

TTS voice corpus reduction for audio-book generation

Meysam Shamsi

▶ **To cite this version:**

Meysam Shamsi. TTS voice corpus reduction for audio-book generation. 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 3 : Rencontre des Étudiants Chercheurs en Informatique pour le TAL, 2020, Nancy, France. pp.193-204. hal-02786200v3

HAL Id: hal-02786200
https://hal.science/hal-02786200v3

Submitted on 23 Jun 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

TTS voice corpus reduction for audio-book generation

Meysam Shamsi[1]
(1) IRISA, CNRS, 22300 Lannion, France
meysam.shamsi@irisa.fr

ABSTRACT

Nowadays, with emerging new voice corpora, voice corpus reduction in expressive TTS becomes more important. In this study a spitting greedy approach is investigated to remove utterances. In the first step by comparing five objective measures, the TTS global cost has been found as the best available metric for approximation of perceptual quality. The greedy algorithm employs this measure to evaluate the candidates in each step and the synthetic quality resulted by its solution. It turned out that reducing voice corpus size until a certain length (1 hour in our experiment) could not degrade the synthetic quality. By modifying the original greedy algorithm, its computation time is reduced to a reasonable duration. Two perceptual tests have been run to compare this greedy method and the random strategy for voice corpus reduction. They revealed that there is no superiority of using the proposed greedy approach for corpus reduction.

RÉSUMÉ

Réduction du corpus vocal pour la génération de livres audio par TTS

Aujourd'hui, avec l'émergence de nouveaux corpus vocaux, la réduction de voix pour la synthèse de parole TTS expressive devient plus importante. Dans cette étude, une approche de type glouton cracheur pour supprimer des phrases est étudiée. Dans la première étape, en comparant cinq mesures objectives, le coût global du TTS s'est révélé être la meilleure mesure disponible pour l'approximation de la qualité perceptuelle. L'algorithme glouton utilise cette mesure pour évaluer les candidats à chaque étape et la qualité synthétique résultant de la solution en construction. Il s'est avéré que réduire la taille du corpus vocal jusqu'à une certaine durée (1 heure dans notre expérience) ne dégradait pas la qualité synthétique. En modifiant l'algorithme glouton d'origine, il produit une solution en temps raisonnable. Deux tests de perception ont été effectués pour comparer cette méthode gloutonne et la stratégie aléatoire de réduction du corpus vocal. Ils ont révélé qu'il n'y a pas de supériorité dans l'utilisation du glouton proposé pour la réduction du corpus.

KEYWORDS: Text-to-speech, voice corpus, greedy algorithm, perceptual test.

MOTS-CLÉS : Synthèse vocale, corpus vocal, algorithme glouton, test de perception.

1 Introduction

In the Text-to-Speech (TTS) framework, the main reasons to design a small speech database via a corpus reduction approach are generally the database scalability, the recording or the labeling cost reduction, and the elimination of destructive data (Baljekar & Black, 2016).

Several studies could be found in the literature dealing with voice corpus optimization. The main

Volume 3 : Rencontre des Étudiants Chercheurs en Informatique pour le TAL, pages 193–204.
HAL : hal-02786200.
Cette œuvre est mise à disposition sous licence Attribution 4.0 International.

purposes of these studies were the pruning of voice corpus in order to respect a parsimony constraint or to extract a more neutral voice, assuming the elimination of expressive voice parts would improve the naturalness of synthetic signals. The idea used by (Krul *et al.*, 2007) was to remove the least selected acoustic units to synthesize signals, using a unit-selection TTS system, for a domain specific application in order to decrease the voice corpus size. This approach was compared with a greedy strategy based on Kullback-Leibler divergence (KLD). The evaluations indicated better achievements for the adaptive domain pruning method than for the KLD based one.

Some studies have been done in order to use voice data stemming from audio-books to provide input for TTS systems. The main goal in this case is to improve naturalness of a neutral voice. An outlier-removal approach has been introduced by (Braunschweiler & Buchholz, 2011; Cooper *et al.*, 2016). The outlier, supposed to involve less natural-sounding speech, has been found out as hypo-articulated utterances and low mean f0 utterances in (Cooper *et al.*, 2016). In (Braunschweiler & Buchholz, 2011), it has been noticed that variety in a speech corpus degrades TTS quality in general task. They discarded sentences based on acoustic features (extreme f0 patterns, too loud or barely audible sentences) and linguistic features (non-neutral style sentences such as quotation, interjections, utterances which start with lowercase, etc). Although TTS systems for a general task need neutral voice corpus (Braunschweiler & Buchholz, 2011; Cooper *et al.*, 2016), the audio-book generation needs an expressive speech synthesis. Found data like available audio-books in the public domain can contain some destructive parts if directly used to build a voice for TTS systems. The strategy consisting on selecting a cleaner subset of speech data may result higher synthetic quality and helps TTS unit selection engine to speed up. In (Baljekar & Black, 2016), two types of errors, misalignment and annotation errors, which degrade TTS quality have been identified to be removed.

