Comparing performance of Different Set-Covering Strategies for Linguistic Content Optimization in Speech Corpora

Nelly Barbot, Olivier Boeffard and Arnaud Delhay

IRISA - University of Rennes 1, Lannion, France
{Nelly.Barbot,Olivier.Boeffard, Arnaud.Delhay}@irisa.fr

Abstract

Set covering algorithms are efficient tools for solving an optimal linguistic corpus reduction. The optimality of such a process is directly related to the descriptive features of the sentences of a reference corpus. This article suggests to verify experimentally the behaviour of three algorithms, a greedy approach and a lagrangian relaxation based one giving importance to rare events and a third one considering the Kullback-Liebler divergence between a reference and the ongoing distribution of events. The analysis of the content of the reduced corpora shows that the both first approaches stay the most effective to compress a corpus while guaranteeing a minimal content. The variant which minimises the Kullback-Liebler divergence guarantees a distribution of events close to a reference distribution as expected; however, the price for this solution is a much more important corpus. In the proposed experiments, we have also evaluated a mixed-approach considering a random complement to the smallest coverings.

Keywords: linguistic corpus reduction, greedy algorithms, Kullback-Leibler divergence, random selection

1. Introduction

Most Text-to-Speech (TTS) synthesis systems rely on techniques that select and concatenate recorded speech segments, such as diphones and more generally n-phones (a n-phone being defined as a sequence of n consecutive phones). The quality of synthetic speech depends on numerous factors, but construction of the speech corpus is a crucial task. A speech corpus must indeed offer the widest phonological variety with a minimum size. The units used during the synthesis stage are not known beforehand, and their length could vary. Additionally, recording a corpus must be done in a minimal time in order to guarantee a consistent quality of the speakers' voice, and to enhance concatenation of two speech segments recorded at distant times.

The elaboration of the linguistic content of such a speech corpus, covering an entire predefined set of phonological units and requiring minimal recording time, is akin to a *Set-Covering Problem -SCP-* which is NP-hard (Garey and Johnson, 1979). Moreover, the events under consideration naturally have a heavy-tailed distribution (few elements are very frequent and a lot of them are very rare) and several variants of the same event are required to guarantee an acceptable quality.

Given the size of the search space of these problems, it is necessary to work out sub-optimal or heuristic algorithmic solutions. The most frequently used algorithm in speech processing is the greedy agglomeration. This iterative algorithm chooses at each step a sentence corresponding to the highest score which quantifies a sentence contribution to the iterated covering. (Gauvain et al., 1990) applies this greedy strategy to build a database for a speech recognition task thanks to hierarchically organized covering attributes. (Van Santen and Buchsbaum, 1997) studies greedy variants of text selection, varying the required unit nature (diphone, duration components, etc.) and the sentence score function according to the application. In (François and Boeffard, 2002), several combinations of greedy algorithms are tested in order to constitute a covering of diphones. According to those works, the best strategy lies in applying a greedy agglomeration phase first, followed by a reverse greedy phase - or *spitting* phase. This algorithm is named *GreedyAS - Greedy algorithm based on Agglomeration and Spitting phases*. As an alternative to the greedy method, (Chevelu et al., 2008) suggests an algorithm based on lagrangian relaxation, called *LamSCP - Lagrangian based algorithm for multi-represented SCP*, to cover 2 and 3-phonemes (sequence of 2 and 3 phonemes). The main advantage of *LamSCP* is to provide a lower bound for the minimal length of a covering, thereby allowing for an evaluation of the absolute quality of the results.

Still based on a greedy approach, (Krul et al., 2007) suggests an interesting solution: the optimization criterion for reducing the covering size is based on minimizing the Kullback-Liebler divergence for a distribution of events (2 or 3-phones) between the reduced corpus and a reference distribution. This approach will be referred to as *GreedyAKL*, *Agglomeration Greedy algorithm based on Kullback-Liebler divergence*.

