Multi-level Exemplar-Based Duration Generation for Expressive Speech Synthesis

Mohamed Abou-Zleikha, Éva Székely, Peter Cahill, Julie Carson-Berndsen

CNGL, School of Computer Science and Informatics, University College Dublin, Dublin, Ireland
{mohamed.abou-zleikha|eva.szekely}@ucdconnect.ie, {peter.cahill|julie.berndsen}@ucd.ie

Abstract

The generation of duration of speech units from linguistic information, as one component of a prosody model, is considered to be a requirement for natural sounding speech synthesis. This paper investigates the use of a multi-level exemplar-based model for duration generation for the purposes of expressive speech synthesis. The multi-level exemplar-based model has been proposed in the literature as a cognitive model for the production of duration. The implementation of this model for duration generation for speech synthesis is not straightforward and requires a set of modifications to the model and that the linguistically related units and the context of the target units should be taken into consideration. The work presented in this paper implements this model and presents a solution to these issues through the use of prosodic-syntactic correlated data, full context information of the input example and corpus exemplars.

Index Terms: speech prosody, duration generation, exemplar-based model

1. Introduction

A model of duration, as part of a prosody model, plays a demonstrable role in the intelligibility and naturalness of synthesised speech. Several studies on speech synthesis have been undertaken which seek to generate the duration information from text where the size of generating unit, the duration representation and the generation algorithm have all been considered. Some of these studies have concluded that the syllable is the most suitable unit for modelling and generation since it is the natural unit in speech, while others used the phone since it is the smallest unit in speech. Some approaches represent duration in terms of absolute unit value, others represent it as z-score values and some represent it as a sequence of states. The sequence of states is suitable for representing the duration in Hidden Markov Model (HMM) and other statistical models where the synthesis unit is the state. Absolute values are more suitable for phone-based approaches and the z-score is most suitable when a higher-level unit than the phone is used or when the units are represented in a different way, in terms of consonant-vowel structure, for example. In this paper, the application of an exemplar-based duration model is investigated which aims to overcome some of the limitations of approaches based on average unit durations.

Several algorithms have been suggested for the generation of duration. Klatt used the phone duration as unit size in one of the first rule-based approaches to duration modelling and generation based on linguistic information [1]. The Campbell duration model uses the z-scores for syllable duration and a neural network to predict it. [2]. Other machine learning techniques have also been investigated, which include the use of the phone and genetic algorithms [3], neural networks for phone prediction [4] and Bayesian belief networks to predict consonant duration[5]. The main problem with these methods is that they use average models to generate the duration information. While this approach captures the general acoustic and prosodic properties in a speech corpus, it is less effective in reflecting fine-grained speaker-specific detail that is an essential element of speaker identity. An exemplar-based model has the potential to more accurately reflect the, often context-dependent, speaker-specific prosodic characteristics and thus better preserve the speaker's identity in the synthetic speech. Another advantage of an exemplar based model is one concerning expressive speech. Duration properties are an important component of expressiveness in speech corpora. If these properties are not averaged over in the duration prediction process, but rather kept intact within a linguistic context, the resulting synthetic speech is likely to produce a closer reflection of the expressive speech characteristics in the corpus.

A multi-level exemplar-based model has been proposed by [6] which incorporated a duration production model. The model shows promising result for the production task, but in order to apply this model to duration generation from a speech synthesis perspective, a set of questions need to be answered. More specifically, it is necessary to find a mechanism to identify a set of potential candidates for a new input, to choose an input duration from set of potential candidates from the exemplar database and to apply a perceptual test on the synthesised speech.

This paper investigates the use of a multi-level exemplar-based approach to generate phone duration for speech synthesis. Based on this approach, the proposed method uses two levels of units, syllables and phones, and aims to find the potential candidates according to the syllable units. If the activation (e.g. the number of similar exemplars that are found in the corpus) of the unit exceeds a threshold, a selection process according to the context of the input units is performed and the duration of the units' phones is extracted. If the unit activation is less than a threshold, the phone units are used where the duration of each phone is extracted using its context. Lastly an update process for the exemplar memories is performed with the generated durations and the units. The duration model has been implemented in Hidden Markov Model-based speech synthesis (HTS)[11].