However the objectives of previous works were mainly the improvement of the naturalness of synthetic signals, expressiveness plays also a crucial role in audio-book generation. The main aim of this study is audio-book generation using a recorded voice of book reading. Although synthetic speech has less quality than natural one, the audio-book generation could be less costly using a TTS system. Thus, the voice corpus design problem considered here is defined as the script selection for recording in order to use the resulting signals as voice to synthesize the rest of the book. This problem has been investigated in previous studies (Shamsi *et al.*, 2019, 2020) by taking into account the linguistic information. This paper introduces a posterior strategy as a voice corpus reduction : the study is conducted directly from a fully recorded audiobook, instead of its textual content. The voice is a subset of this audio-book and its achievements is assessed by the quality of the synthetic vocalisation of a complementary part of the book. This approach permits to save recording phases and to test several script selections. Moreover, voice corpus reduction methods can profit from acoustic and linguistic information. Analyses of selected voice sub-corpus achievement and content could be helpful for script design based on only linguistic information before recording process in future works.

The original recorded voice is composed of high quality expressive signals which is spoken by a professional speaker. As to offer an adequate expressiveness (from the perspective of later recording phase), the corpus reduction is done at utterance level. Since the final product will be an audio-book mixing natural speech utterances (which compose the voice) and synthesized speech signals, and a bigger voice size generally provides a better synthesized speech quality, the goal of this study is to find the best trade-off between the signal quality of the audio-book and the voice size.

This paper is organized as follows. First, section 2 describes the proposed greedy algorithm and its heuristics to achieve voice corpus reduction. This algorithm needs to evaluate synthetic signals and utterance candidates to remove without requiring human listeners and section 3 investigates an

objective measure for ranking candidates. At last, the evaluation results of the proposed algorithm, in comparison with a random method as baseline, are detailed in section 4.

2 Framework

In order to extract a reduced voice from the original voice corpus well-adapted to synthetically vocalize the content of the unselected sentences, two main requirements are needed : a practical heuristic for selecting a subset of the full corpus and an automatic evaluation method to assess the quality of synthetic signals by using the voice subcorpus. By considering the previous works (François & Boeffard, 2002; Espinosa *et al.*, 2010; Barbot *et al.*, 2015), it has been shown that the greedy algorithm selects portions of a corpus in reasonable time, close to the optimal ones.

The greedy algorithm (with spitting or agglomerative policy) is an iterative strategy and needs a score function to rank candidates. In each step of spitting (resp. agglomerative) greedy process, candidates with minimum (resp. maximum) *utility* are selected to be excluded from (resp. added to) the voice corpus under reduction (resp. construction). The *utility* score of each utterance represents the increasing gain of the richness in the voice corpus when this utterance is kept. This metric will be presented in section 3.

The objective is to extract *voice corpus* from a fully recorded audio-book to be used by TTS. This voice corpus selection should provide the highest synthetic quality for the rest of the book which is called *synthetic part*. The process, described in algorithm 1, is based on a spitting greedy approach and the initial voice corpus to reduce is the whole audio-book. A similar process has been implemented in (Espinosa *et al.*, 2010) : the authors proposed to agglomeratively select utterances which causes the least synthetic quality degradation for a target set of utterances. In the case of our problem, the target set is the rest of the book and is modified by the reduction process.

Algorithm 1: Spitting greedy for audio-book generation

```
1  voice corpus = candidate set = all utterances ;
2  synthetic part = ∅ ;
3  while the candidate set has at least one utterance do
4      for All U_i utterance in the candidate set do
5          Remove U_i from the voice corpus ;
6          Synthesis the synthetic part by using the voice corpus ;
7          if synthesizing of the synthetic part failed then
8              Lock U_i and remove from the candidate set ;
9          else
10             Compute the utility of the U_i based on the quality degradation of synthetic part ;
11         end
12         Add U_i to the voice corpus;
13     end
14     Find U_x with the minimum utility from the candidate set;
15     Remove U_x from the voice corpus and the candidate set and add to synthetic part;
16 end
```

Some utterances in the audio-book contain unique units and a concatenative TTS system cannot find

these units in other utterances. These utterances are locked and should not be removed from voice corpus. At each step, the remaining unlocked utterances in voice corpus compose the *candidate set* for the next step. By removing utterances from voice corpus, a voice subset with a reduced size will be achieved. The spitting greedy process is continued until the *candidate set* is empty.

