Script Optimization for the Expressive Synthesis of Audio-books

Meysam Shamsi

Univ Rennes, CNRS, IRISA

meysam.shamsi@irisa.fr

1. Introduction

The synthetic speech quality is strongly affected by the quality of the corpus used to build the voice. Previous studies [1, 2, 3] have shown that a random selection is not efficient to design such speech corpora. Moreover, the corpus should be as small as possible in order to minimize the human cost of high quality recording and labeling checking stages. Removing redundant elements while adding critical ones to the corpus is important. A well-designed corpus combines parsimony and balanced unit coverage in order to gain a satisfactory level of richness with a minimal cost construction.

The main point in this thesis is the design of the recording script to improve the Text-to-Speech (TTS) quality in the specific case of expressive audio-book generation when the target book is known in advance. The script will then be composed of a part of the target book to vocalize. The other part will be vocalized using a TTS system based on the associated recording speech data. Therefore, the audio-book will be a mixing of natural and synthetic parts. The recording script should be optimized to provide the best trade-off between its length (or its human cost construction) and the overall quality of the audio-book.

Covering the linguistic units under a parsimony constraint is the main idea of script corpus design. The most commonly used algorithmic strategy for the set covering problem is the greedy approach which provides solutions close to optimal ones [4]. Some studies [5, 6, 7] investigated the distribution of units in the corpus. [5] suggested to design TTS corpora which minimize the Kullback-Leibler Divergence (KLD) between its diphoneme and triphoneme distribution and a prior distribution.

There are still remaining challenges in the corpus design which will be considered in the next section.

2. Challenges

In audio-book generation, recording a long script by speakers is costly and time-consuming. The main objective of this thesis is to reduce the recording cost with minimum degradation of the quality of final audio-book.

The final audio-book is a mix of recorded and synthetic signals. It is not feasible to test and perceptually evaluate all the combinations of the recorded and synthetic portions for all lengths of recorded parts.

Generally the challenges can be categorized in two main categories considered below: evaluation of the speech quality and speech corpus design methodology, which selects a sub-set of the book for recording.

2.1. Evaluation of quality

Although the perceptual test is inevitable for any final conclusion, it is costly and requires a sufficient number of listeners. An alternative objective measure could be used as an estimation of perceptual quality.

Firstly a general measure of synthetic signal quality is needed which would take into account the naturalness and intelligibility of synthetic signals. However, given the fact that the expressiveness plays an important role for the quality of an audio-book, any synthetic signal should be evaluated in the context of the script. This means that a short synthetic signal is not appropriate to evaluate expressiveness, whereas it is difficult for listeners to evaluate a long signal.

Finally, we should investigate the configuration of whole audio-book as a mix of synthetic and recorded signal. The order of the recorded and synthetic signal could impact on the pleasantness of listening, especially when the synthetic signal is not as good as the recorded signals. It becomes even more challenging in the case of multi-characters story books with imitations.

2.2. Speech corpus design

If we assume that expert speakers produce ideal quality, especially in an expressive domain like audio-book, it can be claimed that the best quality is achieved when recording the entire script. Therefore the basic assumption is that the quality will be degraded with less recorded data.

In our approach, firstly we need to determine the suitable length of recording. It should be long enough to cover the phonological variety but not too long in terms of recording cost. For instance as an initial experiment, we have found that selecting half of the script in a book with about 3000 utterances provides same quality as full script. It means that adding more speech data does not provide a significant quality improvement for listeners. Nevertheless this result could change for other TTS systems or books.

The other challenge to be considered is the length of the shortest part selected for corpus design.The shortest part can be paragraph, utterance, breath groups or a window of words. Very short parts could be ambiguous for speakers in terms of context, whereas, on the other hand long parts could contain unnecessary elements.

As it is mentioned earlier it can not be guaranteed that even a high quality recorded signal would be used for synthesizing the rest. Particularly in our problem, there are two different factors for corpus design; the selected part which should be recorded as the speech corpus, and the rest which would be the target of synthesizing process. It means, the richness of first part and the difficulty of synthesizing the second part are important.

In the following sections, the proposed method and the results, the future works, and contributions will be reviewed.