2. Multi-level exemplar-based model

In multi-level exemplar theory [6], the proposed model contains two databases:

- an exemplar database on the unit level: this database contains the unit exemplars (e.g syllables database).
- an exemplar database on the constituent level: this

database contains the constituent exemplars (e.g phones database).

The model consists of four components:

- *Generation/Perception Interface*: this interface transmits an input that serves as stimulus for the model.
- *Parser*: the parser parses a unit into its constituents (e.g a syllable into its phones).
- *Similarity Calculator*: this module calculates the similarity between the input and the exemplars in the exemplars database, and returns the most similar units to the input. It also returns the level of activation.
- *Decision Component*: this component receives the activation that has been calculated by the similarity calculator. If the activation value is above a threshold, the perception or production will be based on the set of similar units found by the similarity calculator in the unit exemplar database. If the stimulus does not receive sufficient activation in the unit exemplar model, then perception/production is based on the sets of similar constituents (one set per constituent) found by the similarity calculator in the constituent exemplar database.
- *Composer*: the composer composes a sequence of constituents into a unit (compose a syllable duration from the duration of its phones).

For a new stimulus unit, if it receives enough activation, then its exemplars cloud (or exemplars database) is used for production or perception. Otherwise the alternative path is taken where exemplars similar to the individual constituents are used as the basis for the input to the composition function. Inspired by this model, the next section presents further explanations for each of previous components is discussed in the context of a specific workflow for modelling duration for speech synthesis.

3. Multi-level exemplar-based duration generation

Given an input text, the parser component parses this input into a sequence of syllables and the syllables into a sequence of phones. For each syllable and its phones, a set of candidates is selected from the unit and constituent exemplar databases, these candidates are selected from the prosodic and linguistically related data only; if the activation level of unit level exemplars (syllable) exceeds a threshold, the units candidates are used to compose the new duration model for this unit using the composer component, otherwise, the constituent level (phone) candidates are used. Figure 1 illustrates this process.

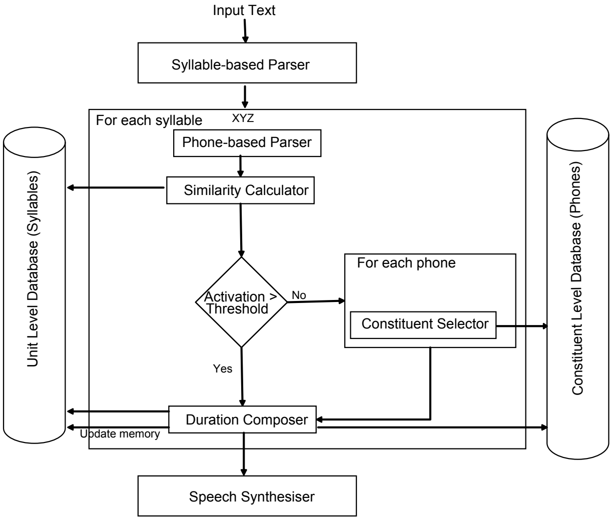

Figure 1: *Duration generation process.*

3.1. Unit-level and constituent-level databases

In the unit-level database, syllables with context and consonant-vowel structures are stored while in the constituent-level database the phones of each syllable is extracted along with their context and stored. The duration of each phone is represented using the z-score which represents the deviation of the value above or below the mean, this value is calculated as :

$$z - score(x) = \frac{x - \mu}{\sigma} \tag{1}$$

where μ is the mean value of the duration of the unit x and σ is the standard deviation. The reason for choosing z-score to represent the duration is to allow for the possibility of choosing a syllable candidate only on the basis of a consonant-vowel structure in which case, a normalised duration model will be needed.

3.2. Parser

The parser extracts the units and its constituents (syllables structures and phones), and extracts the associated context information. This context information is stored in the exemplars database, and used in the composer component.