For each reduction rate, the rest of the book should be synthesized and evaluated in terms of quality degradation. Indeed a small change inside voice corpus could change the synthetic quality of the *synthetic part*. But since synthesizing the whole *synthetic part* each time is computationally expensive, the synthetic signal of selected candidates in each step will be used as approximation of the overall quality degradation. The idea behind this proposition is that the small change of voice corpus by removing an utterance could be ignored and only the quality degradation of final audio-book because of replacing recorded voice of the utterance by its synthetic signal would be taken into account.

In the ideal scenario, in order to investigate the impact of TTS voice corpus reduction on synthetic quality, all combinations of sub-sets should be evaluated perceptually for a given reduction rate. But this is not possible within a reasonable time. In (Espinosa *et al.*, 2010), the measure for ranking utterances and the evaluation of quality are the same (TTS costs). In this way, the quality evaluation process does not need additional computation (algorithm 1 by using the same score in line 10 and 14 follows this idea). This measure will be investigated in section 3. The selection algorithm and the computational problem of the proposed algorithm will be presented in section 4.

The optimization process starts with a full audio-book as a initial corpus. In this study, the initial voice corpus contains 3339 utterances of a French expressive audio-book spoken by a male speaker. The overall length of the speech corpus is 10h44. More information on the annotation process can be found in (Boeffard *et al.*, 2012). The IRISA TTS system (Alain *et al.*, 2017), which is unit selection based, uses voice subcorpus for synthesising.

3 Objective measure for selection

It is impossible to have all synthetic signals evaluated perceptually by listeners. Thus an automatic measure is necessary for quality evaluation of synthetic utterances. This objective measure should be a good approximation of perceptual evaluation. The correlation coefficient or the ranking correlation coefficient could indicate the reliability of an objective measure.

The unit selection TTS costs is used in previous works (Chu & Peng, 2001; Toda *et al.*, 2006; Krul *et al.*, 2007; Espinosa *et al.*, 2010) as the synthetic quality indicator. In this study, the usage of TTS global cost, which is a linear combination of concatenation and target cost, in unit selection TTS is evaluated as the objective measure of synthetic quality. Moreover this objective measure does not need any supplementary computation in the proposed greedy (the result of the line 6 in algorithm 1 can be used directly for line 10).

Beside TTS costs, we propose other objective measures for quality evaluation of synthetic signals. Some measures such as PESQ (Rix *et al.*, 2001) and Dynamic Time Warping (DTW) between two signals need the reference signal. Basically they evaluate similarity between a test signal and its reference. Three DTW based measures are proposed; a DTW between Mel-Frequency Cepstral Coefficients (MFCC) features of the test signal and its natural pair, a DTW between Mel-Generalized Cepstral Coefficients (MGC) features of the test signal and its natural pair, and a DTW between MGC features of test signal which is synthesized signal using voice subcorpus and a reference signal which

is synthesized signal using the whole of voice corpus. The third DTW calculates the degradation quality of an synthetic test signal from the highest possible synthetic quality using the TTS.

3.1 Experimental setup

To investigate the correlation of these objective quality measures with perceptual quality, a listening test (DMOS) is designed. Six different sub-voice corpora of different sizes (75%, 50%, 25%, 10%, 5%, and 1% out of the initial audio-book) are selected randomly. In this experiment, the rest of the book will be synthesized and used for perceptual evaluation. The listeners are asked to evaluate 60 synthetic samples from each synthetic part corresponding to corpus size. By providing the natural voice of each synthetic signal, the quality degradation of synthetic signal are asked on a scale from 1 to 5 (5 means without quality degradation and 1 means the highest degradation).

3.2 Result

The perceptual test resulted in 850 evaluation scores by 17 listeners. A perceptual score is assigned to each sample by getting the average of its scores. Two ranking correlation coefficients (Spearman (Spearman, 1904) and Kendall tau (Kendall, 1948)) and Pearson correlation coefficient (Freedman *et al.*, 2007) are computed between average perceptual score and objective scores of samples. The correlation coefficients between listener scores and 5 objective measures for all voice corpus sizes are compared in table 1.

