DO NOT BUILD YOUR TTS TRAINING CORPUS RANDOMLY

Jonathan Chevelu, Damien Lolive

IRISA - University of Rennes 1, Lannion, France

jonathan.chevelu@irisa.fr, damien.lolive@irisa.fr

ABSTRACT

TTS voice building generally relies on a script extracted from a big text corpus while optimizing the coverage of linguistic and phonological events supposedly related to voice acoustic quality. Previous works have shown differences on objective measures between smartly reduced and random corpora, but not when subjective evaluations are performed. For us, those results do not come from corpus reduction utility but from evaluations that smooth differences. In this article, we highlight those differences in a subjective test, by clustering test corpora according to a distance between signals so as to focus on different synthesized stimuli. The results show that covering appropriate features has a real impact on the perceived quality.

Index Terms— Corpus reduction, Subjective evaluation, Corpus-based Unit Selection TTS

1. INTRODUCTION

In the field of Text-To-Speech (TTS) synthesis, size and quality of the corpus are among the main factors that influence the quality of the signal produced. In particular, this quality depends on different aspects such as speaker's voice, recording quality, annotation accuracy and also phonological richness.

In order to limit the recording and post-processing costs as well as to ensure voice homogeneity, the recording script has to be of reasonable size while guaranteeing phonological coverage, generally of diphone classes. Designing a rich corpus with a minimal size has been widely studied and several approaches exist [1–5]: it consists in reducing a large corpus by selecting a set of sentences which covers a given set of features. For instance, a comparison has been made in [6, 7] between two greedy approaches and a Lagrangian based approach. These studies have shown the suitability of these approaches to reduce a corpus according to a set of phonological features. Moreover, [8] proposes an evaluation of the reduction impact on the quality of a TTS system that tends to show that a randomly selected corpus may achieve a similar output quality to a corpus covering the diphones. A possible explanation is that a random selection of sentences provides a set of units distributed following a natural distribution. Then the units used during synthesis also follow the same distribution and are more frequent which results in less artifacts. Consequently, such a random selection may be sufficient to render the frequent units with correct quality. Moreover, when evaluating the difference between different reduction strategies, the samples are usually selected randomly. The consequence is also that the systems are generally compared using samples which, if not the same, are most of the time equivalent and bias the evaluation results by smoothing marks. We rather think that to compare two systems, one may concentrate on the differences between them. Thus, we propose a new evaluation methodology to compare the systems by clustering test sentences according to an alignment cost.

In this paper, we investigate the quality loss between a full corpus and two reduction strategies. The first one produces a coverage of units described by phoneme labels and positional features useful for synthesis. The second one simply uses diphone coverage that is completed up to the size of the first one by a random set of phrases. Going further the work of [8], which uses only lexical stress information, we introduce other features useful for TTS systems to compute the coverage. By the introduction of synthesis specific features, we want to see if the coverage of such features may help the synthesis system to produce better quality speech.

To do so, we consider a corpus-based speech synthesis system [9]. In this study, we use the speech synthesis system to produce speech using the two corpora presented above and assess the impact of the reduction. Using the proposed methodology, we show that a real difference exists between the two covering strategies.

First, a brief presentation of both the corpus-based TTS system and the reduction approach is proposed in section 2. We insist on the features covered and the systems we compare. In section 3, the proposed methodology as well as the corpora used in the experiments are introduced. Then the experiments and their results are presented in section 4.

2. TOOLS AND SYSTEMS

2.1. Speech synthesis system

The corpus-based TTS system described in [9] is used to produce an acoustic signal from a text input. The main goal of the synthesis system is to find, in a corpus, the best unit sequence

978-0-9928626-3-3/15/$31.00 ©2015 IEEE

that best matches a target unit sequence while minimizing the audible concatenation artifacts. Generally, this problem can be formulated in the following way (as presented in [10]):

$$U^* = \underset{U}{argmin} \left(\sum_{n=1}^{card(U)} C_t(u_n) + \sum_{n=2}^{card(U)} C_c(u_{n-1}, u_n) \right) \quad (1)$$

where U^* is the best candidate unit sequence according to the cost function and u_n the candidate unit intending to match the n^{th} target unit. The sub-cost $C_t(u_n)$ (target cost) represents the distance between the candidate and the corresponding target unit. The second sub-cost, $C_c(u_{n-1}, u_n)$ (concatenation cost), is the distance between the current candidate u_n and the previous one u_{n-1} in the path, used to measure artifacts in concatenation areas. To compute U^*, the Viterbi algorithm is used with no restriction on the lattice size. The size of the units considered during the search is variable. Here, we only consider diphone and triphone units, but finding longer units (if they exist) is insured, as the cost function grants a null cost to consecutive units in the corpus.

