An Analysis of Limited Domains for Speech Synthesis

Robert Batůšek

Faculty of Informatics, Masaryk University, Brno, Czech Republic
xbatusek@fi.muni.cz,

Abstract. This paper deals with the problem of limited domain speech synthesis. Some experiments show that the segment variability is extremely large for unlimited speech synthesis. It seems that it is practically impossible to colllect the text corpus large enough to cover all combinations of even very coarse features. A natural question arises whether restricting the synthesizer to a specific domain can help to increase segment coverage. This paper provides an analysis of several limited domain text corpora and evaluates their applicability to the problem of segment selection for speech synthesis.

1 Introduction

Corpus-based synthesis has become very popular in last years. A significant progress has been done in this field (see [4] for an overview). One of the major remaining problems is the variable quality of the speech output. Van Santen [6] has recently made an analysis that can explain the reasons of the above mentioned problem. He used a text analysis component of its speech synthesizer to generate a feature vector for each diphone detected in a large newspaper textual corpus. He used very simple feature vectors with only a few possible values of each feature. Even with these artifitially coarse vectors and a very large corpus he was not able to achieve the sufficient coverage. Making things even worse, it is not clear what is the final number of feature combinations and how large text corpus should be to ensure full coverage.

Many researches use the corpus-based approach for building speech synthesizers targeted to specific domains [3, 5]. The vocabulary of these domains is typically somehow limited, although not totally closed. As van Santen's results were achieved using a general newspaper corpus, the natural question arises, whether we cannot expect better results for restricted domains. The rest of the paper provides an investigation of several limited domains.

2 Data Collection

We have created several small corpora for experimental purposes. All of them were collected using data publicly available on the Internet. The brief characteristics of the corpora follow:

– **WEATHER** (150,000 phoneme tokens) — daily weather reports from November 15, 1999 till November 15, 2001.
– **RECIPE** (594,000 phoneme tokens) — a collection of 1,300 cooking recipes.
– **ECONOMIC** (165,000 phoneme tokens) — short daily economic reports from December 3, 2001 till April 3, 2002.
– **FAIRYTALE** (172,000 phoneme tokens) — a children fairy tale book.

All data were stored in plain text files and a structural information, e.g. distinguishing recipe names from the ingredients, was ignored. Figure 1 illustrates the vocabulary growth of the corpora. The vocabulary sizes have been normalized, i. e. the value 60 at the point 30 means that the first 30% word tokens of the particular corpus contain 60% phonetically distinct words occurring in the whole corpus. It seems from the graph that the most limited corpus is WEATHER. The vocabulary growth of the ECONOM and FAIRYTALE corpora is nearly linear and it is questionable whether they can be considered limited.

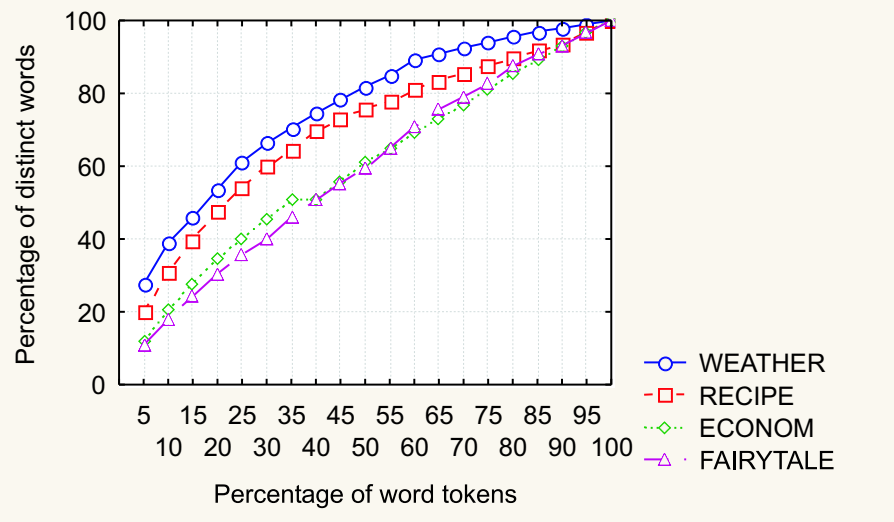

Fig. 1. The vocabulary of four inspected domains.

3 Feature Generation

All experiments have been made using the text analysis components of the Demosthenes speech synthesizer [2]. Typically, one feature vector corresponds to a phoneme-sized unit.

