Unit-Selection Speech Synthesis Adjustments for Audiobook-Based Voices

Jakub Vít(✉) and Jindřich Matoušek

Department of Cybernetics, University of West Bohemia in Pilsen,
Pilsen, Czech Republic
{jvit,jmatouse}@kky.zcu.cz

Abstract. This paper presents easy-to-use modifications to unit-selection speech-synthesis algorithm with voices built from audiobooks. Audiobooks are a very good source of large and high quality audio data for speech synthesis; however, they usually do not meet basic requirements for standard unit-selection synthesis: "neutral" speech properties with no expressive or spontaneous expressions, stable prosodic patterns, careful pronunciation, and consistent voice style during recording. However, if these conditions are taken into consideration, few modifications can be made to adjust the general unit-selection algorithm to make it more robust for synthesis from such audiobook data. Listening test shows that these adjustments increased perceived speech quality and acceptability against a baseline TTS system. Modifications presented here can also allow to exploit audio data variability to control pitch and tempo of synthesized speech.

Keywords: Speech synthesis · Audiobooks · Unit selection · Target cost modification

1 Introduction

Unit selection ranks among the most popular techniques for generating synthetic speech. It is widely used and it is known for its ability to produce high-quality speech. The unit-selection algorithm is based on a concatenation of units from a speech database. Each unit is represented by a set of features describing its prosodic, phonetic, and acoustic parameters. Target cost determines a distance of each unit candidate to its target unit using features such as various positional parameters, phonetic contexts, phrase type, etc. When the algorithm is searching for an optimal sequence of unit, it minimizes total cumulative cost which is composed of the target cost and join cost. Join cost measures the quality of adjacent unit concatenation using prosodic and acoustic features like $F0$, energy, duration and spectral parameters. More detailed explanation of this method can be found in [1].

The work has been supported by the grant of the University of West Bohemia, project No. SGS-2016-039, and by the Technology Agency of the Czech Republic, project No. TA01011264.

P. Sojka et al. (Eds.): TSD 2016, LNAI 9924, pp. 335–342, 2016.
DOI: 10.1007/978-3-319-45510-5_38

The speech database is usually recorded by professional speakers in a sound chamber. Sentences for the recording are selected to cover as many unit combinations in various prosodic contexts as possible. When recording the speech corpus, the speaker is instructed to keep a consistent speech rate, pitch and prosody style during the entire recording. This method produces a very high quality synthetic voice but it is very expensive and time consuming as the number of sentences to record is very high (usually more than ten thousands—approximately 15 h of speech).

Audiobooks offer an alternative data source for building synthetic voices. They are also recorded by professional speakers and they have good sound quality. Unfortunately, they do not meet basic requirements for standard unit-selection synthesis: "neutral"[1] speech properties with no expressive or spontaneous expressions, stable prosodic patterns, careful pronunciation, and consistent voice style during recording. This problem is not so significant for the HMM; however, it greatly reduces quality of unit-selection based synthetic speech.

Unlike [2–5] and [6] where various styles were exploited to build an HMM synthesizer with the capability of generating expressive speech, our primary goal is to build only neutral voice but with highest quality possible. Therefore, unit-selection algorithm was used to ensure naturalness of synthetic speech.

This paper presents adjustment to unit selection algorithm to better cope with non-neutral and inaccurate speech database. It introduces a statistical analysis step to synthesis algorithm which allows to penalize units which would drop quality of speech. This step also allows to partially modify speech prosody parameters. It also summarizes the process of creating synthetic voice from audiobook for unit selection speech synthesis.

2 Audio Corpus Annotation

Unit selection voice requires a speech corpus which is in fact a database of audio files containing sentences and text transcriptions [7]. These sentences have to be aligned on a unit level, i.e., usually on a phone level. Text representation is often available for audiobooks but only in a form of a formatted text, not of a unit-level alignment. Also, the text form is usually optimized for reading, not for the computer analysis, therefore there must be some text preprocessing which removes formatting and converts text to its plain form. It is also necessary to perform text normalization, replace abbreviations, numbers, dates, symbols and delete all texts, which do not correspond to the audio content of a book. Due to the large volume of data, this step is no longer possible to perform by hand; therefore, it must be done automatically or at least semi-automatically.