Objective measures	PESQ	DTW-MGC (Natural ref)	DTW-MFCC (Natural ref)	DTW-MGC (Synthetic ref)	TTS global cost
Pearson C.C.	0.07(p>0.2)	-0.41(p<0.001)	-0.38(p<0.001)	-0.40(p<0.001)	**-0.66(p<0.001)**
Spearman R.C.C.	0.08(p>0.1)	-0.39(p<0.001)	-0.39(p<0.001)	-0.40(p<0.001)	**-0.65(p<0.001)**
Kendall tau R.C.C.	0.05(p>0.1)	-0.28(p<0.001)	-0.28(p<0.001)	-0.28(p<0.001)	**-0.48(p<0.001)**

TABLE 1: Correlation coefficients between objective measures and perceptual evaluation and their p-value.

According to table 1, the TTS global cost has a stronger correlation with perceptual scores than PESQ or DTW on different acoustic features (MFCC, MGC).

While the reported correlation coefficients are calculated on synthetic signals with 6 voice corpus sizes, the mean of perceptual and objective scores on each voice corpus size could reveal more information. The impact of voice corpus length on synthetic quality (with perceptual and objective measures) is investigated. Figure 1 compares the perceptual and objective scores for synthetic utterances with different sub-corpus sizes. The horizontal axis indicates the size of voice subcorpus out of initial voice corpus which is selected randomly.

The increasing trend of MOS score and decreasing trend of TTS global cost for larger voice subcorpus confirm that the quality of synthetic signals will be improved with larger voice corpus. But the perceptual quality of synthetic signals with 25%, 50%, and 75% of the initial corpus (more than 1 hour) are not significantly different. It means that using more data for TTS voice corpus after a threshold would not improve significantly the speech quality.

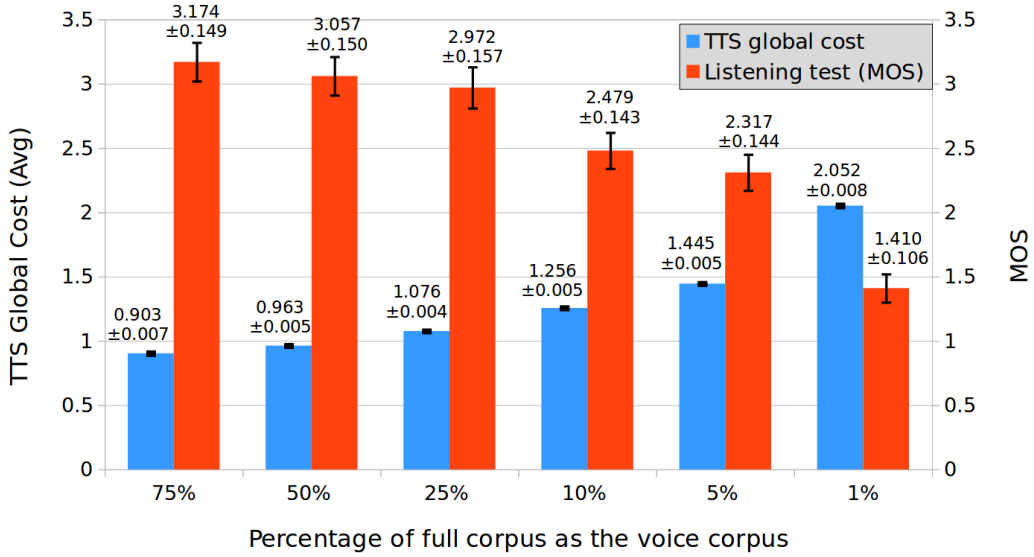


FIGURE 1: The TTS global cost and perceptual score for different voice corpus sizes.

In the remainder of this study, the TTS global cost will be used as an approximation of perceptual quality with the spitting greedy approach, which is evaluated in the next section.

4 Optimization strategy

In this section, a modified greedy process based on algorithm 1 will be compared with a random selection for voice corpus reduction in terms of synthetic quality.

4.1 Modified Greedy algorithm

The computational time of the algorithm 1 makes its use unfeasible for large voice corpora. Calling TTS for synthesizing *the synthetic part* and evaluating its quality are expensive. Consequently, we propose to do two modifications on the original algorithm in order to find a solution in a reasonable time. The first modification is removing bunch of utterances in each step instead of one utterance. And the second one is evaluating the utility of *the candidate set* (algorithm 1, line 10) based on the quality of the synthetic signal of a single utterance instead of the quality degradation of *the synthetic part*. In the following these proposition will be justified.

By analyzing the ranking list of utterances in consecutive steps of the algorithm, it has been observed that the ranking list of candidate utterances does not change a lot from one iteration to the next. Although a simple experiment has showed that this assumption is not completely true, considering the utility of utterances (for being in voice corpus) as a stable rank list helps to get rid of all computational problems. The normalized root-mean-square error (NRMSE) of the TTS global cost for corpus reduction from initial corpus to 70% of initial corpus was 0.012. It shows that, in a big corpus, following the initial ranking list makes the resulting sub-corpus slightly worse than following the original spitting greedy. It leads to a compromise between updating ranking list after each change, which is computationally expensive, and using initial ranking list, which gives a less efficient solution.

