

TEXT DESIGN FOR TTS SPEECH CORPUS BUILDING USING A MODIFIED GREEDY SELECTION

Baris Bozkurt*, Ozlem Ozturk*, Thierry Dutoit⁺,

Multitel ASBL*, TCTS Lab⁺, Faculté Polytechnique de Mons

Initialis Sci. Park, B-7000 Mons, Belgium, {bozkurt, dutoit}@tcts.fpms.ac.be,
oozlem@metu.edu.tr

Abstract

Speech corpora design is one of the key issues in building high quality text to speech synthesis systems. Often read speech is used since it seems to be the easiest way to obtain a recorded speech corpus with highest control of the content. The main topic of this study is designing text for recording read speech corpora for concatenative text to speech systems. We will discuss application of the greedy algorithm for text selection by proposing a new way of implementing it and comparing with the standard implementation. Additionally, a text corpus design for Turkish TTS is presented.

1. Introduction

Text design by random selection of sentences from various topics is one of the most frequently used techniques for speech corpora design. But corpus design is a long and difficult task and therefore some means of optimization are necessary. Especially for building open domain applications, optimization becomes a must since recording every possible speech event is practically impossible.

The coverage concept is very appropriate for reformulating the problem and searching for solutions [1]. The aim can be redefined as optimal design of a text corpus, which has highest coverage for a target synthesis domain. Coverage of a domain is defined via the concept of unit, therefore we discuss our basic unit choice in the following section. Greedy selection and the modified implementation will be discussed in section three. We also present tests comparing the standard greedy selection with the proposed modified version and an application for designing Turkish text corpus for recording text-to-speech voice database.

2. Unit Choice

In concatenative text to speech systems, half phonemes (demiphone), diphones and triphones are most frequently used units. The reason mainly stems on the concatenation problem; concatenation at stationary parts of speech chunks are less problematic for most of the phonemes. Additionally, for some phonemes, concatenation at phoneme boundary is possible without audible discontinuity (for example concatenation at the left phoneme boundary of a plosive). Therefore, using half phonemes as the basic units for unit selection and corpus design may be appropriate.

The advantage of using half phonemes is that the coverage problem is simplified. The number of half phonemes is quite small compared to the number of diphones and triphones. However not all diphones are synthesized with high quality with half phoneme concatenation [2] and a larger unit size is necessary for high quality synthetic speech. Many researchers use diphone as the basic unit for unit selection synthesis and it is to some extent affordable to build a text

corpus with high coverage of diphones. Therefore in our text corpus design, we will consider diphones as basic units.

In case of triphones, full coverage is impractical for most of the domains. A way to obtain high coverage with triphones as basic units is to target coverage of most frequently existing triphones. Figure 1 represents the coverage evolution with the number of frequency-sorted triphones for a large Turkish newspaper text. As it can be easily seen in the figure, a small portion of triphone space is very frequent. Figure 2 represents the coverage evolution with number of frequency-sorted triphones.

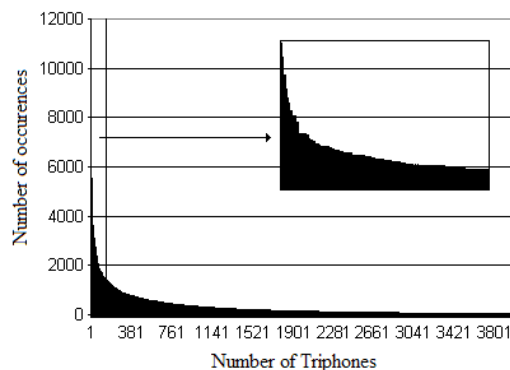


Figure 1: Frequency plot of first 4000 triphones sorted by their frequency (the small figure contains the plot for first 100 triphones).

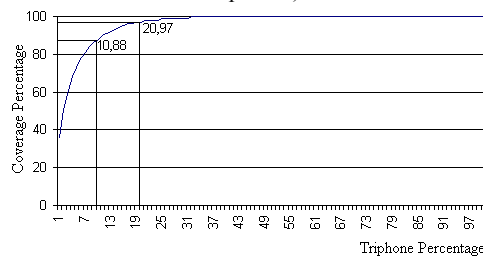


Figure 2: Evolution of cumulative coverage by triphone percentage (triphones sorted by frequency)

However, rare events are common in speech [3]. The vast majority of triphones are extremely rare but occurrence of rare triphones in speech is frequent. Beutnagel and Conkie [2] report that rare units are often preferred in their unit selection synthesis system and that the quality of synthesis is highly increased by including rare units in their database.