Under the title "How (Not) to select your voice corpus: random selection vs. phonologically balanced", (Lambert et al., 2007) presents experiments that show the performance of a reduced-based corpus with covering criteria compared to a random selection of sentences. The experiments are based on the corpus used in the Blizzard Challenge. Considering subjective evaluation results, a synthesized sentence built from a random corpus is better rated than a sentence built from a phonetically-balanced corpus elaborated through a greedy approach.

That result, somehow unsettling, allows for a revision of the criteria used in the selection of "optimal" corpora used in speech synthesis, with respect to the quality of the synthetic signal. One of the main goals of this paper is to seek a compromise between the idea of a corpus defined through minimum coverage algorithms, and a random selection of

sentences to build an equivalent corpus. Although the validation of the outcome with a subjective testing is an interesting option working towards this goal, we also wish to keep our analysis as general as possible. Thus, we suggested phonological and linguistic criteria to qualify the contents of those built corpora. This phonological level is required by the Blizzard Challenge; moreover, these criteria will give us a more general framework that will prevent us from drawing conclusions that would be applicable only in the context of a specific task.

The paper is organized as follows. Section 2. presents the algorithms involved in the experiments to design reduced corpora. Section 3. introduces the evaluation criteria of the reduced corpora. Section 4. describes the experimental methodology allowing for the comparison of various algorithms and section 5. discusses the results.

2. Systems for Corpus Reduction

In this section, we briefly introduce *GreedyAS*, *LamSCP* and *GreedyKL*. Their purpose is to extract from a large reference corpus G a subset of sentences C which contains at least k instances of each attribute u_i, with the attribute set $\{u_1, \ldots, u_n\}$ being defined beforehand. If G does not contain at least k instances of u_i, C has to contain all of its intances present in G.

First, *GreedyAS* aims to derive the shortest possible solution in terms of elocution duration - approximated by the number of phone instances. This method is composed first of an agglomerative greedy process which provides a covering C: initially $C = \emptyset$ and each step adds to C the sentence s with the highest score, until the covering is reached. The score of s corresponds to the number of its attributes that are missing in the ongoing covering, divided by its length. A spitting greedy algorithm is then applied: at each step, the longest redundant sentence of C is excluded of the covering.

Secondly, *LamSCP* has the objectives of minimizing the covering length and deriving a lower bound for the optimal solution size. This lower bound may be not reachable, but it provides useful information. *LamSCP* uses several heuristics based on lagrangian relaxation properties. A large pruning of the search space is then done and several greedy procedures based on a lagrangian cost function are carried out to obtain coverings. From the best of these coverings, a subset of sentences is selected according to the score function. The residual sub-problem is then processed similarly, and the pruning of the search space is updated. The algorithm is stopped as soon as the residual problem is empty or when the associated lower bound indicates that the ongoing covering cannot be better than the current best solution.

At last, *GreedyAKL* builds a covering C of which the 2-phoneme distribution is as close as possible to the ones in G in terms of Kullback-Liebler divergence. The main strategy is an agglomerative greedy one: each step selects the sentence containing missing attributes in the ongoing covering and contributing the least to the Kullback Liebler divergence.

3. Notations and Evaluation Criteria

Our main experimental goal concerns evaluation of the different ways to design a corpus of reasonable size with regard to its phonological content. It is a matter of finding a compromise solution between the shortest possible corpus and the largest possible linguistic and phonological contents indicated by a set of attributes to cover.

