Defining a Global Adaptive Duration Target Cost for Unit Selection Speech Synthesis

David Guennec(✉), Jonathan Chevelu, and Damien Lolive

IRISA - University of Rennes 1, Lannion, France
{david.guennec,jonathan.chevelu,damien.lolive}@irisa.fr
https://www-expression.irisa.fr/

Abstract. Unit selection speech synthesis systems generally rely on target and concatenation costs for selecting a best unit sequence. These costs, though often considering contextual features, mainly include local distances that are accumulated afterwards. In this paper, we describe a new duration target cost that takes a whole sequence into account. It aims at selecting a sequence globally good, instead of a very good sequence almost everywhere but having a few local duration cost leaps that are counter-balanced by other units. The problem of weighting this new duration cost with other sub-costs is also investigated. Experiments showed this new measure performed well on sentences featuring duration artefacts, while not deteriorating others.

Keywords: Target cost · Cost function · Neural networks · Corpus-based TTS · Unit selection

1 Introduction

While new Statistical Parametric Speech Synthesis based TTS techniques are currently emerging, like DNN-based TTS, unit selection and HSMM-based synthesis remain the two most influential methods investigated so far, along with hybrid techniques that try to get the best from both worlds. HMM-based parametric approaches, for which HTS [1] is the main system, are quite recent and have been the framework for many academic work in recent years. These methods offer advanced control on the signal and produces extremely intelligible speech but generated voice lacks naturalness. The historical approach, unit selection, is a refinement of concatenative synthesis [2–7].

In the formulation of the unit selection problem, a unit is a list of contiguous segments (from a speech corpus) fitting a portion of the target sequence of phonemes to synthesize. To discriminate units that fit requirements expressed via the target sequence, the usual method [3] is to search a unit graph with a search algorithm, evaluating the context matching degree (target cost) and the risk of creating an artefact if concatenating the unit (concatenation cost) via balanced cost functions. Alternative ways exist though. For instance, one can also achieve unit selection with genetic algorithms and selection and crossover operators are used along with fitness measures [8].

P. Král and V. Matoušek (Eds.): TSD 2015, LNAI 9302, pp. 149–157, 2015.
DOI: 10.1007/978-3-319-24033-6_17

Speech created using unit selection features naturalness and prosodic quality unmatched by other methods, as it basically concatenate speech actually produced by a human being. For this reason, most industrial TTS systems mainly use either pure unit selection approaches or hybrid ones. However, unit selection offers less control than statistical parametric methods, especially over prosody. Moreover, artefacts may appear in the synthesized signal and penalize intelligibility. While obtaining good speech output with neutral voice is (almost) a solved problem with unit selection, getting prosody right for natural and expressiveness is entirely another matter. Prosody modification methods after selection - like TD-PSOLA for adapting duration - are an option, but for now none has been convincing. The possibility of influencing selection to choose units that are the closest to the required prosody remains. A good state of the art for expressive speech synthesis is made in [9]. As phonetic durations are subject to a lot of changes when considering voices with different levels of expressiveness, controlling duration gets particularly important. Lastly, decision trees have been the most widely used method to predict duration, for instance, in systems like HTS, with only a few mentions to using a target duration cost (e.g. in [10]) within unit selection cost function. Recent approaches where DNNs replace HTS decision tree can also be mentioned [11].

In this article, we propose a new way of computing duration target cost, not only based on the assumption that we want to get units as close as possible to a predicted duration. Thus, we try to find the units that stay the closest to requested duration by optimizing the mean duration error with respect to the previous units. Hence, it prevents inadequate units in terms of duration from being selected if other units are available while not forcing a path with homogeneous durations. The main idea is that it is better to have units globally longer or shorter than to have only one or two units with a big duration error in the synthesized speech. The paper is organized as follows. The TTS system used as a basis for experimentation is presented in section 2. Proposed target cost, along with the underlying duration model are presented in section 3. Experimental evaluation on french corpora including both objective assessments of the model and the target cost (4.2) and subjective evaluation by listeners (4.3) are presented in section 4. Conclusions and future work are presented in section 5.