One other drawback of targeting coverage of most frequent triphones is the existence of important differences in triphone distribution between distinct text genres [4]. Similarly to the tests presented in that study, we have conducted comparative tests for five Turkish text corpora of different genres (play, novel, astronomy, history, news) collected on the internet. For each text corpus, 2000 most

frequent triphones were obtained and we checked to what extent the whole corpus is covered by these triphone subsets. Figure 3 represents the outputs of coverage by most frequent triphone approach. The bars in the figure appear by groups of five, each group corresponding to the percentage of coverage achieved when the 2000 most frequent triphones are selected from a particular corpus (for group 1, the triphones are selected from the first corpus, for group 2 from the second...). Each group of bars indicates that the highest coverage is obtained for that particular corpus from which the triphones are selected, and coverage for the other corpora is lower respectively. We have also compared the most frequent triphone sets obtained from these five corpora and checked to what percentage they appear in common. Table 1 represents the percentage of common triphones for all corpus couples.

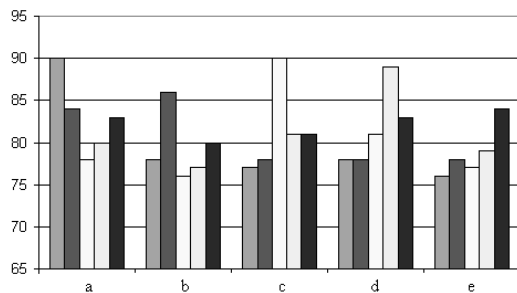


Figure 3 : Triphone coverage percentage the 2000 most frequent triphones obtained from a, b, c, d, and e respectively.

Table 1: Common triphone percentages

	Play	Novel	Astronomy	History	News
Play	100%	75%	68%	70%	73%
Novel	75%	100%	70%	72%	78%
Astronomy	68%	70%	100%	74%	74%
History	70%	72%	74%	100%	79%
News	73%	78%	74%	79%	100%

Additionally, when we compare all the five triphone subsets, we see that only 56% of the triphones are common to all subsets. From figure3 and table 1, we observe that by selecting the most frequent triphones we can achieve high coverage for a given text corpora but high coverage is not guaranteed for all types of new text corpora. Therefore, we conclude that covering most frequently occurring units approach is mostly appropriate for limited domain synthesis speech corpora construction where the synthesizer only needs to cope with text with a rather limited or constrained unit space.

As the unit size or the domain space increases, full coverage becomes harder to achieve. The corpus design is an optimization problem : that of recording the most appropriate set of sentences given practical limitations (like recording time, size of the data on disk, or available manpower in corpus construction). In this section, we discussed the basic phonetic unit choice in corpus design problem. Being limited to only a few hours of speech, we find it appropriate to use diphones as units for building an open domain targeted corpora and triphones for limited domain targeted corpora. Our reasoning for this choice is based on the affordability of full-diphone coverage (around 30 minutes of recording is sufficient for obtaining a corpus with full-diphone coverage for French or Turkish) and on the need to include rare

phonetic units, in addition to detailed discussions presented above.

3. Greedy selection

3.1. The standard Greedy Selection

The greedy algorithm has been used by many researchers for corpus design [1,5,6]. It is a simple iterative technique for constructing a subset of sentences from a large set of sentences to cover the largest unit space with the smallest number of sentences. Prior to the selection process, the target space to be covered needs to be defined by the unit definition, mainly the feature space of a unit. The phonetic content is the most basic feature but more features are desired in a high quality text-to-speech system, mainly some prosodic features (like isStressed, PhraseFinal...) [1].

The greedy selection process involves assigning costs to sentences according to the number of out-of-cover units and in-cover units in the sentence. At each iteration, the algorithm picks the most useful sentence to include in the subset, removes the sentence from the large set and updates target cover space and sentence costs in the large set. Iterating these steps until a termination criterion is reached, the subset is constructed. For comparison of greedy algorithm with various other techniques and combinational implementations, the reader is referred to [5].