The three algorithms, *LamSCP*, *GreedyAS* and *GreedyAKL*, are carried out to reduce the corpus *Gutenberg* (Hart, 2003). *Gutenberg*, denoted G, contains 53,996 sentences in English, covering 57 distinct phonemes (considering two levels of stress) and its size is 1,539,735 phones (*i.e.* "instances" of phonemes). Its phonological content is detailed in Table 1. Every phoneme and 2-phoneme in *Gutenberg* are collected to define the set of attributes to cover. A study based on this corpus is interesting in view of its wide spreading, notably in the context of the Blizzard challenge.

unit name	unit number
phoneme	57
2-phoneme	1,955
3-phoneme	27,477
4-phoneme	149,435
5-phoneme	378,280

Table 1: Phonological content of Gutenberg

In order to clarify the reading of the experimental cases, we introduce the following notations. Let S be a finite set of integers, a k-covering of S-phonemes denotes a covering containing at least k instances of every l-phoneme, for every $l \in S$. Such a covering obtained by *LamSCP* is called $LamSCP_S^{(k)}$. Similar notations are used for *GreedyAS* and *GreedyAKL*. The coverings are evaluated according the following criteria: for $l \in \{2, \ldots, 5\}$

(1) l**-phoneme/#**: number of distinct $\{1, \ldots, l\}$-phonemes,

(2) l**-phoneme/div**: Kullback-Liebler divergence between the distributions of $\{1, \ldots, l\}$-phonemes of G and the reduced corpus C,

(3) l**-phoneme/cov**: representativeness of $\{1, \ldots, l\}$-phonemes of C relatively to their frequencies in G.

Some of these criteria are directly related to covering attributes, and others not. For instance, the criterion indicating the representativeness of 2-phonemes of the reduced corpus relatively to their frequencies in the reference corpus depends on the constraint to cover the entire set of 2-phonemes. In this case, the expected value for this criterion will be 100%. It may be interesting to observe the consequence of an algorithmic choice on a feature that the algorithm does not aim to optimize: it can be a matter, as an example, of checking the representativeness of 3- or 4-phonemes. Indeed, it is illusive to try to build coverings based on constraints that are too complex. For example, in (Chevelu et al., 2008), the $\{1, 2, 3\}$-phoneme coverings are on average 17 times longer than the $\{1, 2\}$-phoneme coverings.

For the calculation of 2- and 3-**phoneme/div** values, the involved distributions are smoothed using Turing-Good method (Katz, 1987). 4- and 5-**phoneme/div** values are not derived, because the necessary smoothing is too drastic.

4. Description of the Experiments

We distinguish two experimental purposes: the first one is to analyse the quality and the stability of the various solutions to reduce G and the second one consists in evaluating the impact of the addition of randomly chosen sentences to the coverings in order to reach a given size. 95% confidence intervals are calcultated with a bootstrap method and are indicated by the symbol $\pm$.

4.1. Experiment 1 - Stability of Coverings

This preliminary experiment, called experiment 1, aims at showing the ability of each algorithm to provide stable solutions with regards to their size and the analysing criteria. Indeed, one of the difficulties of the greedy methodology used in *GreedyAS* and *GreedyAKL* is that their associated score functions have discrete values and several sentences can have the same score. In our implementation, the greedy algorithms choose the first coming sentence out of those that have the best current score. We would like to measure the influence of this *random* choice on the stability of the results. *LamSCP* uses heuristics that make a pre-selection from among sentences according to their lagrangian costs, which are continuous real-value functions (with respect to the lagrangian multipliers). Moreover, *LamSCP* uses greedy strategies based on lagrangian costs, which may be therefore more discriminant than the function. Consequently, the obtained coverings depend on the order of the sentences in the corpus G. A simple solution to evaluate the stability consists in proceeding an important amount of experiments on the same *SCP* instance by randomly modifying the sentence ordering in the initial corpus, at the beginning of each experiment. Considering the computation time, we choose to carry out at least 59 times a 1-covering of the $\{1,2\}$-phonemes on G for each algorithm.

