differences between random samples of the same corpus, our results show that these differences are not nearly as large as those between different text genres. This implies that in the standard test method a relatively high proportion of test tokens does not represent new types, but *new tokens of old types*. As a result, these tests do not tax the generalization capability of the learning system; they test the accuracy with which the system represents types seen in the training data base – which is often quite good for general-purpose statistical techniques. We propose that such tests be carried out using materials from the broadest possible range of corpora.

5. CONCLUSIONS

The goal of this paper was to investigate the feasibility of coverage for several domains of interest for TTS. In all domains investigated, it proved impossible to obtain either complete coverage or at least very high values of the coverage index (indicating near-certainty of coverage) using training data bases of a practically viable size. These results were obtained in analyses where both training and test materials were drawn from the same text corpus. The analyses of the overlap between different text corpora showed that distributional differences between text corpora are large and systematic. This makes obtaining high values of the coverage index even more difficult if the intention is to apply a system trained on one text corpus to other corpora. This is a realistic scenario, however, given the broad and unpredictable range of possible TTS applications.

We conclude that when a learning algorithm is used that is not known to have solid generalization capabilities, it may be critical to investigate whether the TTS input domain can be covered by an affordable training corpus. Initial successes in experiments that do not tax the generalization capability may prove deceptive when the system is confronted with truly new cases.

6. REFERENCES

1. J. Allen, S. Hunnicut, and D.H. Klatt. *From text to speech: The MITalk System*. Cambridge University Press, Cambridge, U.K., 1987.
2. H. I. Kaplan and B.J. (Eds) Sadock. *Comprehensive Textbook of Psychiatry*. Williams & Wilkins, New York, 1985.
3. H. Kucera and W. Francis. *Computational Analysis of Present-Day American English*. Brown University Press, Providence, 1967.
4. A. Maghbouleh. An empirical comparison of automatic decision tree and hand-configured linear models for vowel durations. In *Proceedings of the Second Meeting of the ACL Special Interest Group in Computational Phonology*. Association for Computational Linguistics, 1996.
5. M.D. Riley. Tree-based modeling for speech synthesis. In G. Bailly, C. Benoit, and T.R. Sawallis, editors, *Talking Machines: Theories, Models, and Designs*, pages 265–273. Elsevier, 1992.
6. W. S. Torgerson. *Theory and Methods of Scaling*. Wiley, New York, 1958.
7. J. P. H. van Santen. Assignment of segmental duration in text-to-speech synthesis. *Computer Speech and Language*, 8:95–128, April 1994.