10th International Conference on Speech Prosody 2020
25-28 May 2020, Tokyo, Japan

Corpus design for expressive speech: impact of the utterance length

Meysam Shamsi[1], *Jonathan Chevelu*[1], *Nelly Barbot*[1], *Damien Lolive*[1]

[1] Univ Rennes, CNRS, IRISA

meysam.shamsi@irisa.fr, jonathan.chevelu@irisa.fr, nelly.barbot@irisa.fr, damien.lolive@irisa.fr

Abstract

Voice corpus plays a crucial role in the quality of the synthetic speech generation, specially under a length constraint. Creating a new voice is costly and the recording script selection for an expressive TTS task is generally considered as an optimization problem in order to achieve a rich and parsimonious corpus.

In order to vocalize a given book using a TTS system, we investigate four script selection approaches. Based on preliminary observations, we simply propose to select shortest utterances of the book and compare the achievements of this method with state of the art ones for two books, with different utterance lengths and styles, using two kinds of concatenation based TTS systems.

The study of the TTS costs indicates that selecting the shortest utterances could result in better synthetic quality, which is confirmed by a perceptual test. By investigating usual criteria for corpus design in literature like unit coverage or distribution similarity of units, it turns out that they are not pertinent metrics in the framework of this study.

Index Terms: Text to speech, voice corpus design, utterance length

1. Introduction

The generation of an audiobook generally needs an expensive recording phase with a professional speaker. In order to reduce the recording duration, the use of a Text-to-Speech (TTS) system could be a lead. So that the TTS system provides a high quality expressive voice, for a given speaker, the construction of a specific voice may be necessary. For a given book to vocalize, this problem can be considered as an expressive voice design task for a specific-domain TTS. Although, in TTS, vocoder-based approaches – like end-to-end DNN systems – are more and more prevalent, hybrid or classical unit selection-based systems are still well-adapted to take into account the data parsimony constraint. However, their achievements are very sensible to the voice quality and the impact of the voice is all the stronger as the constraint of parsimony is important [1, 2, 3].

To limit the recording and post-processing costs due to the voice building while guaranteeing a good quality, it is customary to carefully craft the recording script that will be performed by the speaker. The main proposed approaches consist in extracting, from a large textual corpus, for instance the target book to be vocalized, a minimal subset of sentences that maximizes an optimisation criterion. This criterion is often related to the maximization of the linguistic coverage [4, 5, 6] (formalized as a set covering problem) or the closeness to a target linguistic distribution [7, 8]. Different algorithms have been compared and the mainly used approach is the greedy one, providing a good trade-off between the computational time and closeness to the optimal solution [6].

Moreover, during the voice creation process, several kinds of linguistic features were considered but rarely compared. They could be symbolic as diphoneme [4] or triphoneme [9] labels, phonetic ”sandwiches” [10], etc. Ones may add some positional characteristics to these units [2] or some stress information [1]. Some recent works observe an improvement by using an embedding representation of the linguistics markers built thanks to a convolutional neural network (CNN) [11, 12].

Some studies, as in [6], point out that the designed voices tend to be composed of utterances shorter than those of the initial pool. In [4], the set covering problem is dealt with a greedy strategy which selects a sub-corpus with an average length of 20 phonemes per sentence out of an initial corpus with an average length of 74 (this approach will be named *set covering* in the following). It is also noticed in [10], and authors proposed to correct the algorithm to force longer sentences. In these cases, it may be explained by the expert function that is locally optimized by the greedy algorithms. It would then be a bias of the algorithms and not a trend from the data to achieve better quality. On the other hand, the CNN-KMeans method in [11] is completely unsupervised but it also selects shorter sentences nonetheless. Moreover, as illustrated by Figure 1, this trend is reinforced when the reduction constraint is strong. This figure compares the evolution of the average utterance length – in number of phoneme instances – of the subset selected by the set-covering approach with the CNN-KMeans one with respect to the reduction rate of the initial corpus.

