Corpus design based on the Kullback-Leibler divergence for Text-To-Speech synthesis application

Aleksandra Krul [1], *Géraldine Damnati* [1], *François Yvon* [2], *Thierry Moudenc* [1]

[1] France Télécom R&D Division, TECH/SSTP
2, avenue Pierre Marzin, 22307 Lannion Cedex, France
{aleksandra.krul,geraldine.damnati,thierry.moudenc}@francetelecom.com

[2] GET/ENST and CNRS/LTCI
46, rue Barrault, 75624 Paris Cedex 13, France
yvon@enst.fr

Abstract

This paper presents a corpus design method for Text-To-Speech (TTS) synthesis application. The aim of this method is to build a corpus whose unit distribution approximates a given target distribution. Corpus selection can be expressed as a set covering problem, which is known to be NP-complete: we therefore resort to a heuristic approach, based on greedy algorithm. We propose the Kullback-Leibler divergence to guide the iterative selection of candidate sentences: indeed, this criterion gives the possibility to control the unit distribution at each step of the algorithm. We first show how to efficiently update, in an incremental manner, this criterion. We then present and discuss experimental results, where our selection algorithm is compared, for various unit sets, with alternative selection criteria.

Index Terms: speech synthesis, corpus design, Kullback-Leibler divergence.

1. Introduction

Current Text-To-Speech systems are based on concatenative methods. Such systems use a large database of pre-recorded speech from which acoustic units are selected for concatenation. The quality of the system is strongly related to the quality of the recorded textual corpus. Therefore, the corpus construction is a crucial step in building a TTS system.

Corpus design can be formulated as a set covering problem [1, 2, 3, 4]. The target set C is the units to be covered; each sentence in the textual corpus is also a set of units, and the corpus selection problem consists in finding a minimum size set of sentences whose union contains all the units in C. This problem is known to be NP-complete [5], and heuristic approaches have to be considered, such as greedy algorithms. The greedy approach for the set covering problem incrementally builds a corpus by selecting, at each step, the most useful sentence from a large textual corpus, according to a criterion which assesses the benefits of including a new sentence. At each iteration, a score is assigned to each sentence. The sentence with the highest score is selected and removed from the set of candidates, and the set of units to be covered is updated. Other methods for corpus construction have been proposed: for instance, [6] propose the pair exchange method; a spitting method is also considered in [2].

In the context of speech synthesis, the main goal of the corpus construction is to get an optimal coverage of speech units. The selection algorithm therefore aims at finding a subset of sentences which achieves the desired unit coverage condition. Units to be covered can be either diphones, triphones or syllables [2, 3, 4]. However, to ensure a high synthesis quality, it is preferable to consider not only the phonetic nature of the units, but also features which characterize the units: length, stress, syntactic, lexical and phonetic context, etc.

The performance of the aforementioned methods depends on the criteria used to compute the sentence score. Usually, criteria take into account the number of in-cover and out-of cover units. In order to control the sentence length, the total number of units is generally taken into account in the various criteria. Several criteria were presented and evaluated in [2], for example the criteria based on the presence of units useful to the coverage as well as on the presence of rare units in the candidate sentence.

There are different ways to evaluate the score of the candidate sentences. In order to reach coverage of units the basic score computation consists in normalizing the number of out-of-cover units of the candidate sentence by the total number of units in the sentence. To favour rare units, the frequency of units computed on the initial corpus can also be introduced in the sentence score computation. In order to achieve maximum variability of units in the selected subset, [3] redefine the sentence score in the greedy algorithm, taking into account the scores of different realizations of units.

Choosing an optimal coverage depends on the application for which the corpus is designed [3, 7, 1]. For open domain synthesis, it is important to include frequently occurring units as well as rare units. A full coverage has to be obtained at least for the basic units (typically diphones). The coverage that reflects a given domain is more suitable for limited domain synthesis: the most frequent units in the domain have to be included in the corpus. As a consequence, the database is likely to be smaller.