4.2. Experiment 2 - Phonological Content of the Coverings

In order to keep corpora with a reasonable size, experiment 2 aims to compare the phonological content of the 1-coverings of $\{1,2\}$-phonemes with the one of the 2-coverings of $\{1,2\}$-phonemes for each algorithm, and with the 1-coverings randomly augmented at the associated 2-covering level. Since experiment 1 has shown a good robustness of the three algorithms to sentence ordering in G (see below paragraph 5.1.), only one 1-covering and 2-covering of $\{1,2\}$-phonemes have been calculated by each algorithm. For each 1-covering of $\{1,2\}$-phonemes, $LamSCP^{(1)}_{\{1,2\}}$, $GreedyAS^{(1)}_{\{1,2\}}$ and $GreedyAKL^{(1)}_{\{1,2\}}$, 100 randomly augmented corpora have been built, in order to calculate for each criterion its average value and its confidence interval.

4.3. Experiment 3 - Comparison of Reduced Corpora for a Given Length

This last experiment, called experiment 3, aims to compare reduced corpora for a given size. These corpora are built from $LamSCP^{(1)}_{\{1,2\}}$ and $GreedyAS^{(1)}_{\{1,2\}}$. Randomly chosen sentences are added in order to reach a limit of N phones. Given the covering sizes previously obtained, we choose $N \in \{20,000; 25,000; 30,000\}$. These corpora are compared with an equivalent-sized corpus built entirely in a random way, called rand(N). For rand(N) and each random complement, 100 instances have been built.

5. Results and Discussion

5.1. Experiment 1

The results of the first experiment are given in Table 2. In the following, we first comment the results about stability. Then we make some comments on the behaviour of the algorithms on some remarkable points.

Except for the criterion 2-**phoneme/div**, the relative standard deviation values are low and indicate a great stability of the size and phonological content of the solutions derived by each algorithm. Only the criteria 2- and 3-**phoneme/div** are more sensitive to the sentence ordering in G. It may be due to the very small size of the coverings (less than 3% of G) and the low 2- and 3-**phoneme/div** values.

Moreover, among the 59 experiments, the best lower bound of the minimal covering size, derived by *LamSCP*, indicates that the optimal solution contains at least 13,376 phones. Therefore, *LamSCP* and *GreedyAS* have good capacity to calculate a solution close in size to the optimal one: an average of 13,458 phones by *LamSCP* and 14,914 by *GreedyAS*. For a solution 10% longer than that of *LamSCP*, *GreedyAS* improves the number of distinct 3- to 5-phonemes from 7 to 8%, with more rare units, the relative covering rising only of 1%. Concerning the other criteria, the results for solutions of *LamSCP* and *GreedyAS* are similar.

Let us note that the k-**phoneme/div** score for *GreedyAKL* is far better than for the two other algorithms. This result was expected because the minimisation of the Kullback-Liebler divergence between the 2-phoneme distribution in the reduced corpus and the one in G is the optimisation criterion of this algorithm. The other phonological criteria have also better values for *GreedyAKL* than for *LamSCP* and *GreedyAS*, since *GreedyAKL* provides solutions at least twice as long as the ones obtained by *LamSCP* or *GreedyAS*. The cause is probably due to the fact that some frequent units in the original corpus have to be covered with regards to their distribution, then more often than in a simple 1-covering. Instead of choosing preferably sentences containing several uncovered units at each iteration, *GreedyAKL* tends to select more sentences than the other algorithms to reach a 2-phoneme distribution as natural as possible.

5.2. Experiment 2

The results are shown in Table 3. The sub-columns labelled "(k=1)+rand" describe the statistics of the 1-coverings augmented by a sentence sampling at the 2-covering size level. The lower bound of the optimal solution size is 13,352 phones for k=1 and 25,414 phones for k=2. In terms of