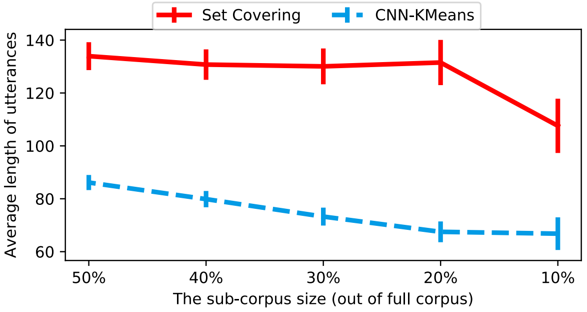


Figure 1: *Average sentence length of sub-corpora provided by two reduction algorithms and at various reduction rates of the Pod corpus (see Section 2 for details on algorithms and corpus). The best system selects shorter utterances. Besides, increasing the parsimony size constraint involves a decreasing of the length of the selected sentences.*

These observations lead us to question ourselves about the impact of the average sentence length of the voice on the final TTS quality: is it a consequence of the optimization or a cause of the good results? Let us assume that it is a cause and name it the ”shortest” hypothesis. If a voice created by selecting only the shortest sentences is less good than a voice from

10.21437/SpeechProsody.2020-195

another strategy, it will allow us to discard this hypothesis. On the contrary, if all attempts show that the “shortest” strategy is better, it will give us clues that this hypothesis may be true and encourage us to investigate further.

In this paper, we will test this “shortest” hypothesis by simulating a voice creation process on different kinds of books (one with long formal sentences, recorded by male speaker and one with shorter and less formal sentences, recorded by a female speaker) with two TTS engines (one expert unit selection TTS and one hybrid TTS). We will compare one of the best reduction strategy proposed so far, i.e. the CNN-KMeans introduced in [11], with a simple ”shortest first” algorithm using automatic measures and perceptual evaluations. Then we will investigate if the classical optimisation criteria – linguistic unit coverage or distribution – can predict or explain the observed results.

2. Data and systems

2.1. Reduction algorithms

To simulate the corpus design process by reducing a full corpus, four approaches are considered and introduced below.

Random: this simple baseline consists in selecting a subset randomly until the requested length is reached. Since this approach is less stable by design, statistics resulting from this method and detailed further are consolidated by repeating this selection process six times. Each utterance in the test section will also be synthesized six times and the associated average score will be taken into account for the objective evaluations.

Set covering: one classic corpus design approach formalises the task as a set covering problem. It can be approximately solved using greedy strategies [4] or Lagrangian relaxation [6]. A greedy based approach is used here, as presented in [11]. The attributes considered for the coverage are the diphone labels enhanced with 20 linguistic features. Those are Boolean variables answering questions like *”it is or not the first/second phone, in the first/last syllable?”*.

CNN-KMeans: this approach, proposed by [11], employs an embedding representation of several linguistic features to characterize utterances. The embedding space is produced by a multi-layer CNN auto-encoder implemented to project the discrete features into a continuous space. Then, for each utterance, the average vector of its embedded unit sequence is computed and used as its representation. A KMeans algorithm clusters utterances and for each cluster, the closest utterance to its center is selected. In [11], this method, called *CNN-KMeans* in the following, outperforms the set covering approach.

Shortest: as presented in the introduction, to synthesize an expressive text like a book, our assumption is to use the shortest utterances first. To assert it, we propose a system named *Shortest*. Its algorithm is basically a simple loop that selects the shortest utterance until the desired length of the selected sub-corpus is reached.

2.2. Corpora

Two French audio-books are used as initial corpora for experiments. The *Pod* corpus is *Albertine disparue* by Marcel Proust [13]. The *Nad* corpus is *La Vampire* by Paul Féval [14]. While *Pod* contains long formal utterances, *Nad* contains more contemporary content with simpler utterances. The average length of utterances in *Nad* is less than half the one in *Pod*. Their main properties are summarized in table 1.

Corpus	**Pod**	**Nad**
Speaker gender	Male	Female
Number of utterances	3339	6032
Average length of utterances	120.1±3.2	54.4±1.2
Duration	10h 44min	10h 02min
Number of distinct diph.	1005	1000
Number of distinct triph.	12655	4693

Table 1: *The initial corpus details*

2.3. TTS engines

Two types of TTS systems are used for synthesis in our experiments.

The first one, is a standard unit selection engine [15] with a beam search algorithm. The global cost function optimised by the TTS is a weighted sum between a concatenation and a target cost. The concatenation cost is a weighted distance between some acoustic features (MFCC, F0, amplitude, etc.). The target cost is a weighted distance between linguistic features (phoneme, syllable, positioning information, etc.). In this system, all weights were manually tuned over time. It then will be called *expert* TTS in the remainder.