While assigning costs to sentences, each unit in the sentence is labeled as out-of-cover or in-cover (i.e. a binary decision is made). The drawback of this technique is as follows; if the number of factors defining a unit is high, the unit set to-be-covered will be very large (due to the factorial structure). Therefore many units in sentences will be likely to be labeled as out-of-cover during the search. If the number of factors is low, then the unit set to-be-covered will be small and many units in sentences will be likely to be labeled as in-cover after a few tens of sentences are selected. Therefore a proper definition of target set to-be-covered is of high importance in using greedy selection for corpus construction, which needs choosing an appropriate feature list that defines a unit.

3.2. Modifying the sentence cost for maximum variability

In this study we propose a new method where we try not to define a discrete unit set to-be-covered but we target maximum variability of unit features in the subset to span largest possible unit space (i.e. we try to choose "orthogonal" sentences in terms of the units involved). To achieve this goal, we redefine unit costs in the greedy algorithm. For units (diphone-sized in our system), we do not assign binary costs but values within the range 0-1 defined by a weighted difference of the unit to its best match found in the subset (1 means no match, and 0 means perfect match). The method is almost the same as using a unit selection system for each unit of the candidate sentence to find the best match in the subset containing already selected sentences and obtain a normalized score of match for the candidate sentence. For assigning a cost to a unit in a sentence, first all matching units in the selected sentences subset are found, then a *MatchScore* is computed for each matched unit with the following function;

$$MatchScore = 1 - \sum_{n=1}^N w(n) * F(n) \quad (3)$$

$$\sum_{n=1}^N w(n) = 1 \quad (4)$$

Where $w(n)$ are weights and $F(n)$ are matching scores of two units for that particular feature. Features are assumed to have complete match or no match (for example stress vs. unstressed) except for the phonetic context match. For the phonetic context matching feature, we allow the user to associate values between 0-1 to similar types of phonemes, like 0.5 for a {p,t} class, for instance. Finally, the lowest *MatchScore* is assigned as the unit's cost. This way, in the selection process by the greedy algorithm, the uncovered units are most favored and then the units with lowest match in already selected sentences set. The sentence cost is calculated by normalizing the sum of unit costs by number of units in the sentence.

Choosing the appropriate weights appears to be a problem (as it is the case in a unit selection system). Ideally, the weights of this function shall match the weights in a unit selection-based speech synthesis system (mainly to the weights of the target cost function). If no unit selection system is available, weights can be either set to be equal or can be set by heuristics. It should be specified here that we propose only a selection procedure in our study, therefore we will not address the weight assignment problem in detail here.

Such an approach makes it possible to use as many features as wanted regardless of their space dimension without the risk of assigning high costs to most of the sentences (as explained in the section 3.1). By setting appropriate termination criteria, the algorithm may be adopted to obtain full coverage of a target space or largest coverage for a given maximum number of sentences to be selected (similarly, criteria based on sentence costs could be used). The advantage of this approach is clear when features which greatly enlarge the target space are to be used. For example, if the phonetic context match is to be used in the binary decision cost assignment approach, the discrete unit space becomes too large to cover (which will end up with very high number of units labeled to the same cost : 1). In the latter approach, it is possible to use this information to favor context variability of units while avoiding explosion of the unit space and typical cost assignment problems.

To be able to explain the cost assignment more clearly, we present an example below. Let's assume we want to calculate the cost of the diphone unit "a-s" in the context "...-p-a-s-a-..." and our feature set contains "left phonetic context ($F1$), right phonetic context($F2$), isstressed($F3$)" in addition to the diphone name($F4$), corresponding weights are equally 0.25 and the unit is unstressed. Let's also assume the diphone "a-s" already appears two times in the selected sentences subset; ...-k-a-s-e-..., in unstressed condition and ...-s-a-s-o-... in stressed condition. The cost calculated for these two matches will be as follows;

$$MatchScore(1)=1-w(1)*0.5-w(2)*0-w(3)*1-w(4)*1=0.375,$$

for ...-k-a-s-e-..., (assuming that {p,k} class was specified to have a score of 0.5 by the user)

$$MatchScore(2)=1-w(1)*0-w(2)*0-w(3)*0-w(4)*1= 0.75, \text{ for } \dots-s-a-s-o-\dots,$$

$MatchScore(1)$ is the lowest score, therefore it will be assigned as the unit cost in the sentence.