Covering	$LamSCP^{(1)}_{\{1,2\}}$		$GreedyAS^{(1)}_{\{1,2\}}$		$GreedyAKL^{(1)}_{\{1,2\}}$	
Statistics	c.i.	rstd (%)	c.i.	rstd (%)	c.i.	rstd (%)
Size (phones)	13,458±2	0.07	14,914±22	0.58	34,184±28	0.33
2**-phoneme/#**	2,012±0	0	2,012±0	0	2,012±0	0
2**-phoneme/div**	0.089±0.003	13.87	0.088±0.003	14.95	0.041±0.002	20.77
2**-phoneme/cov**	100±0	0	100±0	0	100±0	0
3**-phoneme/#**	8,497±3	0.14	8,967±8	0.36	12,055±2	0.07
3**-phoneme/div**	0.251±0.002	3.26	0.240±0.002	3.63	0.121±0.001	4.21
3**-phoneme/cov**	91.27±0.00	0.03	91.90± 0.02	0.85	96.15±0.004	0.01
4**-phoneme/#**	18,033±7	0.15	19,326±22	0.44	31,817±18	0.23
4**-phoneme/cov**	78.17±0.01	0.06	79.03±0.02	0.13	87.04±0.009	0.04
5**-phoneme/#**	28,483±10	0.14	30,670±39	0.49	56,737±43	0.29
5**-phoneme/cov**	66.66±0.01	0.09	67.53±0.02	0.15	76.88±0.01	0.05
The best lower bound of the minimal solution size, found by *LamSCP*, is 13,376 phones The c.i. for the lower bound mean value is 13,356±1 phones						

Table 2: Results of experiment 1. Relative standart deviation (rstd) and 95% confidence interval (c.i.) for each criterion mean value based on 59 1-coverings of $\{1, 2\}$-phonememes derived by *LamSCP*, *GreedyAS* and *GreedyAKL*.

Covering	$LamSCP^{(k)}_{\{1,2\}}$			$GreedyAS^{(k)}_{\{1,2\}}$			$GreedyAKL^{(k)}_{\{1,2\}}$		
Corpus	k=1	k=2	(k=1)+rand	k=1	k=2	(k=1)+rand	k=1	k=2	(k=1)+rand
Size (phones)	13,454	25,585	25,603± 2	14,997	27,615	27,631± 1	34,063	65,377	65,393± 2
2**-phoneme/#**	2,012	2,012	2,012±0	2,012	2,012	2,012±0	2,012	2,012	2,012±0
2**-phoneme/div**	0.071	0.052	0.056±0.002	0.076	0.051	0.054±0.002	0.032	0.012	0.042±0.004
2**-phoneme/cov**	100	100	100±0	100	100	100±0	100	100	100±0
3**-phoneme/#**	8,497	11,366	10,844±8	8,975	11,812	11,260±8	12,046	15,253	14,752±9
3**-phoneme/div**	0.232	0.191	0.154±0.001	0.230	0.190	0.149±0.001	0.116	0.098	0.094±0.002
3**-phoneme/cov**	91.34	94.59	94.82±0.01	91.93	94.94	95.13±0.00	96.13	97.81	97.77±0.00
4**-phoneme/#**	18,066	27,360	26,539±20	19,379	28,774	27,888±20	31,738	46,132	44,899±28
4**-phoneme/cov**	78.30	83.60	84.33±0.01	79.10	84.19	84.88±0.01	87.00	90.89	90.91±0.00
5**-phoneme/#**	28,566	46,061	45,298±33	30,770	48,671	47,789±32	56,557	88,913	86,957±48
5**-phoneme/cov**	66.82	72.53	73.60±0.01	67.62	73.19	74.22±0.01	76.84	81.96	82.04±0.00

Table 3: Results of experiment 2. Statistics for the k-coverings of $\{1, 2\}$-phonemes and the associated random-enlarged coverings.

sizes, we can notice therefore the effective achievements of *LamSCP* and *GreedyAS* to calculate solutions close to the optimal ones in terms of size. The 1- and 2-coverings obtained by *GreedyAS* are about 11% longer than the ones given by *LamSCP* and increase the number of distinct 3- to 5-phonemes by 7 to 8% for k=1, by 4 to 6% for k=2. As for the other criteria, the results of *LamSCP* and *GreedyAS* are close. Concerning the coverings produced by *GreedyAKL*, their 2**-** and 3**-phoneme/div** values are about a factor of 2 to 4 smaller than the other ones but their sizes are at least twice greater. Hence, it seems to be difficult to compare the k-coverings obtained by *GreedyAKL* with the previous ones.