More recently, most unit selection systems shifted to an hybrid architecture that includes DNN to learn the cost functions [16]. Following this trend, the second system for the experiments, called *hybrid* TTS, is inspired by [17]. Its target cost is computed based on an euclidean distance in an embedding space. This embedding is learned from an encoder-decoder trained on the voice.

In the following experiments, the *hybrid* TTS uses only one DNN per speaker to compute the target cost. From our experience, the bias it may introduce is not significant and it allows to directly compare all costs between sub-voices from the same corpus. It also helps to discard noise from the DNN training initialisation.

3. Experimental setup

Two audio-books with almost same length (around 10 hours) are provided as the initial corpora. A 10-fold cross validation without shuffling is used for separating the full corpus (90%) and test section (10%). Each fold is continuous, like a chapter, and the first fold starts with the first utterance in the book. Finally, the initial corpora will be synthesized by different full corpora and sub-corpora. The length of the selected sub-corpus is fixed to 10% of full corpus (about 1 hour).

The remainder of this section will describe the objective measures which are used to approximate the quality of sub-corpora and the synthetic quality.

3.1. Objective measures

It is inevitable to ask listeners for comparing the quality of the synthetic signals but listening tests are costly and need enough listeners. Same as [11], we propose to use TTS costs as the objective measures to approximate the quality of synthetic signals.

The concatenation and global costs of TTS systems provide a good approximation of the perceptual quality. The global cost is a linear combination of the concatenation and target costs.

The global cost and concatenation cost of the synthetic signal of test section utterances are normalized by their length. The

Corpus	Pod	Nad
Full corpus	120.1±3.2	54.4±1.2
Random	121.1±5.0	54.9±1.8
Set covering	163.2±10.3	95.0±4.7
CNN_KMeans	86.7±4.5	38.6±1.5
Shortest	44.5±1.2	22.0±0.5

Table 2: *The average length (number of phones) of selected utterances for 10% of full corpus*

normalized costs average over utterances are used to compare different corpus design methods for each TTS/corpus.

3.2. Perceptual evaluation

The synthetic signals resulting from *CNN-KMeans* as the state of the art method are compared with the *Shortest* method ones. By running 10-fold cross validation a sub-corpus is extracted from each full corpus to synthesize the corresponding test section for each fold. It will provide a synthetic signal of the whole book. As mentioned in table 1, there are 9371 utterances in the two corpora. However, the excessive length of some utterances may be problematic for listeners to compare the signals. For instance, the longest sentence in *Pod* is 238 words long (82 seconds). Consequently, each utterance is split into breath groups. Breath groups shorter than a minimum length (20 phonemes) are merged with the following breath group. It provides 37711 breath groups that have been synthesized according to the corresponding selected voice using *hybrid* and *expert* TTS for each fold of the cross validation. Based on the idea of [18], to avoid smoothing the results, pairs of signals that are too similar (DTW < 1.0) have been removed. Then, 100 sample pairs have been selected randomly from remaining candidates as the listening test samples. Half of these samples has been selected from *Pod* corpus and half from *Nad* corpus. Listeners evaluate 40 pairs of synthetic signals on a 5 points scale. At each step of the test, the script of the full utterance corresponding to the signal is displayed, even if the signal is only a part of the utterance. The pronounced part is highlighted to help listeners evaluate the overall quality of samples by considering the context.

4. Results

Methods mentioned in section 2 have been run using 10-fold cross validation to select 10% of the full corpus. The average length of selected utterances by the different selection methods are compared in table 2. In French, the average length of sentences depends on the context. For instance, the average length of sentences in *Le Monde*, whose context is French newspaper, is around 98 phones [19]. This length for the *SynPaFlex* corpus, which contains novel books and poems, is 48 phones [14].

4.1. Objective measures

The selected sub-corpus voices have been used to synthesize the test section of the 10 folds. The average global cost normalized by length (number of phones) of synthetic signals is shown in figure 2. Given that the same behavior is observed with the concatenation cost, it is not shown here.

The resulting voice from the *Shortest* method succeeds to synthesize signals with lowest global costs. The resulting signals from the *SC* method have higher global costs than even *random* method. It shows that however following a set covering strategy will guarantee all units to be covered, the resulting TTS costs would be worst than random method for big enough voice corpora.