In this paper, we propose an alternative approach to corpus design, based on the Kullback-Leibler divergence. The main idea of this method is to build a textual corpus whose unit distribution is close to an *a priori* distribution. A similar method was used in adaptation text design for speech recognition [8]. During the corpus construction, the proposed criterion compares the obtained

units distribution with the target distribution. In section 2, we introduce the Kullback-Leibler measure and we present an efficient implementation of our unit selection algorithm. We first consider the case where the target distribution is uniform, before explaining how it can be generalized to other kinds of target distributions. Experimental results are presented in section 3, where we compare our criterion with two standard criteria.

2. Method

2.1. The Kullback-Leibler divergence

The KL divergence [9] is a measure which assesses the similarity between two probability distributions. It is defined as:

$$D(P \parallel Q) = \sum_{i=1}^{t} p_i \log \frac{p_i}{q_i} \tag{1}$$

where P and Q are two discrete probability distributions.

The properties of this measure are the following. The divergence is positive or equal to zero. The two probability distributions are identical if and only if the KL divergence is null.

2.2. Sentence selection based on the KL divergence

2.2.1. Algorithm

For text corpus design, a greedy algorithm is used. At each iteration of the algorithm, the sentence which minimizes the KL divergence to the target distribution is picked. Let Q denote the target distribution, and $S = \{s_1, s_2, \ldots, s_l\}$ be the corpus from which the sentences are selected. The corpus built after m iterations of the algorithm is denoted $S'_m = \{s'_1, s'_2, \ldots, s'_m\}$, where $m \leq l$. Let s be a candidate sentence at iteration $m+1$: n_i is the number of occurrences of unit i in $S'_m \cup \{s\}$, and N is the total number of units ($N = \sum_i n_i$). The probability of unit i is simply computed as its relative frequency $p_i = \frac{n_i}{N}$. The score of the candidate sentence s is then:

$$D(P \parallel Q) = \sum_{i, n_i \neq 0} \frac{n_i}{N} \big(\log \frac{n_i}{N} - \log q_i \big) \tag{2}$$

Taking $0 \log \frac{0}{q} = 0$ allows to perform the summation over the complete set of units.

At each step of algorithm 1, the sentence which minimizes the KL divergence is added to the set of previously selected sentences. The newly formed set of sentences has the minimum KL divergence between its unit distribution and the target unit distribution. The algorithm stops after picking L sentences, where L is a predefined limit.

A naive implementation of this algorithm requires to repeatedly estimate the probability distributions P_{jk} for each candidate sentence, and to compute the KL divergence with Q. The overall complexity of these computations is thus $O(L \times |S| \times |C|)$, which is prohibitive when $|S|$ is large.

It is however possible to significantly improve over this naive complexity, by incrementally updating the KL divergence as follows. Between iterations j and $j+1$, the algorithm adds a set of units V, which only contains those units which occur in the candidate sentence. Let N_j and N_{j+1} denote respectively the total number of units in the corpus built at iteration j and the candidate corpus at iteration $j+1$; P_j and P_{j+1} denote the corresponding probability distributions. Let finally n_i^j and n_i^{j+1} be respectively

Algorithm 1 Sentence selection based on the KL divergence

```
Set the target distribution Q
S'_0 = ∅
for j = 1 to L do
    D_min ← +∞
    for Every sentence s_k ∈ S \ S'_{j-1} do
        F_jk = S'_{j-1} ∪ {s_k}
        Estimate the probability distribution P_jk on F_jk
        Compute D(P_jk ‖ Q)
        if D(P_jk ‖ Q) < D_min then
            D_min ← D(P_jk ‖ Q)
            s_best ← s_k
        end if
    end for
    S'_j ← S'_{j-1} ∪ {s_best}
end for
```

the number of occurrences of unit i at iteration j and $j+1$. α_i is defined as $\log q_i$. Simple arithmetic computation show that the KL divergence can be decomposed as follows:

$$D(P_{j+1} \parallel Q) = A_{j+1} + B_{j+1} \tag{3}$$

where A_{j+1} is defined as:

$$A_{j+1} = \frac{N_j}{N_{j+1}} \Big[D(P_j \parallel Q) + \log \frac{N_j}{N_{j+1}} \Big] \tag{4}$$

and B_{j+1} is defined as:

$$\begin{aligned} B_{j+1} &= \sum_{i \in V} \frac{n_i^{j+1}}{N_{j+1}} \Big(\log \frac{n_i^{j+1}}{N_{j+1}} - \alpha_i \Big) \\ &- \sum_{i \in V} \frac{n_i^{j}}{N_{j+1}} \Big(\log \frac{n_i^{j}}{N_{j+1}} - \alpha_i \Big) \end{aligned} \tag{5}$$

In this decomposition, A_{j+1} is related to $D(P_j \parallel Q)$ through a simple affine function. The computation of B_{j+1} only implies a summation over the set of units occurring in the candidate sentence, which considerably reduces the computation cost.