As for the concatenation cost, we only take into account two components which are distances in terms of amplitude and $F0$ between two consecutive units:

$$C_c(u, v) = C_{amp}(u, v) + C_{F0}(u, v) \quad (2)$$

where u and v are the units under comparison, $C_{amp}(u, v)$ is the amplitude cost and $C_{F0}(u, v)$ is the $F0$ cost. The target cost is not considered for this study. The label of the phonemes is filtered to take only the ones that match the target sequence.

2.2. Covering algorithm

In the context of a speech synthesis system, the reduction process is a trade-off between reducing the size of a large corpus and covering a given set of attributes (mainly linguistic ones) interesting for a good synthesis quality. This problem can be seen as a generalization of a set covering problem (SCP) [3]. It is known as a NP-hard problem and the most frequent strategy is to use greedy algorithms to solve it. Considering the distribution of the desired attributes in the linguistic corpora, many types of greedy algorithms have been studied, for example in [11] and [12]. Through the use of Lagrangian relaxation principles, [6] shows that an *Agglomeration* greedy algorithm followed by a *Spitting* greedy *Algorithm* is close to the optimal solution in this framework. This combination, called *ASA*, has been chosen in this paper.

Two reductions are designed from the *Full* corpus using the *ASA* algorithm:

- *TTSCover* covers similarly at least once each successive pair of phonemes taking into account their associated vectors of features containing the following information:
  - Is the phone in the last syllable of its phrase?
  - Is the phone in the last syllable of its sentence?
  - Is the phone in the onset of the syllable?
  - Is the phone in the coda of the syllable?
- *CompRand* is obtained by randomly complementing the shortest set of phrases (in terms of number of label occurrences) covering at least once each diphoneme of *Full* until reaching the same number of phones as *TTSCover*.

3. PROPOSED METHODOLOGY

In this section, we present the proposed evaluation methodology including the different corpora used in the experiments.

3.1. Approach

Generally, the classic approach for subjective evaluations is to synthesize a small set of samples, to propose them to listeners for evaluation, and draw conclusions about the systems based on this small set of samples. In our opinion, this method works for systems that have a large output quality difference and depends greatly on the set of sentences chosen. For us, to reveal the differences between two systems, we have to focus on the differences found in the generated speech signals. Moreover, as the evaluation generally relies on a small set of samples, it is not possible to select the most different output signals. Consequently, we propose the following:

1. Synthesize a large text of a different style/domain with each system;
2. Compute for each pair of samples the alignment cost (e.g. a DTW [13]);
3. Select the most different samples to evaluate the systems.

In this paper, the alignment cost is computed using the DTW cost between the MFCC sequences for each signal, divided by the alignment path length which gives a normalized cost. This measure has the good property of being independent from the systems under evaluation but another one may be used.

3.2. Speech corpus

Since this work is a preliminary study of the expensive task of building a speech corpus for TTS systems from the design of the recording script to the annotation phase, we have chosen to reduce a large speech corpus, in French and spoken by a female speaker. Its initial purpose was the TTS system of an answering automaton in a Telecommunication framework and its annotations are manually checked.

3.3. Reduced corpora

Three corpus-based TTS systems are then built: they respectively select speech segments in *Full*, *TTSCover* and

CompRand, and they are called by the name of their associated corpus, without risk of confusion. Main statistics of the previous corpora are presented in Table 1.

Sub-corpus	*Full*	*TTSCover*	*CompRand*
Duration	7h06'12	59'04	53'19
Size in phrases	7,662	1,010	1,022
Size in labels	259,684	31,994	31,994
Labels	34 phonemes and 2 NSS		
Diphonemes	1,242		

Table 1: Main statistics of used corpora.