Let us compare, for a given algorithm, the statistics of the 1-covering with those of 2-covering. The relative ratios of the sizes of the corpora take values between 1.84 and 1.91. For each algorithm, the 2-covering really improves the analysing criteria in a similar way. However, if *GreedyAKL* is the most well-adapted method to decrease the criterion 2**-phoneme/div**, on the other criteria, *LamSCP* gives the best improvement. *GreedyAKL* improves the criteria 3**-** to 5**-phoneme/#** and 3**-** to 5**-phoneme/cov** less than both other algorithms. It is due to its main objective to provide a distribution of 2-phonemes as "natural" as possible, whereas *LamSCP* and *GreedyAS* favour sentences with rare units.

At last, we compare the 2-covering problem with the randomly enlarged 1-coverings. For each algorithm, the statistics of the 2-covering and the enlarged 1-covering have similar values except for the criterion 3**-phoneme/div**, and for the 2**-phoneme/div** value in the last column. Indeed, the augmented 1-covering offers a more natural distribution of $\{1, 2, 3\}$-phonemes than the 2-covering. As for $GreedyAKL^{(2)}_{\{1,2\}}$, its distribution of $\{1, 2\}$-phonemes is, as expected, more natural than the one relative to the augmented 1-covering.

Finally, considering all criteria, it seems to be interesting to randomly augment a 1-covering of $\{1, 2\}$-phonemes obtained by *LamSCP* or *GreedyAS* as to solve the problem of the 2-covering of $\{1, 2\}$-phonemes. Furthermore, this first strategy permits easier control on the reduced corpus size.

Corpus	$LamSCP^{(1)}_{\{1,2\}}$+rand(N)			$GreedyAS^{(1)}_{\{1,2\}}$+rand(N)			rand(N)
N	20,000	25,000	30,000	20,000	25,000	30,000	30,000
Size (phones)	20,015±2	25,017±2	30,014±2	20,016±2	25,018±2	30,014±2	30,065±2
2-**phoneme/#**	2,012±0	2,012±0	2,012±0	2,012±0	2,012±0	2,012±0	1,403±2
2-**phoneme/div**	0.063±0.001	0.060±0.002	0.052±0.002	0.069±0.002	0.059±0.001	0.057±0.002	0.101±0.008
2-**phoneme/cov**	100±0	100±0	100±0	100±0	100±0	100±0	99.69±0.004
3-**phoneme/#**	9,908±7	10,759±8	11,473±9	10,020±6	10,873±8	11,580±9	10,121±10
3-**phoneme/div**	0.181±0.001	0.157±0.001	0.139±0.001	0.189±0.002	0.160±0.001	0.144±0.001	0.102±0.001
3-**phoneme/cov**	93.65±0.01	94.73±0.05	95.47±0.01	93.63±0.01	94.71±0.01	95.47±0.01	95.69±0.007
4-**phoneme/#**	22,971±16	26,198±20	29,097±23	23,078±14	26,333±19	29,240±23	28,250±29
4-**phoneme/cov**	82.19±0.01	84.15±0.01	85.62±0.01	82.036±0.01	84.04±0.01	85.55±0.01	86.64±0.01
5-**phoneme/#**	38,054±24	44,588±31	50,656±37	37,972±21	44,553±28	50,654±37	51,603±47
5-**phoneme/cov**	71.12±0.01	73.39±0.01	75.14±0.01	70.89±0.01	73.22±0.01	75.02±0.08	76.61±0.01

Table 4: Results of experiment 3. Statistics for random-enlarged coverings and random-built corpus.

5.3. Experiment 3

Table 4 shows results of experiment 3.