The normalized text is then ready to be aligned to phone levels. However, segmentation and alignment of large audio files is not a trivial task [8]. Standard forced-alignment techniques which are used for the alignment of single sentences cannot be used here primarily because of memory and complexity requirements.

[1] In this paper, neutral speech is meant as news broadcasting style, which is very often used by modern commercial TTS system.

Text must be either cut into smaller chunks with some heuristics or annotated with the help of a speech recognizer run in forced-alignment mode. This approach tends to produce much more annotation errors when compared to the alignment of single sentences.

This problem was already dealt with in [8,9] or [10] where new techniques to reduce the number of errors were proposed. However, it must be noted that these errors can still occur and that the corpus database could contain badly aligned or otherwise unsuitable units.

2.1 Automatic Annotation Filtering

For our experiment, a simple procedure was used to check whether text annotation and alignment matches audio data. This procedure helped to remove the worst annotated sentences, i.e. sentences where the speech recognizer was desynchronized or text did not match audio representation. For every sentence from the source text, a sentence with the same text was synthesized using an existing high-quality voice, which was selected to be similar to the voice of the audiobook speaker. These sentences were compared using dynamic time warping (DTW) using mel-frequency cepstrum coefficients (MFCC) and euclidean distance. Distance was then divided by number of phones in a sentence to ensure the final score to be independent on the sentence length. Ten percent of sentences with the worst score were then removed from the speech corpus. Manual inspection confirmed that textual transcription of these sentences did not match the corresponding audio signal. A more sophisticated algorithm for detecting wrong annotation was proposed e.g. in [11].

3 Unit Selection Modification

The following subsections describe various modifications that were made to the described baseline algorithm to achieve better synthesis quality when using voices built from audiobooks.

3.1 Weights Adjustments

The total cost is composed of a large number of features which are precisely tuned to select the best possible sequence of units given the input specification. Source data from audiobooks have different characteristics. To reflect that, features' weights must be adjusted. Due to the higher prosodic variability of audiobook speech data, it is suitable to increase the weight of prosodic features (intonation, tempo, etc.) to keep speech as neutral as possible. Also, having a big database of audio data, it is possible to incorporate more specific features like more distant phonetic contexts or stricter rules for comparing positional parameters.

However, some of the features tend to be problematic in audiobooks, for example, phrasing. Narrators usually do not follow phrasing rules typical for read speech. They adjust their phrasing style based on the current context and actual

sentence meaning. They simply do not use the same prosodic sentence pattern. So, relying on positional parameters is not always useful and its contribution to the cost function should be reduced. It is better to focus more on "neutral" prosody to ensure the requested "neutral" speech.

Audiobooks also contain a lot of sentences with direct speech which are usually pronounced with different (more expressive) style. Moreover, this change of style can also affect neighboring sentences. Such problematic sentences can be either completely removed or penalized with another component of the target cost. If this component is tuned well, it could preserve those direct speech segments which do not have a different style than another parts of the book.

3.2 Target Cost Modification

A typical voice database created for speech synthesis contains precisely annotated units. All of them could be used during synthesis. Audiobooks contain much more "unwanted" data. Some units just do not fit into neutral style because of their dynamic prosodic parameters and some units might be wrongly annotated (see Sect. 2).

During the synthesis, each unit is assigned a set of possible candidates. Target cost is computed for each of the candidates. This cost is composed of many features which evaluate how well a candidate fits this unit. At this point, the algorithm is modified with an another step, in which all candidates are analyzed together. More concretely, their prosodic and spectral features (which are typically used also in join cost) are analyzed. For each individual feature ($F0$, energy, duration, MFCCs) a statistical model is built. The model is described by its mean and variance so that "expected" values for each feature can be predicted.