3. Experiment

3.1. Corpus design method

As a first step, we simplify the problem by dividing the book into a speech corpus and a test section. The objective is to use a part of the corpus to synthesize the test section.

A test set ($\mathcal{T}$) which is randomly selected as a continuous part (10% of the whole corpus). The rest of the audio-book is named the full corpus and is denoted $\mathcal{F}$ in the remainder. The objective is to extract a subset $\mathcal{S}$ from $\mathcal{F}$ with a certain length to synthesize $\mathcal{T}$. In this approach the different corpus design methods can be easily compared by synthesizing a same script from $\mathcal{T}$. We proposed to represent utterances by an embedding vector in a continuous space. This vector is used instead of phonological labels to evaluate the linguistic content of an utterance. The main idea is to select some utterances which cover the diversity of vectors in the embedding space. We proposed to employ Deep Neural Networks (DNNs) and particularly auto-encoders as linguistic feature extraction/selection method. The encoder part of the auto-encoder gives the opportunity of transforming utterances in an embedding space with latent features. In [8], we proposed two methods for selecting ($\mathcal{S}$) from ($\mathcal{F}$). The first idea is applying K-Means algorithm on utterance representation in embedding space to choose the closest utterances to cluster centers. We assume that the center of each cluster represents the information of other utterances of its cluster. The second idea for utterance selection is minimizing KLD between $\mathcal{S}$ and a prior distribution. An agglomerative greedy strategy is used to minimize KLD.

Moreover a set covering method [4] and the classical minimization of KLD [5] have been compared with the proposed method.

We compared the synthetic signals which result from $\mathcal{S}$ based on different neural network architectures. The comparison was objective and perceptual. As an objective measure, we proposed to use the TTS costs (the global cost which is a combination of target and concatenation cost in unit selection TTS).

3.2. Results

In order to have an embedding model, we compared several models such as Deep Convolutional Neural Network (DCNN) auto-encoder, DNN auto-encoder, LSTM sequence to sequence model [9], and Doc2vec model [10] with different hyper parameters.

Utterances of $\mathcal{F}$ have been selected by K-Means and KLD minimization methods for five reduction rate (10%, 20%, 30%, 40%, 50% of the length of $\mathcal{F}$). The TTS global cost demonstrates that CNN auto-encoder which is followed by two selection methods (*CNN-KMeans* and *CNN-KLD*) achieves the lowest values which can be interpreted as the highest synthetic quality. We compared perceptually the synthetic signals from the different systems. The perceptual test showed that listeners prefer the quality of CNN coupled with KMeans or KLD rather than the set covering method [4]. The listening tests also indicated that the TTS global cost can be a pertinent measure to evaluate overall quality of synthetic signals.

Some additional statistics and analyzes of the sub corpus selected by these methods have been reported in [11].

3.3. Supplementary experiment

Based on the first experiment, the embedding space was found interesting for corpus design. We proposed to use the euclidean distance between phones in the embedding space for calculating the TTS target cost during unit selection. The performance of the TTS system using expert target cost and two kinds of embeddings as its target cost is evaluated [12]. We compared a model taking only linguistic information into account with a model using both linguistic and acoustic information.

In this experiment we compared the CNN auto-encoder based on linguistic information at utterance level with a feed-forward DNN using acoustic information at the frame level and its corresponding linguistic information [13]. The listening test results show that the TTS target cost calculated by the embedding model which is derived by only linguistic information has better performance than expert target cost. But the DNN acoustic model is preferred to the linguistic model. The preference of listeners emphasizes the importance of acoustic information in TTS target cost. This result lead us to use acoustic information besides the linguistic information for corpus design.

4. Future work

As future work, we will use the acoustic information in order to define a general acoustic model. This acoustic model can be used as an embedding model for the presented selection methods.

Technically speaking, with recent advances in deep learning, we will implement new models such as attention models [14] in order to have good representation of unit embeddings.

Afterward we will focus on the main problem which is designing a corpus for a specific script. In other words we should evaluate the subset $\mathcal{S}$ according to the synthetic quality of $\mathcal{F}-\mathcal{S}$ and not $\mathcal{T}$. Consequently, the subset selection has to take into account the rest of the book which is not selected for recording.

