How to Compare TTS Systems: A New Subjective Evaluation Methodology Focused on Differences

Jonathan Chevelu[1], *Damien Lolive*[1], *Sébastien Le Maguer*[2], *David Guennec*[1]

[1] IRISA - University of Rennes 1, Lannion, France
[2] Saarland University, Saarbrücken, Germany

jonathan.chevelu@irisa.fr, damien.lolive@irisa.fr,
david.guennec@irisa.fr, slemaguer@coli.uni-saarland.de

Abstract

Subjective evaluation is a crucial problem in the speech processing community and especially for the speech synthesis field, no matter what system is used. Indeed, when trying to assess the effectiveness of a proposed method, researchers usually conduct subjective evaluations by randomly choosing a small set of samples, from the same domain, taken from a baseline system and the proposed one. When selecting them randomly, statistically, samples with almost no differences are evaluated and the global measure is smoothed which may lead to judge the improvement not significant.

To solve this methodological flaw, we propose to compare speech synthesis systems on thousands of generated samples from various domains and to focus subjective evaluations on the most relevant ones by computing a normalized alignment cost between sample pairs. This process has been successfully applied both in the HTS statistical framework and in the corpus-based approach. We have conducted two perceptive experiments by generating more than 27,000 samples for each system under comparison. A comparison between tests involving most different samples and randomly chosen samples shows clearly that the proposed approach reveals significant differences between the systems.

Index Terms: speech synthesis, subjective evaluation

1. Introduction

In the field of Text-To-Speech synthesis (TTS), subjective evaluation is crucial as the main goal is to produce speech targeted at human listeners. Classically, both objective and subjective evaluations can be used. On the one hand, objective evaluations have the good property of being cheap to be made but no matter how pertinent they are, they still cannot replace subjective tests. On the other hand, to be interesting, subjective evaluations need a large number of samples to be evaluated and also a large number of listeners both chosen depending on the application domain of the system.

Several perceptive evaluations are usually used. Among all the methods, we can distinguish preference tests like AB and ABX, score tests like MOS, DMOS and more recently MUSHRA. All these methods serve the same purpose, which is ranking systems according to some subjective criteria.

In the literature, most of the propositions are perceptually evaluated. For instance, for the Blizzard challenge, a large scale evaluation campaign is used [1, 2], but each time the number of utterances under test is restricted. The same is true in the majority of the evaluations done. To cite a few examples, we can mention [3] with 350 sentences, [4] with 7 sentences for 5 systems, [5] with two blocks of 18 stimuli. Usually, the explanation for these low numbers of stimuli is that perceptual evaluations are really time-consuming. Some recent work have questioned the evaluation methodology, like [6] which investigates the impact of listeners mental reference on perceptual tests results, or have proposed protocol modifications as in [5, 7]. Even some alternatives to classic methodologies have also been used, based on crowdsourcing as described in [8] .

More important than the small number of samples chosen, the fact that they are chosen randomly and not for their significance to the evaluated systems may bias the results of evaluations. In this paper, contrary to what is usually done, we propose to synthesize a large number of samples (several thousands), using texts from various domains. Considering the high number of samples, we introduce an alignment cost between samples from paired systems to rank the samples by similarity. Once it is done, we can build a perceptual evaluation using the most different samples. This way, we make no assumption concerning the quality of a system among the other, we simply focus the evaluation on what may make a difference between the systems. Such a strategy enable to reduce the size of a perceptual evaluation to assess the difference significance between systems evaluated. We have used successfully this methodology both with a statistical system (HTS) and a corpus-based one. The results we obtain for AB preference tests are clearly significant while it is not the case when randomly choosing the samples.

The remainder of the paper is organized as follows. In section 3, we present the systems we use in the experiments. Section 4 describes the methodology to build the evaluations. Finally, section 5 presents the experiments as well as the results.

2. Speech corpora

The first corpus is extracted using a fully automatic process presented in [9], from an audiobook in French. The speaker is a male speaker whose reading is moderately expressive and the signal is sampled at 44.1kHz. The full annotated corpus contains 3,339 utterances (10h45 speech). For the experiments, 1h of speech was extracted from the corpus to train the HMM-based synthesis system described later. From now on, this corpus is called *Audiobook*.

The second corpus is spoken by a female speaker in French. Its was initially built for the TTS system of an answering automaton in a Telecommunication framework and its annotations are manually checked. The full corpus contains 7h of speech recorded at 16KHz. From now on, this corpus is called *IVS*.

3. TTS Systems

In order to assess the efficiency of the proposed method, we use two speech synthesis systems. The first one is based on HTS while the second one is a corpus-based speech synthesis system.