We propose to remove a bunch of utterances (100 utterances) based on the ranking list at each step of the spitting greedy. The ranking list is then updated after removing the bunch of utterances. This modification reduces the computational time.

In the second proposition, the computational time of the candidates list evaluation will be reduced. Indeed a change inside the corpus could have an effect on the whole of the *synthetic part*. It means that the *synthetic part*, which U_i has been added to, should be evaluated for U_i's utility in the candidate list. When the evaluation of synthetic part for each U_i candidate is computationally expensive, we propose to consider the TTS global cost of synthesized U_i as an inverse metric for utility. It helps to save synthesizing time of *synthetic part* for evaluating candidates. This idea changes the algorithm 1 by modifying the *synthetic part* to U_i (line 10).

By applying these two modifications on the original greedy algorithm, the computational time will be reduced drastically. The number of utterances in each removal step has a linear relation with computational time. Based on our physical facilities, by removing a bunch of 100 utterances, the experiment will be finished in 2 days, which seems reasonable, and it gives a reduction steps of 3% (or 20 minutes) for voice corpus reduction. Although choosing smaller number of utterances in removal bunch improve the result, it costs computational time.

Comparing the original spitting greedy and modified spitting greedy on a small corpus (334 utterances) shows that the sub-corpus resulting from the proposed greedy synthesizes with higher TTS global cost in a shorter time. It is expected since the modified algorithm is not as efficient as the original one even if it helps to find a solution in a reasonable time.

4.2 Experimental setup

In order to compare perceptually the corpus reduction methods a *test section* (10% of the initial voice corpus) is extracted from the book. Although the main problem in our case is to synthesize the rest of the book, a synthetic part as *test section* would help to compare different methodologies for corpus design. It is assumed that the voice corpus, which is supposed to synthesise the rest of the book, has almost same performance on the *test section*. We assume that since the test part comes from the same book, the synthetic quality of this part can be generalized to the rest of the script. The *test section* is randomly selected as a continuous part with 334 utterances. The remaining part of the audio-book is named the *full corpus*. The *test section* has been synthesized by TTS using 100%, 70%, 40%, 15%, 7%, 3% of the *full corpus*. The voice corpus reduction is done based on the spitting greedy and a random strategy. The initial corpus is the same audio-book as what has been described in previous section (see the end of section 2).

In the following, the evaluation of proposed corpus reduction methods will be detailed. Two perceptual tests are designed to evaluate the quality of signals which are synthesized using resulting voice subcorpora. Based on the previous perceptual test results, the objective measure for ranking utterances in the spitting greedy is the TTS global cost. The first perceptual test is designed to investigate the impact of voice corpus reduction by modified greedy on synthesizing quality. The purpose of the second perceptual test is to compare the performance of proposed greedy and random strategy.

4.3 Synthetic quality degradation by greedy voice reduction

The greedy and random methods provide voice subcorpora with different length to synthesize *test section*. Since the IRISA TTS is unit selection-based, some of the utterances would failed to be synthesized specially in small voice corpus size. After removing these uncommon samples, 70 utterances has been selected randomly. In order to have samples with an acceptable duration, some utterances have been concatenated or cut. More precisely, if the length of selected synthetic signal is less than 4 seconds, the next utterance in script order is concatenated to it. The first 6 seconds of synthetic signals have been cut and used as listening samples. Samples from 6 voice corpus sizes and two corpus reduction methods are used to design a MUSHRA test (Recommendation, 2003). For each step of the perceptual test, the overall quality of 11 synthetic signals, which have been synthesized with different sub-voice corpora, have been evaluated on a scale form 1 to 10 (with one by one increment). Synthetic signals and corresponding natural voice, which have the same script, are available to listeners. The listeners are asked to do 10 steps of MUSHRA test after an introduction step. The estimated time for doing this test is 25 minutes.

This perceptual test has been done by 14 listeners which provides 1441 evaluation scores for synthetic signals. To investigate the impact of corpus size on synthetic signals, the average score for each size/method has been calculated. The figure 2 (left) shows that the average scores for all voice corpus are in a almost same level. It indicates, not only the quality of synthetic signals based on random and greedy strategy dose not have significant different, but also reducing the voice corpus size can not impact on resulted quality at least until 15% of corpus reduction rate. The listeners evaluated 20% of the signals with exact same values.