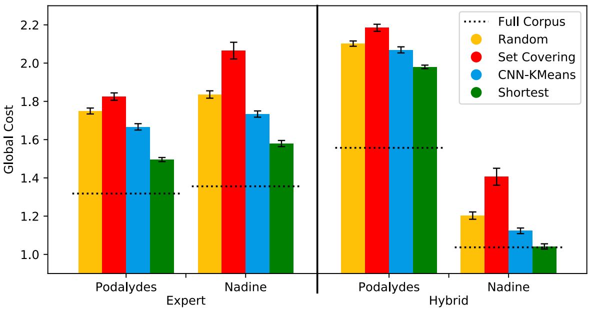


Figure 2: *Average TTS global cost per phone after a* 10*-fold cross validation. Shortest gives the best results in all cases.*

The voice corpus built with short utterances are expected to be less efficient for synthesizing long utterances [20]. To investigate this assumption, the correlation coefficients between the length of utterances and the TTS costs of the corresponding synthetic signals have been calculated. The Pearson correlation coefficients for the global and concatenation cost of both TTS are less than 0.12. This means even by selecting short utterances for voice corpus, TTS systems are able to synthesize long utterances almost with same cost.

4.2. Perceptual evaluation

Based on the TTS cost results, an AB preference test has been conducted to compare two best corpus design methods. 200 synthetic signals have been selected from the *Shortest* and the *CNN-KMeans* methods. For each combination of TTS and book, 50 signals have been chosen as the perceptual test samples.

In total, 12 listeners have compared pairs of synthetic signals. Each pair has been evaluated at least 2 times. Results are shown in figure 3. The perceptual results confirm the results obtained with the TTS costs and the superiority of the *Shortest* method for both corpora and TTS systems.

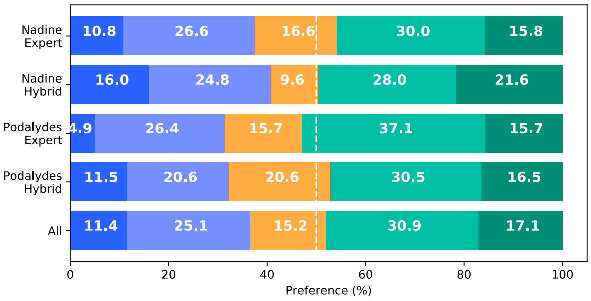


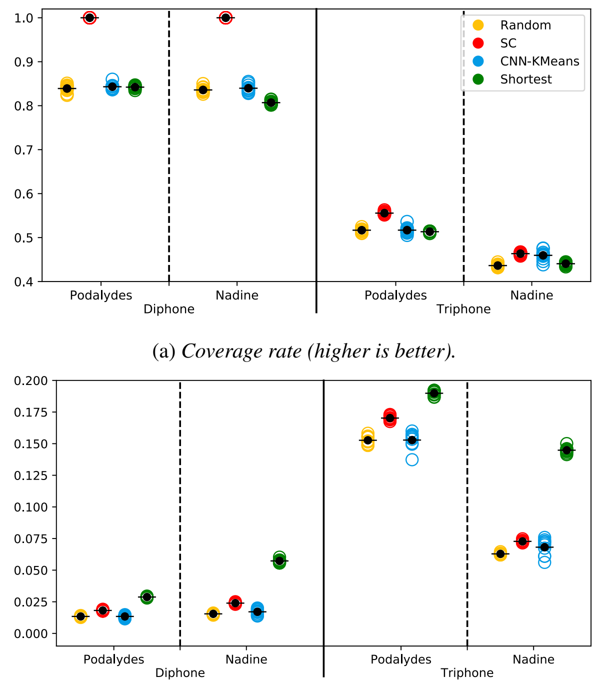
Figure 3: *The Perceptual test results. Right to left: strongly CNN-KMeans (dark blue), slightly CNN-KMeans (light blue), no preference (yellow), slightly Shortest (light green), strongly Shortest (dark green).*

5. Analysis

5.1. Coverage rate and distribution similarity

Other measures need to be considered to evaluate the selection method, such as coverage rate of units, or distribution similarity of units with a target distribution. The first one, the coverage rate, is defined as the number of distinct diphones/triphones which exist in the selected sub-corpus per total number of distinct diphones/triphones in the full corpus. The second one, the distribution similarity of diphones/triphones in the sub-corpus with the full corpus, is evaluated using a Kullback-Leibler Divergence (KLD). The KLD indicates the dissimilarity between two distributions. Some studies claim that a lower KLD with a target distribution will result in better sub-corpora [7, 8].