2 The TTS System

In this work, we use the IRISA corpus-based TTS system[12] as a basis for our experiments. The cost function is built following the traditional equation [3]:

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

where U^* is the best unit sequence according to the cost function and u_n the candidate unit trying to match the n^{th} target unit in the candidate sequence U. In this work, the considered unit size is the diphoneme. $C_t(u_n)$ is the target cost and $C_c(u_{n-1}, u_n)$ is the concatenation cost. W_{tc}, W_{cc}, w_n and v_n are weights for adjusting magnitude for all the parameters. Sub-costs are weighted using corresponding mean cost values in the TTS corpus to compensate magnitudes of all sub-costs. Our concatenation cost is composed of amplitude, MFCC and F0 distances. Current target cost is composed of the new duration cost alone. The algorithm accesses the corpus via a set of preselection filters, preventing units that do not match them to be added to the graph. Their purpose is twofold. First, it considerably prunes the graph explored by the unit selection algorithm, making selection process faster. Second, it serves as a set of binary target cost functions relying on the assumption that if a unit doesn't respect the required set of features, it cannot be used for selection. The preselection filters should therefore be seen as part of the cost for a node. In our system, when no corpus unit respects a given set of preselection filters, the set is relaxed (removing features that seem the least helpfull one by one) until units are found. This mechanism ensures finding a path in all cases provided that the corpus has a full covering of diphonemes. The set of preselection filters we use in this work is the following:

1. Unit label (cannot be relaxed).
2. Is the unit a Non Speech Sound (cannot be relaxed)?
3. Is the phone in the last syllable of its sentence?
4. Is the phone in the last syllable of its breath group?
5. Is the current syllable in word end?

3 An Adaptive Duration Target Cost

3.1 Neural Network

Prediction of phoneme duration has a long history in the TTS field. It was first performed by creating expert hand-made rules that were integrated in rules-based (formant synthesis) and concatenation synthesizers. Over last years, decision trees have been the most widely used method to predict duration, for instance, in systems like HTS. In particular, the use of neural networks for phoneme duration prediction starts in the early '90s. A TTS system using a set of ANNs (one for each phoneme) trained on cepstral coefficients can be cited [13]. A TDNN (Time Delay Neural Network) has also proven to be very efficient for predicting duration, though the learning set was small [14]. In following years, major improvements in the technique were obtained mainly by increasing the number of input features and the size of the learning corpus. The advantage of neural networks is that, contrary to decision trees, they do not cluster predicted values (at least when properly trained). When the network faces an unknown set of features, the predicted value is not the assimilated result for the closest feature set, which can result in much better results [15]. Recent work in speech

synthesis is now focusing on deep approaches (DNNs, DBNs, DRNs). For duration prediction, we did not think such deep approaches were necessary. Thus, we use a MLP (Multi-Layer Perceptron) with batch gradient descent. Input data is composed of a set of 50 features by phoneme, mainly phonetic and linguistic parameters. We also take into account the contextual information for the two preceding and following phonemes. Thus, the network has a topology of 250 input neurons, 1 rectified linear hidden layer of 512 neurons and one output linear Gaussian neuron (directly predicting durations in *ms* as other measures like *log ms* were not performing better). These parameters were the best among the different configurations tested.

3.2 Duration Target Cost

The proposed duration target cost aims at influencing selection so that selected units are, on average, at the same distance of the predicted unit durations. Defining the cost that way means we prefer a sequence moderately close to predicted values, but homogeneous in the repartition of the duration distance among units, to a sequence of perfect elements featuring one unit with dramatic cost. The cost for target unit n in the sequence U (see eq. 1) is as follows:

$$D_e = |D_t(u_n) - D(u_n)| \tag{2}$$

$$C_d(u_n) = |\Delta(u_{n-1}) - D_e| \tag{3}$$

$$\Delta(u_n) = \frac{\Delta(u_{n-1}) * (n-1) + D_e}{n} \tag{4}$$

with Δ_{u_n} being the mean distance to predicted duration for previous target units in the sequence (from u_1 to u_n), $D_t(u_n)$ the target duration for unit u_n, $D(u_n)$ the duration of u_n and $C_d(u_n)$ the target duration cost for unit u_n.

Equation (2) computes the local cost between the target duration and the current unit. This cost is then used to compute the duration target cost in equation (3), which takes into account the mean distance to predicted duration for all the previous units. Finally, the mean duration error is updated using equation (4). Thus, the quality of the current unit depends on the quality of previous units. In other words, it means that if u_n is longer (resp. shorter) that desired, the target cost will be low if the previous units are also longer (resp. shorter). This way, we want to keep the consistency between the different units which might be better than inconsistency and perhaps produce a credible speaking rate slow-down or speed-up.