The algorithm automatically favors non-existing phonetic units by assigning $MatchScore=1$. The practical difference of this approach to that of a conventional greedy selection is obvious after a certain size is reached in the subset. In a conventional greedy selection process, we observe that most of the lately selected sentences are selected because they contain a few rare units and the sentence costs are quite low. In the case of the latter approach, sentence costs are

calculated by including costs of different realizations of the in-cover units, and sentence costs decline with a lower slope. The conventional greedy selection makes no preference for two sentences having the same number of diphones and 2 out-of-cover diphones, the latter approach will however try to favor the sentence according to highest variability of in-cover units compared to their matches in the cover.

The termination criterion we use is the number of sentences (which corresponds to practical limits) or the sentence cost. As the next step, out-of-cover diphones are extracted and manually designed sentences including out-of-cover diphones are added to the sentence list. At the end of this step, full diphone coverage is achieved.

4. Tests and Results

For evaluation of our technique, we have conducted a series of tests by running selections on two 2500-sentence size Turkish text corpora (the size of corpora was limited due to time limitations for testing). The first text was obtained directly from a Turkish novel (the first 2500 sentences were selected) and the second text was obtained from a very large text corpus (containing 115000 sentences from news, novels, scientific text) by greedy selection for multiple diphone coverage. For comparison, 500 sentences were selected using the two methods;

1- Classical greedy selection for diphone coverage

2- Greedy selection for diphone coverage with maximum variability of phonetic context (the proposed modification)

The basic unit size is diphone and only phonetic context variability is used as additional feature to test usability of a feature, which would explode the target unit space in a standard greedy selection. In the standard greedy selection, context variability is often assumed to be a by-product of selection for phonetic coverage.

The first results to point out is that the total size in terms of number of diphones of selected texts were very close; the subsets constructed with the second method had 0.4% and 0.8% more diphones respectively for the two text material. Although the difference is very small, we will use percentages in representation of the results.

For comparing the selected subsets, we performed a search for each unit of each unselected sentence and obtained the best possible match in the selected sentences. Then the units were classified according to the number of matching features for their best matches. Since what we look for is a context match, we also used context similarity groups according to place of articulation of the phonemes (for example {p,t} was one of the similarity groups). The results are represented in percentages in figure 4 and 5.

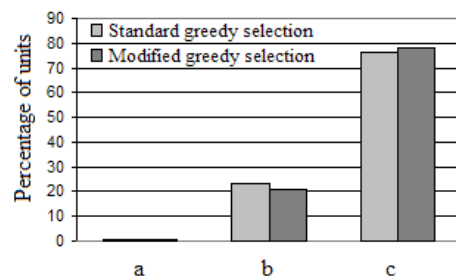


Figure 4 : Results of selection from novel text set : Percentage of diphones matching; a)with no context match, b)with a single context match, c)with two context matches.

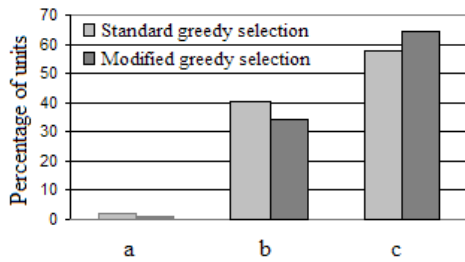


Figure 5 : Results of selection from phonetically rich text set : Percentage of diphones matching; a)with no context match, b)with a single context match, c)with two context matches.

Additionally, we have calculated the triphone coverage of the resultant two subsets for the second text material. The subset selected by the modified greedy selection contained 6.2% more distinct triphones than the subset selected with the standard greedy selection.