3.3. Similarity calculator

The exemplars database contains several candidates for each unit and constituent. The purpose of the similarity calculator is to determine these possible candidates. The similarity function is defined as

$$VCS(input\_unit) = database\_syllables \tag{2}$$

where VCS is the function that converts the syllable to its consonant-vowel structure.

The relationship between the syntactic information and duration information was addressed previously where a mapping was defined as a probabilistic relationship [7] rather than a one-to-one relationship. The research results showed that the mapping is determined by calculating the probability of a sample utterance chosen from the set which represents the closest $\alpha\%$ to an element on the text level being also an element of the set which represents the closest $\beta\%$ to that element on the duration level (i.e. the conditional membership probability). By optimising the value of this probability, these $\alpha\%$ from the data are determined and used for the solution evaluation. To guarantee that only the prosodically relevant units will be selected, the units are selected from the closest utterances to the input according to tree edit distance between the syntactic tree of the input and syntactic trees of the database.

3.4. Decision component

The similarity calculator on the syllable level returns a set of candidates and the purpose of the decision stage is to decide whether the syllable level or phone level will be used. The decision depends on two factors: the activation level (which in this case represents the number of activated candidates), and a threshold. If the activation level exceeds the threshold, a unit level (syllable) is used; otherwise the constituent level (phones) is used.

3.5. Composer

The composer takes a set of candidates for each unit and finds the most suitable one according to the context of that unit. The purpose of the composer at the first stage is to find the units that minimise the formula:

$$\underset{i}{\operatorname{argmin}} f(X_i, Y) = \sum_{1}^{k} \gamma_k dis(X_{i,k}, Y_k) \tag{3}$$

where X_i is the candidate with the index i in the candidates set, $X_{i,k}$ is the context feature that has index k in the exemplar X_i, γ_k the weight of feature k, Y is the target unit, and Y_k is the feature that has index k in the target unit context. As a result of this formula, the candidates that minimise the distance between the context features are selected. If the unit is a syllable, the phones of this syllable are selected from the phones database, and the duration is calculated. If the unit is a phone, the duration of that phone is selected. In the case of taking the phone level, the syllable and phone exemplar memories are updated with the composed new syllable with its context and its phones.

4. Model implementation in HTS

4.1. Corpus design

The corpus used in the experiment is part of an open source audiobook originally published on librovox.org, read by John Greenman. The segmented audio was made available for the Blizzard Challenge 2012 by Toshiba Research Europe Ltd, Cambridge Research Laboratory. The method used to align the audio with the corresponding text and segment it into smaller utterances is described in [8]. Two Mark Twain books, *A Tramp Abroad* and *The man that corrupted Hadleyburg* were selected for this experiment. A corpus of approximately 5 hours was created from the utterances of the audiobooks that were no longer than 5 seconds long. A Self-Organizing Feature Map with input features calculated from glottal source parameters [9] per speech segment was used to identify the variety of expressive speaking styles in the corpus. This method is described in [10]. The 5-hour corpus was then divided into 3 subcorpora, featuring broad categories of three different voice styles. A brief perceptual characterisation of these subcorpora is as follows:

- *Subcorpus A*: Soft, lax voice, featuring relatively low pitch ranges
- *Subcorpus B*: Tense, louder voice, with all pitch ranges represented
- *Subcorpus C*: Very expressive, intense voice, with mid to high pitch ranges

4.2. Voice building and model integration

Three voices have been built from these subcorpora. Two duration models were used, the integrated duration model in HTS (HSMM), and the proposed exemplar based method (Exemplar-based). Both models used HTS version 2.2 [11]. The typical HTS f0 generation method (5 states multi-space probability distributions (MSD) HMM) and MFCC were used for both models. The multi-level exemplar based model was integrated by explicitly forcing the generation algorithm to use the phone duration predicted by that model [12]. Using this approach, the state-level duration model is used to extract the distribution of phone duration among the states. The main advantage of this integration approach is that it eliminates the effect of duration on modeling the other speech parameters (e.g. pitch, spectrum).