In this study, we take as target distribution the uniform distribution. This will be discussed in section 4. It is however important to realize that our algorithm is able to accommodate a variety of alternative target distributions: as is clear from equation 5, the complexity of our procedure does not depend on the choice of Q.

2.2.2. Coverage

The condition to achieve a desired coverage is only indirectly introduced in our criterion. Indeed, the algorithm attempts to include all distinct units. However, as it selects entire sentences, the resulting distribution inevitably reflects the characteristics of the original distribution. It is not difficult to see that our method does not make it possible to quickly achieve a full coverage of units. Actually, after a certain point, some sentences that contain no out-of-cover units can happen to be better candidates than sentences with out-of-cover units but occurring for example with already highly represented other units. We have therefore included the following constraint, which has the additional benefit to speed up the selection process: at each iteration, we only evaluate sentences which contain at least one "new" unit. Once the full unit coverage is achieved, the full selection process resumes, evaluating all available sentences.

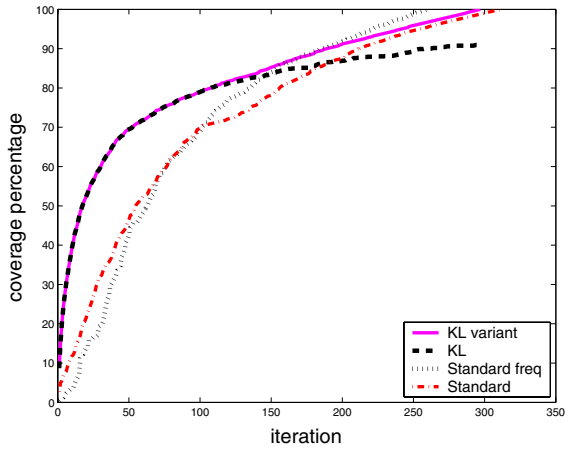

Figure 1: *Evolution of diphone coverage.*

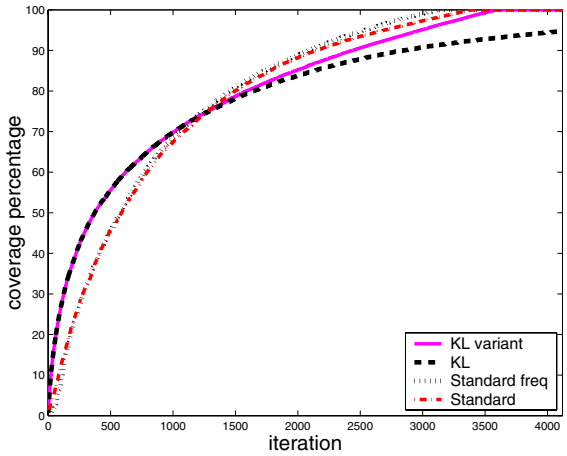

Figure 2: *Evolution of triphone coverage.*

3. Experimental Results

3.1. Data

The corpus used for this experiment contains about 7000 sentences collected from the "Le Monde" newspaper. About hundred sentences used in the vocal applications are also included in the corpus. This corpus was built in order to create the speech database for open domain synthesis application. It has been design with the aim of covering all distinct diphones, 90% of contextual diphones and 80% of distinct triphones observed in the initial exerpt from "Le Monde". We assume that the corpus is phonetically balanced. The maximal sentence length is 27 words. There are 1170 distinct diphones and 14907 distinct triphones in the corpus. As we want to compare the behaviour of different criteria, and as we only consider basic units (typically diphones and triphones) we presume that the corpus size is sufficient.