We start with a very simple feature set consisting of the phoneme label only. The elementary feature set is then refined in various ways. The newly added features may include textual features, e.g. position in the utterance, phonetic features like left and right neighbors or prosodic features, e.g. an estimated segment duration. For each feature set and each corpus, the appropriate list of

Feature set	WEATHER	RECIPE	ECONOM	FAIRYTALE
{phid}	1	1	1	1
{phid,previd}	0.92	0.99	0.99	0.98
{phid,previd,nextid}	0.41	0.73	0.33	0.62
{phid,inwordpos}	0.99	1	1	1
{phid,previd,inphrasepos}	0.89	0.98	0.97	0.97

Table 1. Coverage indices of several feature sets.

feature vectors is generated. A criterion is computed for each list determining the level of domain coverage of the given text corpus with respect to the given feature set.

The motivation behind this approach is the following. The utterances for recording the database are usually selected from the text corpus specific to the target domain. Text selection methods typically select the utterances (locally) maximizing the feature vector variability. However, it makes no sense to select utterances containing the maximum possible number of feature combinations, when the corpus **itself** does not include many combinations probably occurring in the target domain. When the corpus is recognized as insufficient for the requested feature set, it is possible to use one of the following two techniques.

First, more text data can be collected and a larger corpus built. This solution is often useful for limited domains, while it may be problematic for general purpose text corpora. The other possibility is to use a more general feature set for segment selection. To estimate whether the particular feature set is suitable for feature selection, another method must be used, e.g. the one presented in the next section.

4 Domain Coverage

So-called *coverage index* has been recently introduced as a criterion of domain coverage [6]. It is defined as a probability that all feature vectors in a randomly selected test sentence appeared in the training set. We used 90% of each corpus for training and 10% for evaluation. The results are summarized in Table 1. Feature types are similar to those used in [1].

Of course, the number of possible feature sets and feature values is practically infinite, and one needs to decide which set will be used for segment selection. We propose the following technique. The first step is to specify a treshold for the coverage index. All feature sets with coverage index below this treshold are considered inappropriate for segment selection. The second step is to generate some reference feature vectors containing as many details as possible. Typically, these vectors consist of the complete phonetic and prosodic specification of each unit. Finally, the last step is to estimate the unspecified feature variability for each feature set considered as a candidade set for segment selection. This variability is defined as an average distance between two undistinguishable vectors

of the candidade set. The distance computation is based on the feature vectors from the detailed feature set.

We have made such an analysis for the WEATHER domain. The results are shown in Figure 2.

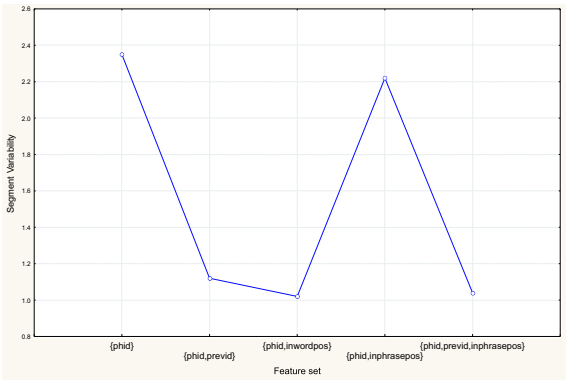

Fig. 2. An analysis of the expected variability of several feature sets for WEATHER domain. Feature sets with the coverage index less than 0.8 were not analyzed. The most suitable feature set is {phid,inwordpos} with variability 1.02.

5 Conclusions

The analysis of several restricted text corpora shows that they suffer from the extremely odd distribution of segmental features as well as the unrestricted corpora. Even the corpus of weather reports may not be considered limited, when it should be used to cover the domain with respect to a more detailed feature set. Thus, we provide here some techniques how to estimate the convenience of the particular feature set for the utterance selection.

References

1. R. Batůšek. A duration model for czech text-to-speech synthesis. In *Proceedings of Speech Prosody 2002*, Aix-en-Provence, France, Apr. 2002.
2. R. Batůšek and J. Dvořák. Text preprocessing for czech speech synthesis. In *Proceedings of TSD'99*, Pilsen, Czech Republic, Sept. 1999.
3. A. W. Black and K. Lenzo. Limited domain synthesis. In *Proceedings of ICSLP*, Beijing, China, 2000.
4. B. Möbius. Corpus-based speech synthesis: methods and challenges. Technical Report 4, Stuttgart University, 2000.
5. K. Stober, T. Portele, P. Wagner, and W. Hess. Synthesis by word concatenation. In *Proceedings of the Eurospeech'99*, pages 619–622, Budapest, Hungary, Sept. 1999.
6. J. P. H. van Santen. Combinatorial issues in text-to-speech synthesis. In *Proceedings of Eurospeech'99*, Budapest, Hungary, 1999.