5. Evaluation

To validate this model, an AB comparison test between the exemplar-based model and a baseline Hidden semi Markov Model (HSMM) was undertaken for each voice separately.

5.1. Experiment design

Traditionally prosody model evaluation has focused on the aspect of naturalness. Evaluating prosody effects is considered a difficult task due to its perceptual and contextual factors. It is desirable to include aspects such as pleasantness and suitability to specific listening contexts [13]. This evaluation was designed to incorporate contextual effects and investigate fine-grained speaker specific detail. The following 4 questions were asked of the utterances to elicit this:

- Which sounds more like an actual person?
- Which one is more likely to be part of a conversation?
- Which one is more fluent?
- Which one reflects the speaker's intention better?

A set of 40 sentences were used for the evaluation, each of these utterances was synthesised using both the exemplar-based and HSMM duration models applied on each sub-corpus. Data was collected online from 17 participants who took part in the test. Each participant was presented with 30 pairs of utterances: out of the 40 sentences, each listener was given a random selection of 10 utterances for each voice using the two duration models, resulting in 60 speech samples to evaluate per person. For each question, listeners could select one of the following four options: 1) sample A preferred, 2) sample B preferred (one of which featured the exemplar-based duration model, the other HSMM, in randomized order), 3) both samples sounded the same regarding the aspect in question, 4) neither of the two samples measured up to an acceptable standard regarding the aspect in question. Figure 2 presents the results of the evaluation. The four possible answers are labelled ”Exemplar-based”, ”HSMM”, ”Same” and ”Neither”, respectively.

5.2. Results

The percentage of voting has been calculated for each question and for each voice. In order to check if the reported voting for the two models is significant (one of the models is statistically better than the other), a Friedman test was performed on the preference values and significance was measured with a $p{-}value < 0.05$. The results showed a significantly higher fluency for voices A and B using exemplar-based model compared with the HSMM, but a small preference for HSMM for voice C. The exemplar-based model achieves a significantly better preference for the part of conversation speech according to voices A and B and the same preference according to voice C. The exemplar-based model was preferred for the question regarding actual person for voices A and B, but HSMM is preferred for voice C. The results show that the exemplar-based model reflects the speaker intention more than the HSMM model in

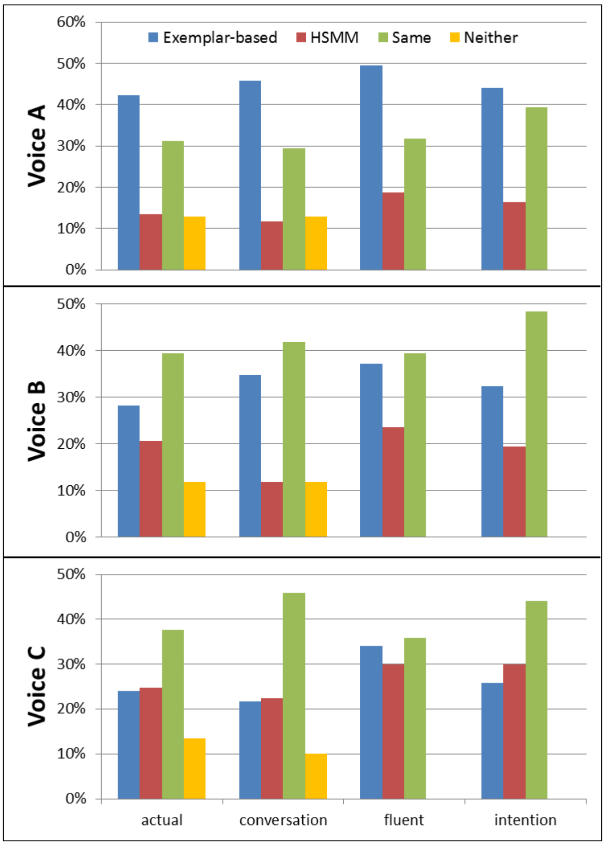

Figure 2: *Voting percentage for voices A, B and C.*

voices A and B, but HSMM reflects it more than the exemplar-based model in C. Figure 2 illustrate the results of voting.