3.1. HMM-based synthesis

Over the past decade, the popular HTS framework has been widely used for various studies. Hidden Markov Models based (HMM-Based) speech synthesis [10, 11] has proven to be a very flexible methodology to produce speech. This statistical framework relies on the Hidden semi-Markov model structure to model Mel-Generalized Cepstral (MGC) coefficients, aperiodicity, fundamental frequency (F0) as separate streams using decision trees and a single set of features [11]. We have used the HTS version 2.3 alpha with 50 MGC coefficients, 25 band aperiodicity (BAP) coefficients and log F0.

In this paper, we voluntarily focus on two simple feature sets constituted only by the phonemes labels including the current phoneme label and the context labels using either [-1,1] or [-2,2] windows (i.e. one or two phonemes before and one or two phonemes after the current phoneme). These configurations are chosen according to [12] which evaluates the features used by the standard HTS framework. In this paper, they are found to be the most relevant features but without a high difference during perceptual evaluation. Nevertheless, by applying the methodology we propose, we will show that a significant difference exists.

Using the *Audiobook* corpus we have trained two HTS systems :

- *HMM-p3*: use only current, previous and next phoneme labels as features
- *HMM-p5*: features from *HMM-p3* + phoneme label before previous and after next.

3.2. Corpus-based synthesis

3.2.1. Baseline system

The corpus-based TTS system used in this study is the one described in [13]. The concatenation cost we use in this study takes into account three components which are distances in terms of MFCC, amplitude and $F0$ between two consecutive units. To improve the search speed, a preselection step is done to filter candidate units as proposed in [14]. The filters used act as a binary target cost within the system and the cost function optimized is reduced to a concatenation cost. We assume that two units passing the preselection step are equivalent with respect to the target cost and the target unit. The following features are used in the baseline TTS system:

- Is the phone in the last syllable of its sentence?
- Is the phone in the last syllable of its phrase?
- Is the phone in the last syllable of its word?
- Is the phone in a syllable with rising pitch ?

3.2.2. Systems under comparison

In the context of a speech synthesis system, corpus reduction is a general problem that is of broad interest. As shown in literature, several papers have studied ways to reduce corpora of either speech or text. In particular, [15] proposes an evaluation of the reduction impact on the quality of a TTS system. They show that a randomly selected corpus seems to achieve a similar output quality compared to a corpus built to cover the diphones. This particular point is investigated in the following experiment using our methodology.

Table 1: Main statistics of used corpora.

Sub-corpus	*Full*	*TTSCover*	*CompRand*
Duration	7h06'12	3h11'15	3h04'19
Size in phrases	7,662	3,238	3,350
Size in labels	259,684	112,324	112,324
Labels	34 phonemes and 2 NSS		
Diphonemes	1,242		

The corpus reduction problem can be seen as a set covering problem (SCP) [16]. 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 [17] and [18]. Through the use of Lagrangian relaxation principles, [19] shows that an *Agglomeration* greedy algorithm followed by a *Spitting* greedy *Algorithm* is close to the optimal solution in this framework (this combination is called *ASA* in the remainder of the paper).

To evaluate our methodology, we propose to reduce a speech corpus (the *Full* corpus) following two methods:

- *TTSCover* covers similarly at least once each successive pair of phonemes taking with an associated vectors of features. Features considered are those used in TTS system described in 3.2.1.
- *CompRand* is obtained by randomly complementing a diphoneme covering of *Full* (around 300 sentences) until reaching the same size as *TTSCover* (in number of phones).

Using the *IVS* corpus, two corpus-based TTS systems are then built: they respectively select speech segments in *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.

4. Evaluation Methodology

In this section, we present the proposed evaluation methodology including the text corpus used in the experiments.

4.1. 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 Dynamic Time Warping (DTW) [20]);
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.

4.2. 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 [21]). Since the same phonetizer is used for systems, sentences that introduce phonetization mistakes are filtered (with non-standard symbols or proper nouns). Finally, $27,030$ sentences are extracted randomly from the complete set of sentences to build the test corpus that has to be synthesized.