Figure 4 compares the coverage rate and distribution similarity of four methods. The top figure is the diphones and triphones coverage rate in selected sub-corpus by the different methods for the two corpora. The bottom figure is the KLD between diphone/triphone distribution of sub-corpus and the full corpus. As the KLD value decreases, the selected sub-corpus distribution is increasingly similar to the one of the full corpus. Each colored circle indicates a selected part from one fold of the full corpus.



(a) *Coverage rate (higher is better).*

(b) *KLD (lower is more similar to target distribution).*

Figure 4: *Coverage rate and KLD of diphone/triphone for 10-fold cross validation. Average of each column is in black.*

Based on table 1, however the number of distinct diphones in two corpora are similar, the number of the distinct triphones in *Pod* corpus is almost three times higher than in *Nad* corpus. According to figure 4a, the coverage rate of the *Shortest* method is almost same as *CNN-KMeans* and *Random* methods. It means the short utterances do not contain a set of specific units and they are as good as random in terms of unit coverage for 1 hour of sub-corpus. However the diphone coverage of *Nad* corpus with the *Shortest* method is slightly lower than others.

Based on Figure 4b, it could be observed that the *Shortest* method does not respect the general distribution of corpora. While the random selection method achieves the lowest KLD, the *Shortest* method results in the highest KLD in both corpora. It is not surprising to have the same distribution as full corpus by the random selection.

5.2. Properties of short utterances

As it is mentioned in [20, 21], short utterances are often more expressive and have a different prosodic delivery. Contrary to [22], the main idea in *Shortest* method is to have more possible prosodic variation in the voice corpus.

However the *Shortest* method can not guarantee the coverage of all diphones or phones, we hope the sub-corpus length is long enough to cover all needed phones. The alternative solution would be replacing the not selected shortest utterance which contain the missed phones with the longest utterances in the selected sub-corpus.

We can mention that the short utterances are easy to read in the recording process. A drawback is that the *Shortest* method will select repetitive sentences. However, in term of linguistic information, same utterances do not add new units to corpus, they can contain different acoustic information. For example there are 5 utterances with same script (*"Ah"*) but they are completely different in terms of intonation.

As a first investigation, we find more variation of F0 in the voice corpus obtained with the *Shortest* method than others. It emphasizes the importance of acoustic and prosodic variation of the sub-corpora containing short utterances.

6. Conclusion

In this study four methods for TTS voice corpus design have been compared. These methods are evaluated with two kinds of TTS and for synthesizing two french audio-books. The synthetic signals obtained these methods have been compared objectively using TTS costs and perceptually by listeners.

The experimental results show a simple method like selecting short utterances could work well for TTS voice corpus design. This method works better than *CNN-KMeans* method in *hybrid* and *expert* TTS for audio-book with long and short utterances. The results show that the coverage of units as the classical method does not work even as good as random selection in a large enough voice corpus. By comparing the TTS cost and the coverage rate and KLD as unit distribution similarity, it revealed that the previous strategies of corpus design [7, 6] does not lead to the best voice corpus. They are not necessarily a good metric of corpus design for big enough voice corpora in TTS.

The results and the performance of the *Shortest* method should be tested with more corpora with different average utterance length. As future work, a combined method can be proposed which takes into account the average length of utterances in book. In other words, it could be more efficient to adapt the selection process to the context and the characteristics of book.

7. Acknowledgements

This research was conducted under the ANR (French National Research Agency) project SynPaFlex ANR-15-CE23-0015.

8. References

[1] T. Lambert, N. Braunschweiler, and S. Buchholz, "How (not) to select your voice corpus: random selection vs. phonologically balanced," in *Sixth ISCA Workshop on Speech Synthesis (SSW6)*, Bonn, Germany, August 2007, pp. 264–269.

[2] J. Chevelu and D. Lolive, "Do not build your TTS training corpus randomly," in *Proceedings of the 23rd European Signal Processing Conference (EUSIPCO)*. Nice, France: IEEE, September 2015, pp. 350–354.

[3] K. Szklanny and S. Koszuta, "Implementation and verification of speech database for unit selection speech synthesis," in *Proceedings of the Federated Conference on Computer Science and Information Systems (FedCSIS)*. Prague, Czech Republic: IEEE, September 2017, pp. 1262–1267.

[4] H. François and O. Boëffard, "Design of an optimal continuous speech database for text-to-speech synthesis considered as a set covering problem," in *Seventh European Conference on Speech Communication and Technology*, Aalborg, Denmark, September 2001.