In Table 4, if we compare the statistics of the random corpus with 1-coverings completed up to an equal size level, we can notice that these mixed corpora improve to at least 43% the number of distinct $\{1,2\}$-phonemes, resp. 13% for $\{1,2,3\}$-phonemes, resp. 3% for $\{1,2,3,4\}$-phonemes and these additional units are rare in regards to the n-**phoneme/cov** values.

Given Tables 3 and 4, the values of the divergence criteria generally decrease when the length of the corpora increases, except for the above-mentioned case of the enlarged 1-covering using *GreedyAKL*, and for the 2-**phoneme/div** value in the randomly built corpus case. Indeed, a corpus design in a random way does not guarantee a good covering of 2-phonemes (more than 30% of uncovered $\{1,2\}$-phonemes in rand(30,000)), even for a size of a randomly designed corpus twice as great as the minimal covering size. As for the 3-**phoneme/div** criterion, we can notice that its lowest value in Table 4 is reached by rand(30,000). A first explanation may be that, since no corpus in these experiments is built with regards to the 3-phoneme covering and distribution, the larger is the randomly built part, the more natural is the 3-phoneme distribution and then, the lower is the associated 3-**phoneme/div** value (the influence of the 3-phoneme distribution on 3-**phoneme/div** is certainly greater than the $\{1,2\}$-distribution). However, since a complete covering of $\{1,2\}$-phonemes increases the number of rare 3-phonemes to the detriment to frequent 3-phonemes, this phenomenon weights the tail of the 3-phoneme distribution.

If the choice of the covering strategy remains open for building a small corpus, it seems better to randomly complete the $GreedyAS^{(1)}_{\{1,2\}}$. This solution offers the best counts of $\{3,4,5\}$-phonemes for equivalent covering scores.

6. Conclusion

The work presented in this paper focuses on analyzing experimental behavior of different reduction algorithms for building linguistic corpora. A first set of experiments shows that a set-covering algorithm, *LamSCP* or *GreedyAS*, can be used in order to guarantee a good representation of the shortest units. Moreover, these algorithms have side effect considering criteria which are not taken into account by covering constraints. As for the longest units representation, it turns out that the main parameter is the corpus size. Conversely, complementing a corpus reduced by a set-covering technique is a tempting solution to achieve a desired corpus length and improve representativeness indicators.

7. References

Chevelu, J., N. Barbot, O. Boeffard, and A. Delhay, 2008. Comparing set-covering strategies for optimal corpus design. In *Proceedings of the 6th International Language Resources and Evaluation (LREC)*.

François, H. and O. Boeffard, 2002. The greedy algorithm and its application to the construction of a continuous speech database. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC)*, volume 5.

Garey, M.R. and D.S. Johnson, 1979. *Computers and intractability: a guide to the theory of NP-completeness*. Freeman.

Gauvain, J.-L., L.F. Lamel, and M. Eskenazi, 1990. Design considerations and text selection for bref, a large french readspeech corpus. In *Proceedings of the 1st International Conference of Spoken Language Processing (ICSLP)*.

Hart, M., 2003. Project gutenberg. *http://www.gutenberg.org/*.

Katz, S.M., 1987. Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE transactions on Acoustics, Speech and Signal Processing*, 35:400–401.

Krul, A., G. Damnati, F. Yvon, Boidin C., and T. Moudenc, 2007. Adaptive database reduction for domain specific speech synthesis. In *Proceedings of the 6th ISCA Tutorial and Research Workshop on Speech Synthesis (SSW6)*.

Lambert, T., N. Braunschweiler, and S. Buchholz, 2007. How (not) to select your voice corpus: Random selection

vs. phonologically balanced. In *Proceedings of the 6th ISCA Tutorial and Research Workshop on Speech Synthesis (SSW6)*.

Van Santen, J.P.H. and A.L. Buchsbaum, 1997. Methods for optimal text selection. In *Proceedings of the 5th European Conference on Speech Communication and Technology (Eurospeech)*.