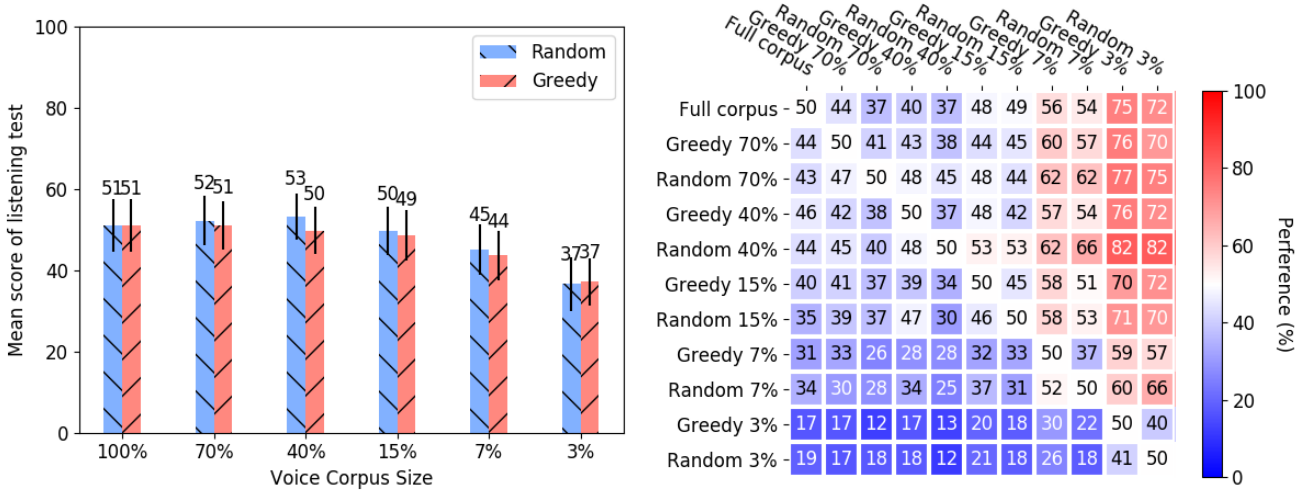


FIGURE 2: The average MUSHRA scores for different voice corpus and size (left). The preference of listeners to assign higher score for signals in compare to others (right).

According to listeners feedback, comparing 11 samples is not an easy task and they had been exhausted. This problem encourages us to estimate the preference of listeners as if they were asked to compare two signals. So the resulting scores from MUSHRA test are used to simulate an AB test. Concretely, each two signals are compared based on their perceptual score. The score values of two signals are converted to a simple comparison in order to simulate the preference of listeners and if the scores are equal, it would be assumed that the listeners does not have any preference. The result is shown on figure 2 (left). The numbers in the heatmap table indicate the preference percentage of

vertical labels against horizontal labels. Based on this figure, the preference of synthetic signals with small voice corpus (left-down) is lower than synthetic signals with large voice corpus (right-up). It confirms that voice corpus reduction decreases the TTS synthetic quality. By looking at cells in large corpus size (left-up), it can be observed that the preference numbers for corpus sizes larger than 15% are around 50. This observation confirms the hypothesis in section 3. It means after a certain voice corpus size the quality of synthetic signal would not be improved perceptually by increasing the voice corpus size.

Figure 2 does not show superiority of spitting greedy in comparison with random strategy. These is contrary to what we expected based on previous studies such as (Chevelu & Lolive, 2015). While it was reported by listeners that the MUSHRA test is not an easy task for this comparison, another perceptual test is proposed for comparing the performance of these two corpus reduction methods.

4.4 The performance of greedy strategy v.s. random selection

Based on listeners' feedback from previous perceptual test, some modifications have been done on listening samples preparation and the platform. While we use the same *test section* and reduction rates, the final listening signals are prepared in a different way. The utterances have been synthesized from the beginning until the first speech pause after 90 diphones. In this way, all samples for sizes/methods will have same content. The duration time of samples are between 5 to 10 seconds. Among 334 utterances of the *test section*, 70 samples have been selected for the listening test according to the highest acoustic distance (Chevelu *et al.*, 2015). The acoustic distance is computed by calculating DTW on the MGC features of two signals. This selection method helps to focus on the most different samples.

An AB test has been prepared with 40 comparison steps. For each step, listeners are asked to give their preference in terms of overall quality between two synthetic signals. These signals have been synthesized by using different voice subcorpora but with same size. Voice subcorpora are a sub part of the *full corpus* selected by the random strategy or the proposed spitting greedy. The reference signal is not provided which lets listeners decide what is the best quality. We hope it makes the task easier. The estimated time for doing the whole test is 15 minutes.