5. Experiments and results

5.1. Alignment costs repartition

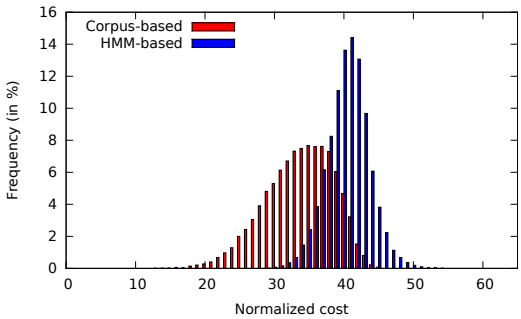
Figure 1 shows the distribution of the DTW costs on the $27,030$ sentences when comparing *TTSCover* and *CompRand* (in red) and when comparing *HMM-p3* and *HMM-p5* (in blue). Considering both the histogram and the density function, we can observe a gaussian-like behavior. The consequence is that when selecting randomly the samples, the resulting set used for perceptive evaluations should contain a high number of equivalent samples. And then, the results of the perceptive evaluation are smoothed by those samples and systems may be considered equivalent.

Note that the costs for HMM-based systems are bigger in mean than those for corpus-based systems. It seems logical since for a sentence from two corpus-based systems build from the same voice, output signals can share significant parts which is not the case for HMM-based systems. So unfortunately, finding a universal threshold on DTW costs from which we could say signals are significantly different could be difficult.

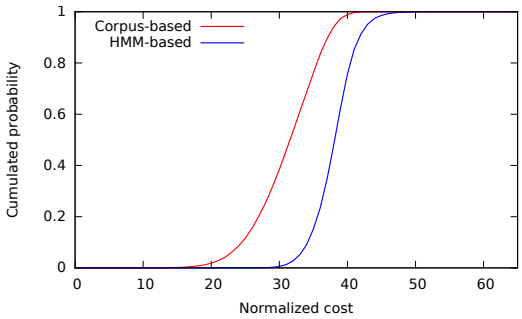
5.2. Perceptive evaluations

To assess the proposed methodology, we conducted separate evaluations for the corpus-based and the HMM-based systems. In the first case, we evaluated three sampling methods. The first test consists of selecting the most similar speech samples according to the proposed measure and is made to verify that the measure correlates to perception in terms of similarity. The second one is the classic method used, i.e. by selecting randomly a subset of samples. Finally, the third one is based on the selection of the most different speech samples. In the second case, for HMM-based systems, we only evaluated a random subset of samples and a subset composed of the most different speech samples. Statistics of each test corpus are presented in table 2. They show a significant difference between the maximum distance selection and the other two methods.

Considering the previous configurations, we extracted 100 samples per system used to build the AB preference tests. At each step, two signals generated from the same sentence but by different systems are presented in a random order. 10 listeners were asked to choose their preferred signal (three answers were provided: A, B and *Indifferent*). The results are presented in tables 3a and 3b.



(a) Cost values histogram.



(b) Cumulative density function for the distance measure.

Figure 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.

Table 2: Statistics of the evaluations corpus.

(a) Corpus-based systems evaluations sets.

Test set	No. of sent.	Mean cost (std. dev.)
Min. dist. corpus	100	15.0 (1.6)
Random corpus	100	31.2 (4.7)
Max. dist. corpus	100	41.6 (0.5)
Full corpus	$27,030$	31.2 (4.9)

(b) HMM-based systems evaluations sets.

Test set	No. of sent.	Mean cost (std. dev.)
Random corpus	100	38.6 (3.3)
Max. dist. corpus	100	48.5 (1.2)
Full corpus	$27,030$	34.0 (3.0)

First, we can observe that when selecting the samples randomly, the systems are not distinguishable and the preference is equally distributed between the three possible answers. This is true both for HMM and corpus-based systems. Moreover, in these cases, the difference between the systems is not significant, according to a binomial test in order to reach a 95% confidence level. A possible explanation is a random selection tends to select samples containing the most frequent events. One may further assume that, on the most frequent events, two comparable systems may behave the same way.

When we select the samples using the ranking method we propose and keep the most different ones, results show clearly a

Table 3: Preference test results

(a) Results for the corpus-based systems. Three AB tests are made by selecting the most similar samples according to the methodology proposed, random samples and the most different samples.

Preferred system	Min. dist. corpus	Random corpus	Max. dist. corpus
TTSCover	27	34	52
CompRand	27	37	32
Indifferent	46	29	16
Significant difference	No	No	**Yes**

(b) Results for the HMM-based systems. Two AB tests are made by selecting random samples and the most different samples.

Preferred system	Random corpus	Max. dist. corpus
HMM-p3	31	26
HMM-p5	41	51
Indifferent	28	23
Significant difference	No	**Yes**

preference for one system. Moreover, in both cases, the number of *Indifferent* answers decreases drastically (e.g divided by 2 for corpus-based systems). For both systems, the results of the perceptive tests are now significant. Consequently, the proposed ranking enables to focus on a subset of samples for which the differences at the acoustic level permit to discriminate the systems evaluated. Furthermore, we can note that no assumption has been made on the quality of the output for the systems.