3.2. Criteria comparison

In this section we compare our criterion and its modified version (cf. 2.2.2) with two standard criteria. The first standard criterion is computed as follows: the number of "new" distinct units present in the candidate sentence is normalized by the total number of units in the sentence. The second standard criteria is similar to the first one, with the difference that it favours rare units. The score of each new unit in the candidate sentence is thus weighted by the inverse of the unit frequency computed on the initial corpus. As we consider basic units we target full distinct unit coverage. We first examine the diphones and triphones coverage obtained at each iteration of the sentence selection process. The evolution of coverage is displayed on figure 1 for diphones and on figure 2 for triphones.

The full diphone coverage is quickly attained by standard methods ("Standard" and "Standard freq") and by our method ("KL variant"). However, the full coverage is only achieved at the end of the process using the algorithm 1. This phenomena was observed in [4]. The KL method prefers to pick sentences which make the built distribution more balanced rather than sentences which contain "new" units: in fact, sentences which contain the out-of-cover units do not necessarily minimize the KL divergence and are not likely to be selected. By introducing a constraint on the candidates the algorithm can achieve quickly the full coverage.

In both cases (diphones and triphones), the coverage rate increases faster with KL-based criteria. This is encouraging for further studies where we wish to include more constraints on units and for those applications where corpus size has to be limited.

In order to go further in the comparison, we have focused our study on the built corpus when the full unit coverage is achieved with two criteria. Let C_{KL}^{dip} be the built corpus using modified KL method related to diphones and C_{freq}^{dip} be the corpus built with the "Standard freq" method. C_{KL}^{trip} and C_{KL}^{trip} are the corresponding corpora with the full triphones coverage. Table 1 presents some statistics computed on the selected corpora when the full coverage is reached. As can be seen from these numbers, the modified KL algorithm picks long sentences in order to include a large number of distinct units. As for the "Standard freq" method, short sentences with rare units are selected in priority.

Table 1: State of built corpora with the full unit coverage.

	Number of sentences	Number of units	Average number of units per sentence
C_{KL}^{dip}	296	9150	31,0
C_{freq}^{dip}	260	6941	26,7
C_{KL}^{trip}	3586	121861	34,0
C_{freq}^{trip}	3131	106244	34,0

For C_{KL}^{dip} the total occurrence number of units is 31% higher than for C_{freq}^{dip}. This difference goes down to 14,7% for corpora related to triphones, where the average sentence lengths are comparable.

4. Discussion

In this study, we have taken as target distribution the uniform distribution: our algorithm will thus try to find a set of sentences such that the unit distribution in this corpus has a maximum entropy. At first sight, this may look quite counter intuitive, and it might seem more natural to target a distribution which is close to the "natural"

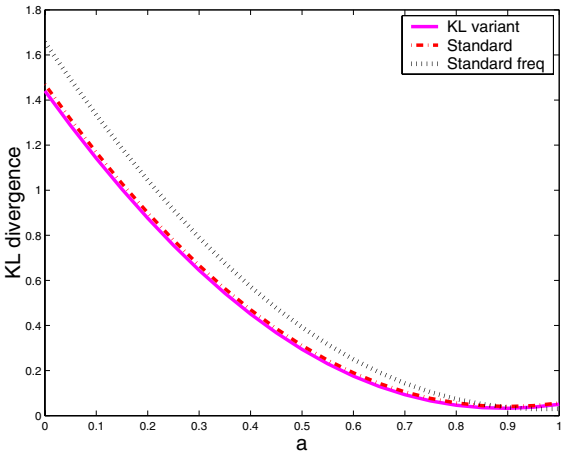

Figure 3: *KL divergence to $q_i(a)$ distribution.*

distribution of units, estimated from a large corpus.

However, in the corpus from which the sentences are selected, units tend to be distributed according a Zipf law [10]: a small number of units are very frequent, while a large number of units are very rare. Any selection algorithm picking entire sentence is thus very likely to get a complete coverage of the frequent units. By targeting the uniform distribution, the algorithm tries to reduce the number of frequent unit while maximizing the selection of rare units. This behaviour of our algorithm is evidenced by the following experiment.