The listening test has been done by 9 listeners. For each voice corpus size between 66 and 70 comparisons have been achieved. Out of 340 comparisons in total, 132 times random strategy has been preferred, 118 times greedy strategy, and 90 times listeners selected no preference. The figure 3 (left) shows the percentage of preference for corpus reduction methods in different voice corpus size.

The figure 3 (left) does not reveal any significant superiority of the modified greedy. Even the synthetic signals for 15% of full voice corpus with random strategy has been evaluated slightly better than modified greedy.

The TTS global costs of the AB test's samples is displayed in the figure 3 (right). This figure shows the synthetic quality of the listening test signals in terms of TTS costs. The TTS global cost given by random selection are not significantly different from those given by the proposed greedy. While the initial problem was synthesizing the rest of the book instead of test section, the the rest of the book has been synthesized by extracted sub-corpus as TTS voice. A same trend as figure 3 is observed for rest of the book (synthetic part). It means that the listening test signals have same synthetic quality as the rest of the book.

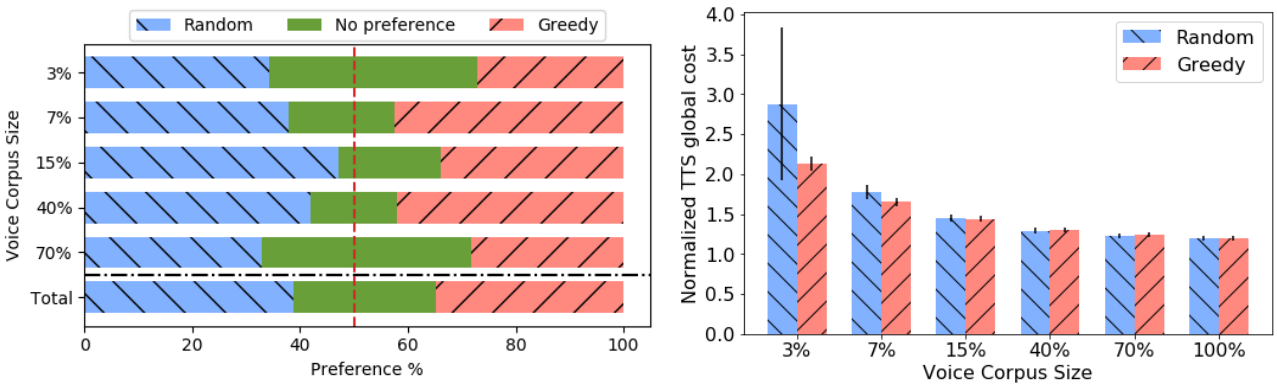


FIGURE 3: The AB test result for different voice corpus size and the total preference of listeners for random and proposed greedy (left). The normalized TTS global cost of listening test samples (right).

It could be concluded that the random reduction works as well as the proposed spitting greedy. The explanation could be the approximation level of the proposed method. It means that reducing the computational time costs a lot in terms of the efficiency of subset solution. Hence the performance of voice subcorpus resulted by proposed greedy becomes close to a random selection.

5 Conclusion

The computational time is the main challenge in voice corpus reduction by greedy algorithm. By modifying the original spitting greedy, its complexity has been reduced to a reasonable time. However this approximation level probably costs lower efficiency and makes the solutions closer to the random strategy.

In the first step, some objective measures like PESQ, DTW between synthetic signal and voice signal, and TTS global cost have been investigated. A perceptual listening test (DMOS) has been designed to evaluate the synthetic signals using different voice subcorpus sizes. A higher correlation between objective measures and perceptual quality confirmed that TTS global cost is the best available metric to estimate perceptual quality.

Hence the TTS global cost has been employed in greedy algorithm for ranking candidates in each reduction step. By modifying the original greedy algorithm, the computational time reduced to a reasonable time. Although these modifications cause some level of inefficiency. The proposed greedy has been compared with random strategy for different voice corpus sizes in a MUSHRA test. It has revealed that after a certain size of voice (1 hour of our audio-book), the voice corpus is big enough and the difference of synthetic signals because of voice corpus size can not be distinguished perceptually. Moreover it has not been observed any difference between random and proposed greedy. In order to evaluate the performance of the proposed greedy, an AB preference test has been run. The result of this listening test confirmed that listeners did not prefer the signals which are synthesized using voice subcorpus obtained with the proposed greedy.

Acknowledgments

Thanks to Damien Lolive, Jonathan Chevelu, and Nelly Barbot for their appreciable help as supervisors. This study has been realized under the ANR (French National Research Agency) project SynPaFlex ANR-15-CE23-0015.