5. Discussions

As shown in the figures, with the modified version of the greedy selection, a subset which would cover the diphone unit space with more phonetic context matches (and therefore having larger triphone coverage) can be designed with the same amount of text selected. The two text material were quite different in terms of phonetic richness, the first material (randomly selected sentences from a novel) contained more restricted unit space. The performance improvement is better in the second case, which shows that method is more advantageous for designing the text of a text-to-speech corpus where phonetic richness is desired.

However, the computational load of the new method is very high compared to the standard method since at each iteration, a unit selection process is performed for all unselected sentences on the selected sentences. Therefore, for text selection from large text corpora, we find it more appropriate to perform two selection operations in sequence; first reducing the size of the very large text corpus to a smaller size corpus with standard greedy selection and then selecting the final set with the proposed method.

With these tests, we have presented how context variability could be improved in the designed text by letting the modified greedy algorithm use an additional feature: phonetic context. With additional features, the algorithm optimizes the coverage space according to given weights.

6. Design of the Turkish text

There exists a lack of shared materials and tools for Turkish speech synthesis. For this reason, we decided to apply our method to design a text corpus for Turkish and share it.

We followed the procedure below for building our Turkish speech corpora for synthesis.

- Text collection from internet (a total of 115000 sentences were reached)
- Preprocessing and rejection of sentences including some unexpected phoneme sequences (which we expect to stem from non-Turkish names and mistyping)
- Sentence selection for multiple diphone coverage with standard greedy algorithm (selecting 20000 sentences)

- A first manual correction of the text
- Modified greedy selection (selecting 5000 sentences)
- A more detailed second manual correction of text
- Modified greedy selection (selecting 2500 sentences)
- Manual design of special sentences for including uncovered diphones

We have a very simple natural language processing module for Turkish (preprocessing and text-to-phoneme transcription which utilizes the minimum set of phonemes for intelligible speech synthesis, namely the phonemes corresponding to the Turkish alphabet), therefore no prosodic features could be used for selecting sentences.

We share the resultant text material at the following web page : <<http://www.tcts.fpm.ac.be/~bozkurt/turkishtts/>>.

7. Conclusion

In this study, we have presented a simple way of modifying the greedy selection method to achieve maximum variability of units in a text design for speech corpora construction problem and improve the content richness of text corpus obtained by greedy selection. The method is used to create a corpus for Turkish text-to-speech synthesis and the corpus is shared freely on the internet.

8. Acknowledgements

The following freely available news text corpus was used as the biggest part of the initial text material;

<<ftp://cs.bilkent.edu.tr/pub/Turklang/corpus>>

We would like to thank Bilkent University (Ankara/Turkey), for making this material available.

9. References

- [1] van Santen, J P. H. and Buchsbaum, A. L., "Methods for optimal text selection ", *Proc. of Eurospeech*, p. 553-556, Rhodes, Greece, 1997.
- [2] Beutnagel, M. and Conkie, A., "Interaction of Units in a Unit Selection Database", *Proc. of Eurospeech*, Budapest, Hungary, 1999, p. 1063-1066.
- [3] Möbius, B., " Rare Events and Closed Domains: Two Delicate Concepts in Speech Synthesis ", *Proc. of 4th ISCA Speech Synthesis Workshop*, Scotland, 2001.
- [4] van Santen, J P. H., "Combinatorial issues in text-to-speech synthesis", *Proc. of Eurospeech*, p. 2511-2514, Rhodes, Greece, 1997.
- [5] François, H. and Boëffard, O., "The Greedy Algorithm and its Application to the Construction of a Continuous Speech Database", *Proc. of LREC*, Las Palmas de Gran Canaria, Spain, 2002.
- [6] François, H. and Boëffard, O., "Design of an Optimal Continuous Speech Database for Text-To-Speech Synthesis Considered as a Set Covering Problem ", *Proc. of Eurospeech*, Aalborg, Denmark, 2001.
- [7] Gauvain, J.F., Lamel, L.F., and Eskenazi, M., " Design Considerations and Text Selection for BREF, a Large French Read Speech Corpus ", *Proc. of ICSLP*, Kobe, Japan, 1990.
- [8] Zhu, W., Zhang, W., Shi, Q., Chen, F., Li, H., Ma, X. and Shen, L., "Corpus Building for Data-Driven TTS System", *Proc. of the IEEE TTS 2002 Workshop*, Santa Monica, USA 2002.