4 Experiments

We have conducted experiments aiming at (i) testing the accuracy of our ANN, (ii) measuring the impact of the new target cost on the unit selection algorithm and (iii) subjectively assessing the improvement in produced speech.

4.1 Corpus Description

Two corpora were used, both as learning sets for ANNs and TTS voices. All are in french language. The first one, *Audiobook*, is extracted from an highly expressive audiobook. The speaker is a male and the mean F0 value for voiced segments is low, at only 87Hz in the corpus. Data was automatically annotated using the process described in [16]. *Audiobook learning* has 353,691 phonemes and 22,727 Non Speech Sounds and is 10 hours long. Its diphoneme covering is not full (78%) but all the most commonly used diphonemes are present. The second corpus, named *IVS*, was recorded for TTS purposes within an Interactive Vocal System with a hand-made recording script which aim was to cover all diphonemes present in French and comprises most used words in the telecommunications field. It features a neutral Female voice sampled at 16kHz (lossless encoding, 1 channel) with a mean F0 at 163Hz for voiced segments. The corpus is composed of 7,662 utterances, 239,260 phonemes and 20,424 Non Speech Sounds for 7h05' speech and is manually annotated. Both corpora are managed using the ROOTS toolkit [17].

200 utterances were removed from each corpus to create four 100 utterances corpora: *Audiobook test*, *Audiobook validation*, *IVS test* and *IVS validation*. Test corpora are used during the train process on ANNs to control training quality at each epoch. Validation sets were used to verify the efficiency of the model after training.

Finally, we used 100 french sentences (i.e. sometimes featuring more than 1 utterance) corpus extracted from a wide variety of audiobooks, featuring very different styles, many being far from *IVS* and *Audiobook*'s styles. It served as our TTS test corpus. All TTS generations in our experiments used these as target sentences. In the following, it will be referred to as *TTS test* corpus.

4.2 Objective Analysis

Neural Network. The mean RMS error for *IVS* voice is slightly better (RMS=24.24, std=9.07) than for *Audiobook* (RMS=26.58, std=6.61). Pearson scores show that predictions are strongly correlated to real values, and the probability of error on the Pearson score is extremely weak. A detailed analysis on a per phoneme basis shows that the worst phonemes are those having very few representations in the learning corpus, for each voice. For instance, /ɲ/ has only 2 realizations in the *Audiobook* corpus, and only one in *Audiobook validation*. Finally, when looking at real and predicted centroids for each phoneme, most of them are very close, if not identical. Given these results, which we consider as fair, and knowing we do not need extremely accurate predictions as they are solely used to influence selection, these models have been kept as is.

Behavior of the Cost Function. To evaluate the impact of duration cost and its interactions with concatenation costs, we considered all $\{W_{tc}, W_{cc}\}$ couples in the $[0, 100]$ interval with a pace of 10. For each weight configuration, we generated the 100 sentences in our *TTS test* corpus. Sentence (not utterance)

based measures were extracted for each configuration. In this section, we will only discuss these measures on *IVS* voice, but exactly the same patterns are observed on *Audiobook* voice. Only small variation in magnitudes are observed between the two voices. It is important to point out that costs presented here are obtained *without* applying W_{tc} and W_{cc} weights. Magnitudes due to these weights have been removed to get raw costs.

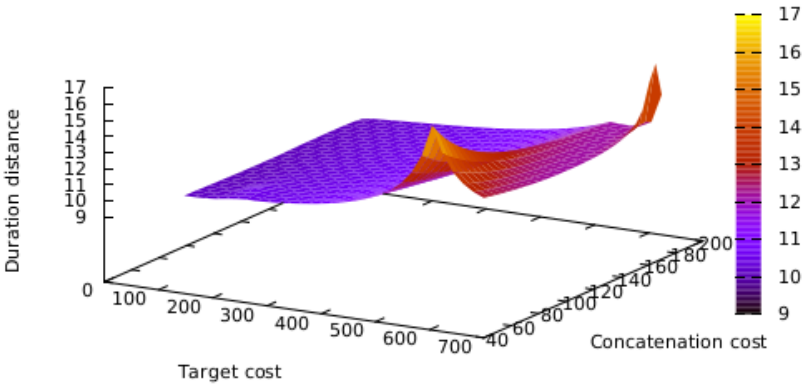

Fig. 1. Duration delta between model predictions and synthesized durations evolution when target and concatenation costs vary. Distance, per phoneme, is given in ms. Data computed using synthesis from *TTS test* corpus.