References

ALAIN P., BARBOT N., CHEVELU J., LECORVE G., SIMON C. & TAHON M. (2017). The IRISA text-to-speech system for the blizzard challenge 2017. In *Blizzard Challenge workshop. International Speech Communication Association (ISCA)*, Stockholm, Sweden.

BALJEKAR P. & BLACK A. W. (2016). Utterance selection techniques for TTS systems using found speech. In *9th ISCA Workshop on Speech Synthesis (SSW9)*, p. 184–189, Sunnyvale, USA.

BARBOT N., BOEFFARD O., CHEVELU J. & DELHAY A. (2015). Large linguistic corpus reduction with SCP algorithms. *Computational Linguistics*, **41**(3), 355–383.

BOEFFARD O., CHARONNAT L., LE MAGUER S., LOLIVE D. & VIDAL G. (2012). Towards fully automatic annotation of audio books for tts. In *Eighth International Conference on Language Resources and Evaluation (LREC)*, p. 975–980, Istanbul, Turkey.

BRAUNSCHWEILER N. & BUCHHOLZ S. (2011). Automatic sentence selection from speech corpora including diverse speech for improved HMM-TTS synthesis quality. In *Twelfth Annual Conference of the International Speech Communication Association (InterSpeech)*, p. 1821–1824, Florence, Italy.

CHEVELU J. & LOLIVE D. (2015). Do not build your TTS training corpus randomly. In *Proceedings of the 23rd European Signal Processing Conference (EUSIPCO)*, p. 350–354, Nice, France : IEEE.

CHEVELU J., LOLIVE D., MAGUER S. L. & GUENNEC D. (2015). How to compare TTS systems : a new subjective evaluation methodology focused on differences. In *Sixteenth Annual Conference of the International Speech Communication Association (InterSpeech)*, p. 3481–3485, Dresden, Germany.

CHU M. & PENG H. (2001). An objective measure for estimating MOS of synthesized speech. In *Seventh European Conference on Speech Communication and Technology (EuroSpeech)*, p. 2087–2090, Aalborg, Denmark.

COOPER E., CHANG A., LEVITAN Y. & HIRSCHBERG J. (2016). Data selection and adaptation for naturalness in hmm-based speech synthesis. In *Seventeenth Annual Conference of the International Speech Communication Association (InterSpeech)*, p. 357–361, San Francisco, USA.

ESPINOSA D., WHITE M., FOSLER-LUSSIER E. & BREW C. (2010). Machine learning for text selection with expressive unit-selection voices. In *Eleventh Annual Conference of the International Speech Communication Association (InterSpeech)*, p. 1125–1128, Makuhari, Japan.

FRANÇOIS H. & BOEFFARD O. (2002). The greedy algorithm and its application to the construction of a continuous speech database. In *Third International Conference on Language Resources and Evaluation (LREC)*, volume 5, p. 1420–1426, Las Palmas, Spain.

FREEDMAN D., PISANI R. & PURVES R. (2007). Statistics (international student edition). *Pisani, R. Purves, 4th edn. WW Norton & Company, New York*.

KENDALL M. G. (1948). Rank correlation methods.

KRUL A., DAMNATI G., YVON F., BOIDIN C. & MOUDENC T. (2007). Approaches for adaptive database reduction for text-to-speech synthesis. In *Eighth Annual Conference of the International Speech Communication Association (InterSpeech)*, p. 2881–2884, Antwerp, Belgium.

RECOMMENDATION I. (2003). 1534-1 : Method for the subjective assessment of intermediate quality level of coding systems. *International Telecommunication Union*.

RIX A. W., BEERENDS J. G., HOLLIER M. P. & HEKSTRA A. P. (2001). Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 2, p. 749–752, Salt Lake City, USA : IEEE.

SHAMSI M., CHEVELU J., LOLIVE D. & BARBOT N. (2020). Corpus design for expressive speech : impact of the utterance length. In *International Conference of Speech Prosody*.

SHAMSI M., LOLIVE D., BARBOT N. & CHEVELU J. (2019). Corpus design using convolutional auto-encoder embeddings for audio-book synthesis. In *Twentieth Annual Conference of the International Speech Communication Association (InterSpeech)*, p. 1531–1535, Graz, Austria.

SPEARMAN C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology*, **15**(1), 72–101.

TODA T., KAWAI H., TSUZAKI M. & SHIKANO K. (2006). An evaluation of cost functions sensitively capturing local degradation of naturalness for segment selection in concatenative speech synthesis. *Speech Communication*, **48**(1), 45–56.