The target cost of every candidate is then modified with a value representing how much its features differ from the its statistical model. For each candidate, the *modified target cost* T_{modif} is then computed as the sum of original target costs T and the sum of features' *diversity penalty* $|d_i - 0.5|$

$$T_{modif} = T + \sum_{i=1}^{n_f} w_i \cdot |d_i - 0.5| \tag{1}$$

$$d_i = \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{f_i - \mu_i}{\sigma_i\sqrt{2}}\right)\right] \tag{2}$$

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2}\,\mathrm{d}t \tag{3}$$

where f_i is the i-th feature value of the candidate, μ_i and σ_i are mean and standard deviation of i-th feature across all candidates, d_i is a value of cumulative distribution function of the i-th feature in its statistical model. The value $d_i = 0.5$ means that $f_i = \mu_i$, w_i is a weight for the i-th feature and n_f is the number of features. Function $\operatorname{erf}(x)$ stands for an error function. Scheme of the target cost modification is shown in Fig. 1.

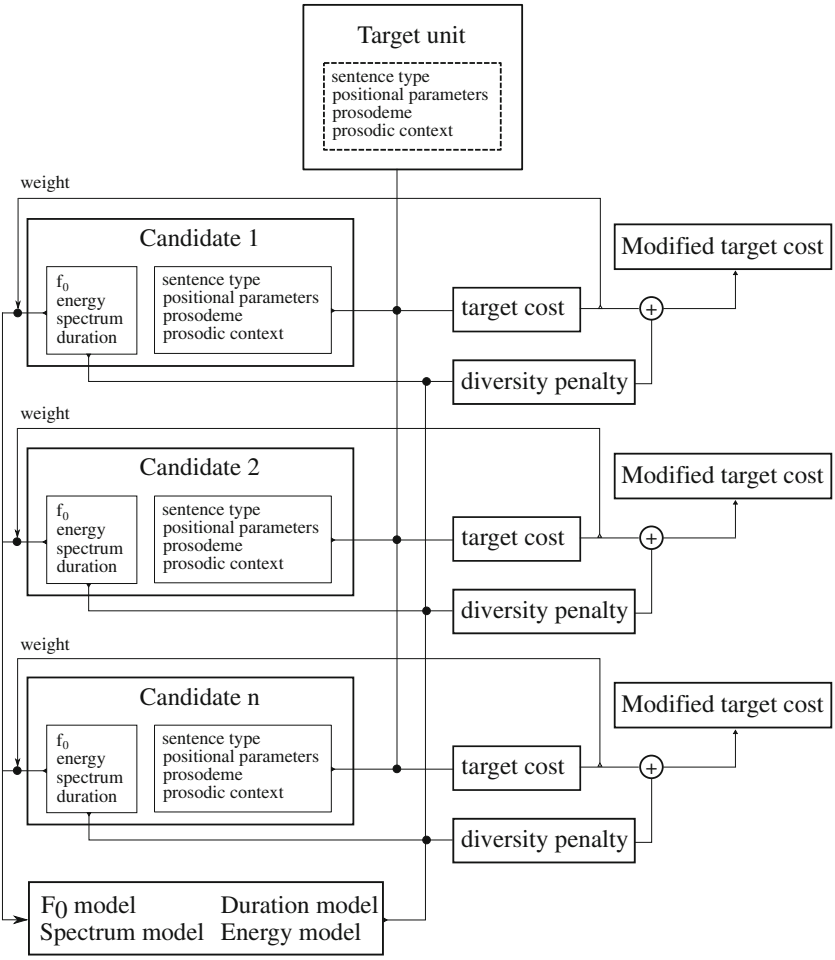

Fig. 1. Scheme of target cost modification.

Since prosodic feature values of candidates change a lot with different phonetic contexts, it would be unwise to build a model of the unit from all candidates with the same weight. Therefore, a weighted mean and variance were used with the weight being the inverted value of the target cost. If the candidate has a low target cost (meaning it fits unit nicely) its significance to mean and variance calculation is high. The weighted formula for mean value calculation is:

$$\mu_i = \frac{1}{c} \cdot \sum_{k=1}^{n_c} \frac{1}{1+T_k} \cdot f_{k,i} \tag{4}$$

$$c = \sum_{k=1}^{n_c} \frac{1}{1+T_k} \tag{5}$$

where T_k is the original target cost, n_c is the number of candidates, $f_{k,i}$ is the i-th feature of the k-th candidate.