To study the relationship between the questioned aspects, a Pearson product-moment correlation coefficient between each pair of questions for each voice was calculated.The statistical test showed significant correlation between all the pairs of questions.The results on voice A showed a strong correlation between the fluency aspect and actual person speech and speaker intention aspects. The actual person speech aspect shows also a strong correlation with the conversation speech and speaker intention aspects. In addition to the previous relations, the speaker intention aspect showed also a strong correlation with conversation speech aspect. Using voice B all the four aspects, the native speaker aspect is not strongly correlated with all other aspect; but the rest of the aspects are fully correlated. The results using voice C showed that the fluency is correlated with the intention aspect. In addition, the actual person aspect is correlated with the conversation speech aspect, and the intention aspect is correlated with the actual person aspect, the conversation speech and the fluency aspect.

6. Conclusions

This paper presents a multi-level exemplar-based model for duration generation in expressive speech synthesis. A perceptual evaluation was carried out with 17 listeners evaluating three expressive voices. The voices were evaluated with respect to sounding like an actual person, reflecting aspects of conversational speech, fluency and speaker intention. Results show that the exemplar based method is generally preferred in each of these categories. In one of the voices the participants did not find significant differences between the stimuli, which warrants further investigation and analysis. Future work also includes investigating the effects of data size on the model performance.

7. Acknowledgements

This research is supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Center for Next Generation Localisation (www.cngl.ie) at University College Dublin, Ireland. The opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of Science Foundation Ireland. The authors would like to thank Prof. Nick Campbell for valuable discussion on evaluation methodology.

8. References

[1] Klatt, D., "Interaction between two factors that influence vowel duration," *The Journal of the Acoustical Society of America*, vol. 54, p. 1102, 1973.

[2] Campbell, W.,"Syllable-based segmental duration," *Talking machines: Theories, models, and designs*, pp. 211–224, 1992.

[3] Morais, E. and Violaro, F.,"Exploratory analysis of linguistic data based on genetic algorithm for robust modeling of the segmental duration of speech," in *Ninth European Conference on Speech Communication and Technology*, 2005.

[4] Córdoba, R., Vallejo, J., Montero, J., Gutierrez-Arriola, J., López, M. and Pardo, J. "Automatic modeling of duration in a spanish text-to-speech system using neural networks," in *Sixth European Conference on Speech Communication and Technology*, 1999.

[5] Goubanova, O. and King, S.,"Predicting consonant duration with bayesian belief networks," in *Proc. of the Interspeech*, 2005, pp. 1941–1944.

[6] Walsh, M., Möbius, B., Wade, T. and Schütze, H., "Multilevel exemplar theory," *Cognitive Science*, vol. 34, no. 4, pp. 537–582, 2010.

[7] Abou-Zleikha, M. and Carson-Berndsen, J.,"Exemplar-based complex features prediction framework," in *7th International Conference on Natural Language Processingand Knowledge Engineering (NLPKE11)*, 2011.

[8] Braunschweiler, N., Gales, M. and Buchholz, S.,"Lightly supervised recognition for automatic alignment of large coherent speech recordings," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[9] Cabral, J., Renals, S., Richmond, K. and Yamagishi, J.,"Towards an improved modeling of the glottal source in statistical parametric speech synthesis,",*Proc. ISCA SSW6* 2007.

[10] Székely, E., Cabral, J., Cahill, P. and Carson-Berndsen, J.,"Clustering expressive speech styles in audiobooks using glottal source parameters,", *Proceedings of Interspeech*,2011.

[11] HTS working group, "HMM-based Speech Synthesis System (HTS)," 2011. [Online]. Available: http://hts.sp.nitech.ac.jp/. [Accessed: 20-Oct-2011].

[12] Masuko, T.,"Hmm-based speech synthesis and its applications," Ph.D. dissertation, 2002.

[13] Campbell, N.,"Evaluation of Speech Synthesis From Reading Machines to Talking Machines," *Evaluation of Text and Speech Systems*, Springer Netherlands, 2007.