Figure (1) shows the evolution of the mean delta per phoneme in *ms* between predictions by the network and final produced durations in relation to target and concatenation costs magnitudes. As it can be seen, the general trend is that distance increases when the target cost increases, which shows a good functioning of our target cost. Moreover, when getting the worst target cost, the delta largely increases. An unexpected result is the relation between the delta and concatenation cost when target cost is high which seems to suggest that concatenation cost excludes units with worst duration, improving the delta. When concatenation cost increases again, the delta dramatically increases again too. We can further note that duration delta at high target costs and low concatenation costs, while being good, remains much higher than the delta we get at lower target costs (this time independently of concatenation cost).

This result led us to think it would be worth investigating the behavior of a system where the duration target cost would be activated only on certain conditions, like for high concatenation cost or when confronted to a drastic relaxation of preselection filters.

4.3 Subjective Evaluation

Based on precedent measures, we selected the configuration $\{W_{tc} = 30, W_{cc} = 70\}$ for listening tests. This choice was motivated by the low variability in terms of duration costs when getting over $W_{tc} = 30$ and the fact that concatenation cost alteration at this level is low. The same reasoning led us to $W_{cc} = 70$. In consequence, listening tests were performed using two system configurations: the baseline system, with weight configuration $\{W_{tc} = 0, W_{cc} = 100\}$ which we call *Uncontrolled*; and the configuration incorporating our duration distance, $\{W_{tc} = 30, W_{cc} = 70\}$, called *Controlled*.

We performed two AB tests involving 13 testers for the first and 11 for the second (half of which were experts) on the *Uncontrolled* and *Controlled* systems. Tests follow recommendations in [18]. Three answers were proposed to the listeners: A, B and "Indifferent". Both *Audiobook* and *IVS* voices were mixed in each test.

The first test presented 20 stimuli for each voice, taken randomly in the TTS test set. The testers were asked to assess the rhythm of speech and select the best system. On raw results, systems were getting almost as much votes (43% for *Uncontrolled* and 38% for *Controlled* with overlapping intervals). We spotted extremely different scales of notation among testers, with none seeming to have the same way of performing the test. Thus, no hard conclusion can be derived from this test. Nonetheless, it suggests the two systems are on par. It is important to underline that post-analysis of the stimuli presented for this test showed that very few samples had strong duration incoherences.

An important point is that *IVS* corpus featuring only neutral voice, duration artefacts are less serious and less frequent. On the contrary, *Audiobook*, being very expressive, features much more minor duration issues. Major duration problems are also much more frequent.

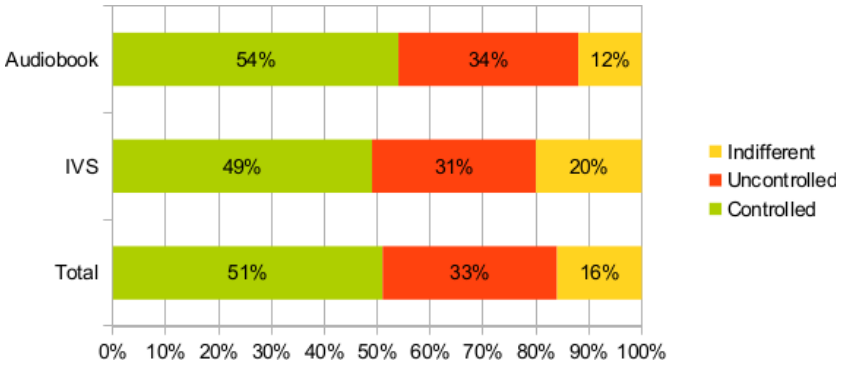

Fig. 2. AB test results. *Uncontrolled* featuring duration artefacts is opposed to *Controlled* system. First and second row are a decomposition of the third one. *Controlled* is clearly preferred.

The second test was focused on sentences having audible duration artefacts. 22 different sentences featuring duration artefacts (of various amplitudes but all being audible) were extracted from *Uncontrolled* synthesis (11 for each voice). They were confronted to their equivalent with *Controlled* system. The testers were asked to say which system has the most natural voice. The testers were also asked to pay particular attention to rhythm (but not exclusively).

Results for this second test are presented on figure 2. First row shows results for *Audiobook* voice only, second for *IVS* only while the third one is the global result. In this test, *Controlled* is strongly preferred by testers, especially for *Audiobook* voice which is normal as it is the voice the most likely to generate artefacts. It was also interesting to see that testers all followed the same trend, placing *Controlled* ahead with different levels of preference. Experts especially had a strong preference for *Controlled* when using expressive voice *Audiobook*, and less for *IVS*.