Let f_i be the relative frequency of unit i in the initial corpus. Let $q_i(a)$ be defined for $0 \leq a \leq 1$ as:

$$q_i(a) = \frac{f_i^a}{\sum_i f_i^a} \tag{6}$$

This distribution corresponds to the case where all unit frequencies are leveraged. Frequent units are more affected in absolute. We are interested here in estimating the KL-divergence between the distribution of units estimated on our automatically selected corpora and this $q_i(a)$ distribution for different values of a. The units under consideration are triphones and we are making evaluations on the corpora selected at the iteration when the full triphone coverage is achieved. The question is: what is the value of a for which the built distribution is closer to $q_i(a)$?

For $a = 0$ the KL divergence to the uniform distribution is evaluated; for $a = 1$ the KL divergence to the initial unit distribution is computed.

As can be seen on figure 3, for all methods the resulting unit distribution is much closer to the initial unit distribution than to the uniform distribution. As whole sentences are selected, the unit distribution can not be totally flattened: units tend to remain distributed according a Zipf law. The distribution obtained with our method is however flatter than those obtained with the other methods (a minimum can be observed for a value of a equal to 0.9).

This corpus design method remains to be experimented with larger corpora, as the results might be different when we have more sentences to choose from: the selected corpus size will undoubtedly be much smaller than the current selected corpus (modulo the initial corpora sizes). In any case, further investigations should be carried out in order to try to force the distributions to be flatter in a reasonable way.

5. Conclusions

In this study, we have presented a method based on the Kullback-Leibler divergence for corpus design. The proposed criterion gives the possibility to globally control the unit distribution in the built corpus. We have also proposed an efficient implementation of this method which incrementally update the KL divergence in the sentence selection process. As a consequence, the computation cost of the method is reduced.

For this study we have targeted the uniform unit distribution but the advantage of our approach is that the proposed algorithm is flexible and it is able to accommodate different distributions which may prove more for domain specific TTS synthesis applications. The textual corpus defining a domain has specific unit distribution. For text adaptation, sentences whose unit distribution resembles the task distribution have to be selected. The adaptation of the textual corpus to various distributions is easy to implement: what is only required is to obtain Q from a given domain specific corpus and to set it as the target distribution in our algorithm.

Our future plans also include exploring this method on larger corpora and examining other types of units, for instance contextual units. Finally, the speech synthesis quality evaluation has to be performed.

6. References

[1] J.P.H van Santen and A. L. Buchsbaum, "Methods for Optimal Text Selection," in 5^{th} *European Conf. on Speech Communication and Technology (Eurospeech)*, Rhodes, Greece, Sept. 1997, pp. 553–556.

[2] H. François, *Synthèse de la parole par concaténation d'unités acoustiques : construction et exploitation d'une base de parole continue*, Ph.D. thesis, Université de Rennes 1, 2002.

[3] B. Boozkurt, O. Ozturk, and T. Dutoit, "Text design for TTS speech corpus building using a modified greedy selection," in 8^{th} *European Conf. on Speech Communication and Technology (Eurospeech)*, Geneva, Switzerland, Sept. 2003, pp. 277–280.

[4] Y. Feng, "Selection of text script for text-to-speech synthsesis," in 5^{th} *IASTED Int. Conf. SIGNAL AND IMAGE PROCESSSING*, Honolulu, Hawaii, USA, Aug. 2003.

[5] Michael R. Garey and David S. Johnson, *Computers and intractability: a guide to the theory of NP-completeness*, W.H Freeman and company, New York, 1979.

[6] H. Kawai, S. Yamamoto, N. Higuchi, and T. Shimizu, "A Design Method of Speech Corpus for Text-to-Speech Synthesis Taking Account of Prosody," in 6^{th} *Int. Conf. on Spoken Language Processing (ICSLP)*, Beijing, China, Sept. 2000, pp. 277–280.

[7] A. W. Black and K. A. Lenzo, "Optimal Data Selection for Unit Selection Synthesis," in 4^{rd} *ESCA Workshop on Speech Synthesis*, Scotland, 2001.

[8] X. Cui and A. Alwan, "Efficient adaptation text design based on the Kullback-Leibler measure," in *Int. Conf. on Acoustics, Signal and Speech Processing*, Orlando, USA, Sept. 2002.

[9] T.M Cover and J.A Thomas, *Elements of Information Theory*, Wiley Series in Telecommunications, 1991.

[10] R.H. Baayen, *Word Frequency Distributions*, Kluwer Academic Publishers, 2001.