The configuration of the partitions is another problem that should be considered in overall quality evaluation of audio-books [15]. The order of synthetic and natural voice in the final audio-book could affect the preference of listeners or users. This point could be generalized to expressive utterances in audio-book context. It means a synthetic utterance would need to be uttered with a specific emotion or style depending on the context in the audio-book.

5. Contributions

we have presented an end-to-end method for sentence selection. We have shown that a CNN auto-encoder can be used successfully to extract linguistic information in TTS corpus design. The K-Means clustering and the KLD methods work properly using embedded representations achieving better results than random, or even than the best methods in state of the art such as set covering. The proposed method could be applied to other sub-set selection problems, especially for sets of sequential data.

The perceptual test shows that the TTS global cost can be used as an alternative to synthetic overall quality.

We have investigated the relation between the corpus design process and an hybrid TTS. The TTS voice corpus has been selected based on an embedding model which uses the phonological information of the full corpus. This embedding model can be applied instead of the expert TTS cost or an acoustic model of phonemes. It has then be used to build an hybrid system by computing the target cost function as the euclidean distance between units in the embedding space.

Results of these experiments have been reported in the following papers [8, 11, 12].

6. Acknowledgements

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

7. References

[1] B. Bozkurt, O. Ozturk, and T. Dutoit, “Text design for TTS speech corpus building using a modified greedy selection,” in *8^th European Conference on Speech Communication and Technology*, 2003, pp. 277–280.

[2] M. Isogai and H. Mizuno, “Speech database reduction method for corpus-based TTS system,” in *11^th Annual Conference of the International Speech Communication Association*, 2010, pp. 158–161.

[3] J. Chevelu and D. Lolive, “Do not build your TTS training corpus randomly,” in *23^rd European Signal Processing Conference (EUSIPCO)*. IEEE, 2015, pp. 350–354.

[4] 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.

[5] A. Krul, G. Damnati, F. Yvon, and T. Moudenc, “Corpus design based on the kullback-leibler divergence for text-to-speech synthesis application,” in *9^th International Conference on Spoken Language Processing (ICSLP)*, 2006, pp. 2030–2033.

[6] Y. Shinohara, “A submodular optimization approach to sentence set selection,” in *International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE*. IEEE, 2014, pp. 4112–4115.

[7] T. Nose, Y. Arao, T. Kobayashi, K. Sugiura, and Y. Shiga, “Sentence selection based on extended entropy using phonetic and prosodic contexts for statistical parametric speech synthesis,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 5, pp. 1107–1116, 2017.

[8] M. Shamsi, D. Lolive, N. Barbot, and J. Chevelu, “Corpus design using convolutional auto-encoder embeddings for audio-book synthesis,” in *INTERSPEECH*, 2019.

[9] I. Sutskever, O. Vinyals, and Q. V. Le, “Sequence to sequence learning with neural networks,” in *Advances in neural information processing systems*, 2014, pp. 3104–3112.

[10] Q. Le and T. Mikolov, “Distributed representations of sentences and documents,” in *International Conference on Machine Learning*, 2014, pp. 1188–1196.

[11] M. Shamsi, D. Lolive, N. Barbot, and J. Chevelu, “Script selection using convolutional auto-encoder for tts speech corpus,” in *21^st International Conference on Speech and Computer (SPECOM)*, 2019.

[12] ——, “Investigating the relation between voice corpus design and hybrid synthesis under reduction constraint,” in *7^th International Conference on Statistical Language and Speech Processing (SLSP)*, 2019.

[13] A. Perquin, G. Lecorvé, D. Lolive, and L. Amsaleg, “Phone-level embeddings for unit selection speech synthesis,” in *6^th International Conference on Statistical Language and Speech Processing (SLSP)*, P. G. Dutoit T., Martín-Vide C., Ed., vol. 11171. Springer, Cham, 2018, pp. 21–31.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[15] L. Gong and J. Lai, “To mix or not to mix synthetic speech and human speech? contrasting impact on judge-rated task performance versus self-rated performance and attitudinal responses,” *International Journal of Speech Technology*, vol. 6, no. 2, pp. 123–131, 2003.