3.4. Evaluation corpus

To be independent from the speech corpus chosen, we have used a different textual corpus. It is composed of a set of sentences extracted from a collection of 50 e-books covering many topics and writing styles. The resulting sentences are then filtered to keep those that have between 30 and 60 phonemes in order to produce outputs roughly between 3 and 6 seconds (as recommended in [14]). Finally, $10,000$ sentences are extracted randomly from the complete set of $68,710$ sentences to build the test corpus that has to be synthesized.

4. EXPERIMENTS

4.1. Classic subjective evaluation

We have first conducted a MUSHRA [15] based evaluation to compare the two systems (*TTSCover* and *CompRand*) by taking randomly selected samples among the $10,000$ generated phrases. We ensure that they have no phonetisation issues. We also add signals from *Full* as baseline. This approach corresponds to the standard evaluation process. The test is composed of 15 steps with one additional introduction step. Each step presents a sample from each system. 10 listeners were asked to evaluate the overall quality of the stimuli and give a mark from 0 to 100 (by steps of 5 points).

The results are presented in Table 2. The only system significantly judged different from the others is *Full* (both on mean score and mean rank). We also note that *TTSCover* and *CompRand* are judged as strictly equivalent for this first experiment. Even if we use a different TTS engine, different features to cover and another language, those results are similar to those from [8].

This result can be explained by considering the distribution of units in a corpus. When you take randomly chosen sentences from a corpus, the resulting distribution of the units selected follows the natural distribution of the units in the corpus. This is true both for the reduced corpus with random selection and for the sentences used for this subjective evaluation. Consequently, such a process tends to favor the random reduction strategy by taking the most frequent units. Thus, even if samples taken from one or the other system may exhibit differences, the mean score achieved is the same. In this case, it does not mean that the systems are equivalent but that we have evaluated the systems on what they do best, i.e. the most frequent events.

The problem is that when a rare unit is selected, the possible artifact generated degrades the global perception of the synthesized sentence. Thus, a TTS system should provide good results even for rare events, which is not evaluated by this classical methodology.

		score		rank	
		mean	conf. int.	mean	conf. int.
System	*Full*	66.5	± 2.7	1.2	± 0.1
	TTSCover	44.5	± 2.7	2.3	± 0.1
	CompRand	44.4	± 2.8	2.3	± 0.1

Table 2: MUSHRA evaluation results for a random selection of samples. 95% confidence intervals are also presented.

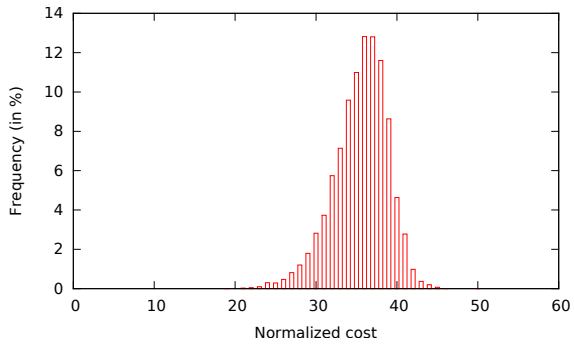
4.2. Alignment costs repartition

Figure 1 shows the distribution of the DTW costs on the $10,000$ sentences when comparing *TTSCover* and *CompRand*. Both with the histogram and the density function, we can observe a gaussian-like behavior. The consequence is that when selecting randomly the samples to build a perceptive evaluation, one selects samples that fit this distribution which means, with a high number of equivalent samples. The results of the perceptive evaluation is then quite understandable when it finds nothing significantly different between the systems.

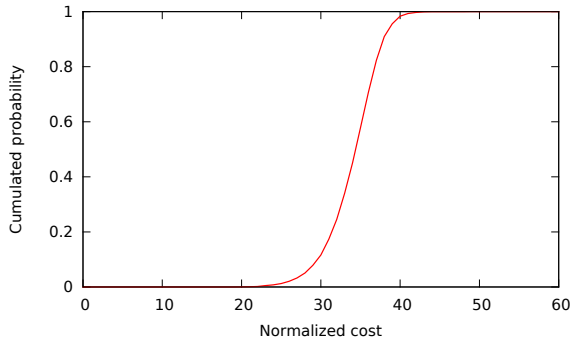
In order to compare the systems, we suggest concentrating the evaluation where the systems are measured as different, under the assumption of a given measure (i.e. DTW here). Then, to build the subjective test, we may choose sample pairs that give a high difference measure (on the right side of the distribution given in figure 1). Note that the measure used only reflects the differences between signals and does not provide any information about the signal quality. Likewise, it does not favor any system.