[5] B. Bozkurt, O. Ozturk, and T. Dutoit, "Text design for TTS speech corpus building using a modified greedy selection," in *Eighth European Conference on Speech Communication and Technology*, Geneva, Switzerland, September 2003, pp. 277–280.

[6] N. Barbot, O. Boëffard, J. Chevelu, and A. Delhay, "Large linguistic corpus reduction with SCP algorithms," *Computational Linguistics*, vol. 41, no. 3, pp. 355–383, 2015.

[7] A. Krul, G. Damnati, F. Yvon, and T. Moudenc, "Corpus design based on the kullback-leibler divergence for text-to-speech synthesis application," in *Ninth International Conference on Spoken Language Processing (ICSLP)*, Pittsburgh, PA, USA, September 2006, pp. 2030–2033.

[8] Y. Shinohara, "A submodular optimization approach to sentence set selection," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Florence, Italy: IEEE, May 2014, pp. 4112–4115.

[9] M. Isogai and H. Mizuno, "Speech database reduction method for corpus-based TTS system," in *Eleventh Annual Conference of the International Speech Communication Association (InterSpeech)*, Makuhari, Chiba, Japan, September 2010, pp. 158–161.

[10] D. Cadic, C. Boidin, and C. d'Alessandro, "Towards optimal TTS corpora," in *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, Malta, May 2010, pp. 99–104.

[11] M. Shamsi, D. Lolive, N. Barbot, and J. Chevelu, "Corpus design using convolutional auto-encoder embeddings for audiobook synthesis," in *Twentieth Annual Conference of the International Speech Communication Association (InterSpeech)*, Graz, Austria, September 2019, pp. 1531–1535.

[12] ——, "Investigating the relation between voice corpus design and hybrid synthesis under reduction constraint." in *International Conference on Statistical Language and Speech Processing, LNCS/LNAI*, vol. 11816. Springer, Cham, October 2019, pp. 162–173.

[13] O. Boeffard, L. Charonnat, S. L. Maguer, D. Lolive, and G. Vidal, "Towards fully automatic annotation of audio books for TTS." in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, May 2012, pp. 975–980.

[14] A. Sini, D. Lolive, G. Vidal, M. Tahon, and E. Delais-Roussarie, "Synpaflex-corpus: An expressive french audiobooks corpus dedicated to expressive speech synthesis." in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC'18)*, Miyazaki, Japan, May 2018, pp. 4289–4296.

[15] P. Alain, N. Barbot, J. Chevelu, G. Lecorvé, C. Simon, and M. Tahon, "The IRISA text-to-speech system for the blizzard challenge 2017," in *Blizzard Challenge 2017 workshop*, Stockholm, Sweden, August 2017.

[16] S. King, J. Crumlish, A. Martin, and L. Wihlborg, "The blizzard challenge 2018," in *Blizzard Challenge 2018 workshop*, Hyderabad, India, September 2018.

[17] A. Perquin, G. Lecorvé, D. Lolive, and L. Amsaleg, "Phone-level embeddings for unit selection speech synthesis," in *International Conference on Statistical Language and Speech Processing, LNCS/LNAI*, vol. 11171. Springer, Cham, October 2018, pp. 21–31.

[18] J. Chevelu, D. Lolive, L. S. Maguer, and D. Guennec, "How to compare TTS systems: a new subjective evaluation methodology focused on differences," in *Sixteenth Annual Conference of the International Speech Communication Association (InterSpeech)*, Dresden, Germany, September 2015, pp. 3481–3485.

[19] L. F. Larnel, J.-L. Gauvain, and M. Eskenazi, "BREF, a large vocabulary spoken corpus for french," in *Second European Conference on Speech Communication and Technology*, Genova, Italy, September 1991, pp. 505–508.

[20] J. Kominek and A. W. Black, "CMU arctic databases for speech synthesis," Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, USA, Tech. Rep. CMU-LTI-03-177, 2003. [Online]. Available: $http://www.festvox.org/cmu-arctic$

[21] M. Charfuelan and M. Schröder, "Correlation analysis of sentiment analysis scores and acoustic features in audiobook narratives," in *Fourth International Workshop on Corpora for Research on Emotion Sentiment & Social Signals (ES3)*. Istanbul, Turkey: Citeseer, 2012, pp. 99–103.

[22] N. Braunschweiler and S. Buchholz, "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)*, Florence, Italy, September 2011, pp. 1821–1824.