This is also the reason why candidates cannot be preprocessed offline. Phonetic context of units is different every time and therefore weights are different resulting in different model parameters for each sentence.

By selecting candidates which feature values are close to the values predicted by the corresponding model, the outliers (i.e., candidates with a different voice style, with an unusual pronunciation or with wrong segmentation) are effectively filtered out. If a unit candidate has a bad annotation, some of its features (e.g. duration) would very probably differ from its expected values. Prosodic feature values will also differ if expressive voice style was used.

Being statistical based, this approach works only if there is a lot of data in a speech corpus. Otherwise, the model is not reliable enough.

3.3 Prosody Modification

The modification presented in Sect. 3.2 is used primarily to filter out outliers. Candidates whose feature values differ significantly from its statistical model are penalized. In Formula 1, 0.5 is used as an ideal reference value. This value is a logical choice but other values can be used to modify (prosodic) properties of output speech. For instance, if the duration reference is set to 0.6 (60 % quantile), the unit-selection algorithm will tend to select longer units, and the resulting speech will be slower in average.

Let us note that there is no need for absolute values of these features. The prosodic characteristics of the output speech (pitch, duration, energy) can be controlled using relative probability distribution function values on the interval <0,1>.

The amount of this modification can be controlled by tuning the weight ratio of this penalty against other components in the target cost. The described approach works even in the standard unit-selection framework. However, as a standard speech corpus does not contain so much variable data, the modification will be not so powerful.

Prosody modification worked very well in our experiments. Even very low weight w_i in Formula (1) was enough for prosody modification to work, especially for pitch and tempo. These parameters were changing on very large scale from very slow to very fast speech (or very low to very high pitch). On the other hand, energy modification was not so useful.

4 Evaluation

A three-point Comparison Category Rating (CCR) listening test was carried out to verify whether the modifications presented in this paper improved the quality of speech synthesized using audiobook-based voices. Ten listeners participated in the test. Half of them had experience with speech synthesis. Each participant got 70 pairs of synthesized sentences. In each pair, one sentence was synthesized by the baseline TTS system and the other one was synthesized by the modified TTS system. The set of possible answers was following: A is much better, A is

slightly better, A and B are the same, B is slightly better and B is much better than A. Order of A and B was shuffled. Listeners were instructed to choose a sentence with better general quality and naturalness.

Listening test results are shown in Figs. 2 and 3. The results show that proposed modification helped to improve the quality of synthesized speech. Nearly 70 % of all answers preferred the new system with the proposed modifications. Average answer was 0.85 on interval $<-2.0, 2.0>$. The results emphasize the importance of the modifications proposed in the paper when such variable speech data as the one from audiobooks are used. On other hand, the baseline unit-selection system does not take the speech data variability into account; the synthetic speech then suffers from more frequent occurrence of audible artifacts and inconsistent prosodic characteristics.

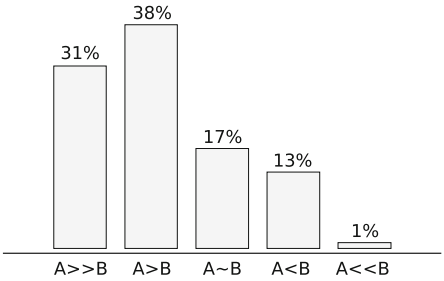

Fig. 2. Listening tests: distribution of answers. (A is the proposed system, B is the baseline system).

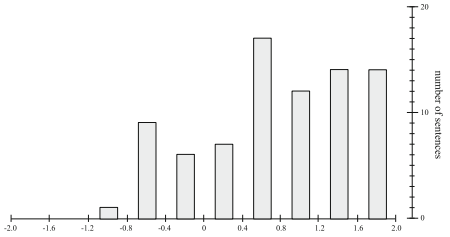

Fig. 3. Histogram of sentence average answers.