Given these results, it can be derived that our target costs behaves well in enhancing durations when needed and only when needed, while not deteriorating synthesis on other aspects.

5 Conclusion

In this paper, we presented a new duration target cost for unit selection. This cost aims at selecting the whole unit sequence that best minimizes duration distance with predicted values rather than choosing the sequence containing units that individually minimize a duration distance. This is intended to avoid cases like excellent synthesis penalized by few very bad units. Experiments showed that this new measure performs well on speech samples that feature durations issues, especially on expressive voices. Furthermore, the new measure does not seem to affect synthesized samples that have good durations from the beginning. An extension of this work we are investigating is to test activating the duration cost only on some sub-parts of the target sequence, under particular conditions suggesting the target cost is needed like a strong relaxation of preselection filters or high concatenation cost. A distinct pause duration model, which could use the same specifications as presented in this paper should also be added. Implementing an intonation target cost relying on a F0 contour prediction model is also part of our next work.

References

1. Yamagishi, J., Ling, Z., King, S.: Robustness of HMM-based speech synthesis. In: Ninth Annual Conference of the International Speech Communication Association, pp. 2–5 (2008)
2. Sagisaka, Y.: Speech synthesis by rule using an optimal selection of non-uniform synthesis units. In: Proc. of ICASSP, pp. 679–682. IEEE (1988)
3. Black, A., Taylor, P.: Chatr: a generic speech synthesis system. In: Proc. of Coling, Association for Computational Linguistics (1994)

4. Hunt, A., Black, A.: Unit selection in a concatenative speech synthesis system using a large speech database. In: Proc. of ICASSP, pp. 373–376. IEEE (1996)
5. Taylor, P., Black, A., Caley, R.: The architecture of the festival speech synthesis system. In: Proc. of the ESCA Workshop in Speech Synthesis, pp. 147–151 (1998)
6. Breen, A., Jackson, P.: Non-uniform unit selection and the similarity metric within bts laureate tts system. In: Proc. of the ESCA Workshop on Speech Synthesis, pp. 373–376. Citeseer (1998)
7. Clark, R., Richmond, K., King, S.: Multisyn: Open-domain unit selection for the festival speech synthesis system. Speech Communication, 317–330 (2007)
8. Kumar, R.: A genetic algorithm for unit selection based speech synthesis. In: Eighth International Conference on Spoken Language Processing (2004)
9. Schröder, M.: Expressive Speech Synthesis: Past, Present, and Possible Futures. In: Affective Information Processing, pp. 111–126. Springer, London (2009)
10. Alías, F., Formiga, L., Llorá, X.: Efficient and reliable perceptual weight tuning for unit-selection text-to-speech synthesis based on active interactive genetic algorithms: A proof-of-concept. Speech Communication, 786–800 (May 2011)
11. Hashimoto, K., Oura, K., Nankaku, Y., Tokuda, K.: The effect of neural networks in statistical parametric speech synthesis. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 4455–4459 (2015)
12. Guennec, D., Lolive, D.: Unit selection cost function exploration using an A* based text-to-speech system. In: Sojka, P., Horák, A., Kopeček, I., Pala, K. (eds.) TSD 2014. LNCS, vol. 8655, pp. 432–440. Springer, Heidelberg (2014)
13. Tuerk, C., Robinson, T.: Speech synthesis using artificial neural networks trained on cepstral coefficients. In: Proc. of EUROSPEECH, pp. 4–7 (1993)
14. Karaali, O., Corrigan, G., Gerson, I.: Speech synthesis with neural networks. In: Proc. of World Congress on Neural Networks, pp. 45–50 (1996)
15. Taylor, P.: The target cost formulation in unit selection speech synthesis. In: Proc. of Stress, pp. 2038–2041 (2006)
16. Boeffard, O., Charonnat, L., Le Maguer, S., Lolive, D., Vidal, G.: Towards fully automatic annotation of audio books for tts. In: Proc. of LREC, pp. 975–980 (2012)
17. Chevelu, J., Lecorvé, G., Lolive, D.: Roots: a toolkit for easy, fast and consistent processing of large sequential annotated data collections. In: Proc. of LREC, pp. 619–626 (2014)
18. ITU-T: Itu-t recommendation p. 800: Methods for subjective determination of transmission quality (1996)