4.3. Proposed subjective evaluation

A second MUSHRA evaluation has been conducted to compare the two systems (*TTSCover* and *CompRand*) by taking the samples with the highest costs among the $10,000$ generated phrases. Once again, we add signals from *Full* as baseline. The test is composed of 15 steps with one additional



(a) Cost values histogram.



(b) Cumulative density function for the distance measure.

Fig. 1: Distribution of the DTW costs computed between the two evaluated systems. These figures show that the cost distribution is gaussian-like and have a high number of equivalent samples, based on the distance measure computed.

introduction step. Each step presents a sample from each system. 10 listeners were asked to evaluate the overall quality of the stimuli and give a mark from 0 to 100 (by steps of 5 points).

To validate this new methodology, we perform two other evaluations following the same protocol by taking samples according to the difference measure. A first set is composed of sentences with the lowest DTW costs and a second one, of sentences with a median cost. Once again, phrases with phonetisation issues are removed. Statistics of each test sets are in Table 3. In this table, one can observe the mean DTW cost values and the mean ranks of the different evaluated sets. In particular, it shows that the random set of sentences contains sentences with a medium alignment cost.

The results for this experiment are presented in Table 4. Results normalized according to the full system's Mushra score are presented in Figure 2. As expected, for the corpus composed of samples with the lowest DTW score, evaluations for both reduced systems are almost indistinguishable. For the phrases with a median cost, the mean Mushra scores are not significantly different. At last, contrary to other evaluation, we notice that the three systems are significantly

Test set	Nb sent.	Mean cost (std. dev.)	Mean rank (std. dev.)
Min. cost	15	21.3 (0.8)	9,959.1 (6.6)
Med. cost	15	34.4 (0.0)	4,982.5 (8.8)
Rand. cost	15	35.0 (4.4)	3,886.1 (3,575.1)
Max. cost	15	43.5 (1.4)	10.2 (5.3)
Full	10,000	34.0 (3.4)	5,000.5 (2,886.9)

Table 3: Statistics of the evaluations sets.

different when looking at phrases with the highest alignment distance.

In this case, the *Full* system is still as good as before with a score of 64.5. The second system is *TTSCover* for which some specific features are covered (score of 47.1). This time, there is a significant drop in evaluations of the third system, *CompRand* which is rated with a score of 30.6.

By neglecting some less frequent units, the TTS engine performs well on most of the phrases but the quality may bring down. However, with a training corpus build to cover a broad selection of events, performances are more stable. Consequently, for a given corpus size, applying the covering method is more efficient that a random choice. These results show that covering a useful set of features has a real impact on the perceived quality of the TTS, contrary to what people may expect by just selecting random evaluation samples. It appears that by selecting carefully the samples, we can compare more precisely the two reduced systems and evaluate their differences.

		System		
		Full	*TTSCover*	*CompRand*
Test set	Min. cost	67.3 ± 2.7	50.9 ± 3.0	50.5 ± 3.1
	Med. cost	69.4 ± 3.1	42.5 ± 3.1	47.6 ± 2.5
	Rand. cost	66.5 ± 2.7	44.5 ± 2.7	44.4 ± 2.8
	Max. cost	64.5 ± 3.2	47.1 ± 3.2	30.6 ± 2.6

Table 4: MUSHRA evaluation results for each test set. 95% confidence intervals are also presented.

5. CONCLUSION

In this paper, we have evaluated the impact of the reduction of a spoken corpus onto the perceived quality when it is used in a TTS system. As randomly selecting samples to build