5 Conclusion

In this paper, we have described a modification of unit-selection algorithm for audiobook-based voices. As audiobooks do not always meet strict requirements for voices used by a standard unit-selection algorithm, our primary goal was to maximize quality and naturalness of speech synthesized from such variable speech sources. We presented modifications of the baseline unit-selection scheme which allow to lower these requirements by introducing a special step into the unit-selection procedure. In this step, candidates' prosodic features are analyzed and statistical models are built to describe expected feature values. Then, each candidate is penalized based on how much its feature values are different from its corresponding model. This approach leads to a penalization of units with different properties such as the ones frequent in audiobooks (e.g., non-neutral, expressive, and spontaneous speech properties, dynamic prosodic patterns, etc.). The proposed modification can also handle errors resulted from a mismatch between text and audio representation of an audiobook (i.e. annotation errors). Since this modification affects only target cost, which is far less computationally expensive than join cost, the computational complexity of the proposed algorithm is not significantly affected.

We also showed a way of taking advantage of this modification. The proposed algorithm enables to control output speech properties like pitch or tempo. According to a listening test, people preferred the proposed system much more than the baseline system. The test proved that this modification was beneficial to speech quality.

In future work, more detailed analysis could be done to identify benefits of individual adjustments to overall quality of speech. Also, expressive part of audiobooks could be used to build synthesis with expressive capability as proposed for instance by Zhao et al. [4]. Lastly, as more audiobooks narrated by the same speaker offer a possibility to use more data at the expense of introducing more inconsistencies, we plan to investigate how a mixture of recordings from more audiobooks will affect the quality of the resulting synthetic speech.

References

1. Dutoit, T.: Corpus-based speech synthesis. In: Benesty, J., Sondhi, M., Huang, Y. (eds.) Springer Handbook of Speech Processing, pp. 437–455. Springer, Dordrecht (2008)
2. Charfuelan, M., Steiner, I.: Expressive speech synthesis in MARY TTS using audiobook data and EmotionML. In: Proceedings of INTERSPEECH (2013)
3. Eyben, F., Buchholz, S., Braunschweiler, N., Latorre, J., Wan, V., Gales, M., Knill, K.: Unsupervised clustering of emotion and voice styles for expressive TTS. In: ICASSP, pp. 4009–4012 (2012)
4. Zhao, Y., Peng, D., Wang, L., Chu, M., Chen, Y., Yu, P., Guo, J.: Constructing stylistic synthesis databases from audio books. In: INTERSPEECH, Pittsburgh, PA, USA (2006)
5. Székely, E., Cabral, J.P., Cahill, P., Carson-Berndsen, J.: Clustering expressive speech styles in audiobooks using glottal source parameters. In: INTERSPEECH, pp. 2409–2412 (2011)
6. Székely, E., Cabral, J.P., Abou-Zleikha, M., Cahill, P., Carson-Berndsen, J.: Evaluating expressive speech synthesis from audiobook corpora for conversational phrases. In: Proceedings of LREC 2012 (2012)
7. Matoušek, J., Tihelka, D., Romportl, J.: Building of a speech corpus optimised for unit selection TTS synthesis. In: Proceedings of LREC 2008 (2008)
8. Prahallad, K., Toth, A.R., Black, A.W.: Automatic building of synthetic voices from large multi-paragraph speech databases. In: INTERSPEECH, pp. 2901–2904 (2007)
9. Braunschweiler, N., Buchholz, S.: Automatic sentence selection from speech corpora including diverse speech for improved HMM-TTS synthesis quality. In: INTERSPEECH, pp. 1821–1824 (2011)
10. Prahallad, K., Black, A.W.: Handling large audio files in audio books for building synthetic voices. In: The Seventh ISCA Tutorial and Research Workshop on Speech Synthesis, pp. 148–153, Japan, Kyoto (2010)
11. Matoušek, J., Tihelka, D.: Annotation errors detection in TTS corpora. In: Proceedings of INTERSPEECH, pp. 1511–1515, Lyon, France (2013)