To complete the evaluation of the method, we have verified on corpus-based systems that taking the most similar samples gives coherent results. In table 3a, we can observe that in this case a large number of samples are judged equivalent (46 Indifferent votes). The rest of the votes are equally distributed between *TTSCover* and *CompRand*. Again, the measure applied does not give any hint on the quality of the samples. To conclude, these results show clearly that carefully selecting the samples is important to obtain significant results.

6. Conclusion

In this paper we have presented a new perceptive evaluation methodology based on a large test set (thousands of samples) and a measure used to rank the paired samples in terms of differences. Then, we suggest that the selected samples have to be the most different ones in order to be able to increase the significance of a perceptual evaluation. This new idea has been applied successfully on both HMM-based systems and corpus-based systems, with different learning voices (an expressive one from a male speaker and a neutral one from a female speaker). Perceptive evaluations have been conducted to compare the classic random selection method to the proposed one. The results show clearly an improvement of the significance of the results and a decrease of the "Indifferent" answers.

This new but simple methodology can then help to effectively validate improvements on TTS systems. It can also be used in an industrial process in order to organize, with a lower cost, non-regression tests between systems versions by spotting sentences with major modifications.

By now, the method has been applied to pair of systems, and future work will be made to extend the method to a higher number of systems. A possible way of doing this is to make a pairwise comparison between the systems and then take the mean rank of the samples to select the globally most different ones. We also plan to compare DTW with other signal distances that could be more correlated with perceptive evaluations.

7. References

[1] S. King and V. Karaiskos, "The blizzard challenge 2012," in *Proc. Blizzard Challenge workshop 2012*, 2012.

[2] K. Prahallad, A. Vadapalli, S. Kesiraju, H. A. Murthy, S. Lata, T. Nagarajan, M. Prasanna, H. Patil, A. K. Sao, S. King, A. W. Black, and K. Tokuda, "The blizzard challenge 2014," in *Proc. Blizzard Challenge workshop 2014*, 2014.

[3] I. Sainz, E. Navas, I. Hernaez, A. Bonafonte, and F. Campillo, "Tts evaluation campaign with a common spanish database," in *LREC*, 2014, pp. 2155–2160.

[4] M.-n. Garcia, C. D'Alessandro, G. Bailly, P. Boula De Mareüil, and M. Morel, "A joint prosody evaluation of french text-to-speech synthesis systems," in *LREC*, 2006, pp. 55–57.

[5] F. Hinterleitner, G. Neitzel, S. Moller, and C. Norrenbrock, "An evaluation protocol for the subjective assessment of text-to-speech in audiobook reading tasks," in *Proc. Blizzard Challenge Workshop*, 2011.

[6] J. Latorre, K. Yanagisawa, V. Wan, B. Kolluru, and M. J. Gales, "Speech intonation for tts: Study on evaluation methodology," in *Proceedings of Interspeech*, 2014.

[7] M. Viswanathan and M. Viswanathan, "Measuring speech quality for text-to-speech systems: development and assessment of a modified mean opinion score (mos) scale," *Computer Speech & Language*, vol. 19, no. 1, pp. 55–83, 2005.

[8] S. Buchholz, J. Latorre, and K. Yanagisawa, "Crowdsourced assessment of speech synthesis," *Crowdsourcing for Speech Processing*, pp. 173–216, 2013.

[9] O. Boeffard, L. Charonnat, S. L. Maguer, and D. Lolive, "Towards fully automatic annotation of audio books for tts," in *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul, Turkey: European Language Resources Association (ELRA), may 2012.

[10] K. Tokuda, H. Zen, and A. W. Black, "An HMM-based speech synthesis system applied to English," in *IEEE Workshop 2002*, 2002.

[11] H. Zen, K. Tokuda, and A. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[12] S. Le Maguer, N. Barbot, O. Boëffard *et al.*, "Evaluation of contextual descriptors for hmm-based speech synthesis in french," in *ISCA Speech Synthesis Workshop (SSW8)*, 2013.

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

[14] A. Conkie, M. C. Beutnagel, A. K. Syrdal, and P. E. Brown, "Preselection of candidate units in a unit selection-based text-to-speech synthesis system," in *Proc. of ICSLP*, vol. 3, 2000, pp. 314–317.

[15] T. Lambert, N. Braunschweiler, and S. Buchholz, "How (not) to select your voice corpus: Random selection vs. phonologically balanced," in *Proc. of SSW6*, 2007.

[16] H. François and O. 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.

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

[18] 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.

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

[20] 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.

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