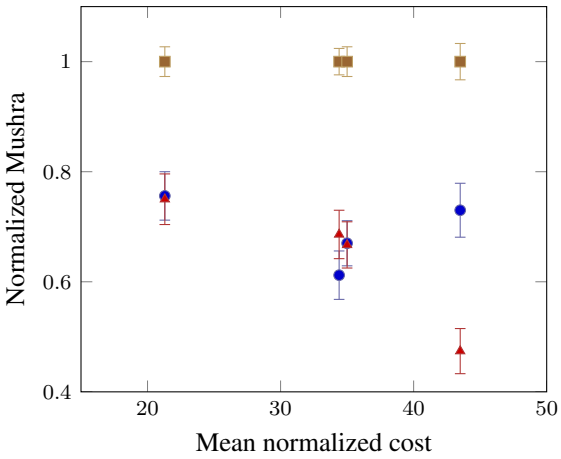


Fig. 2: Results of the perceptive evaluations (and associated 95% confidence intervals) with different samples ordered by the mean normalized cost of the set. Mushra scores are normalized by the score of the system *Full* (squares). *TTSCover* results are represented by the circles and *CompRand* by the triangles

perceptive tests does not enable to make the differences between two systems appear, we have proposed a new way to build subjective tests. The basic principle is to use a measure to evaluate the acoustic differences between the TTS systems output samples. Then to build a test, we have proposed to select the most different samples, thus making no assumption on which one is the best system. We have applied successfully this method with DTW normalized by the length of the alignment path. The results are promising since this method enables to see significant differences between the two systems and conclude that covering a carefully selected set of features has a real impact on the output of a TTS systems. Moreover, the method is applicable on a variety of situations and in particular to evaluate slight modifications on the parameters of a TTS system (concatenative or parametric). Several questions rise from this work and need further investigation. First, the choice of normalized DTW on MFFC vectors as a difference measure has to be compared to other methods. Second, extending the approach to three or more systems would be helpful. Finally, even if this method enables the comparison and distinction between two systems, it is not enough to quantify the performance gap which should be possible by taking into account the score distribution.

REFERENCES

[1] J.-L. Gauvain, L.F. Lamel, and M. Eskenazi, "Design considerations and text selection for bref, a large french readspeech corpus," in *Proc. of ICSLP*, 1990, pp. 1097–1100.

[2] J.P.H. Van Santen and A.L. Buchsbaum, "Methods for optimal text selection," in *Proc. of Eurospeech*, 1997, pp. 553–556.

[3] Hélene François and Olivier Boëffard, "Design of an optimal continuous speech database for text-to-speech synthesis considered as a set covering problem.," in *Proc. of INTERSPEECH*, 2001, pp. 829–832.

[4] A. Krul, G. Damnati, F. Yvon, and T. Moudenc, "Corpus design based on the kullback-leibler divergence for text-to-speech synthesis application," in *Proc. of ICSLP*, 2006, pp. 2030–2033.

[5] Didier Cadic and Christophe D'Alessandro, "Towards optimal tts corpora," in *Proc. of LREC*, 2010, pp. 99–104.

[6] J. Chevelu, N. Barbot, O. Boëffard, and A. Delhay, "Comparing set-covering strategies for optimal corpus design," in *Proc. of LREC*, 2008.

[7] N. Barbot, O. Boëffard, and A. Delhay, "Comparing performance of different set-covering strategies for linguistic content optimization in speech corpora," in *Proc. of LREC*, 2012.

[8] Tanya Lambert, Norbert Braunschweiler, and Sabine Buchholz, "How (not) to select your voice corpus: Random selection vs. phonologically balanced," in *Proc. of SSW6*, 2007.

[9] David Guennec and Damien Lolive, "Unit Selection Cost Function Exploration Using an A* based Text-to-Speech System," in *Proc. TSD*, 2014, pp. 449–457.

[10] Alan W. Black and Paul Taylor, "Chatr: a generic speech synthesis system," in *Proc. of the 15th conference on Computational linguistics*, 1994, pp. 983–986.

[11] H. François and O. Boeffard, "The greedy algorithm and its application to the construction of a continuous speech database," in *Proc. of LREC*, 2002, vol. 5, pp. 1420–1426.

[12] A. Krul, G. Damnati, F. Yvon, C. Boidin, and T. Moudenc, "Adaptive database reduction for domain specific speech synthesis," in *Proc. of the ISCA Research Workshop on Speech Synthesis (SSW6)*, 2007, pp. 217–222.

[13] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 26, no. 1, pp. 43–49, 1978.

[14] ITU-T, "ITU-T recommendation p.800: Methods for subjective determination of transmission quality," 1996.

[15] ITU-R, "ITU-R recommendation bs.1534: Method for the subjective assessment of intermediate quality levels of coding systems," 2003.