Prosodically Modifying Speech for Unit Selection Speech Synthesis Databases

Ladan Golipour, Alistair Conkie, Ann Syrdal

AT&T Shannon Laboratory, Florham Park, NJ, USA

{ladan,adc,syrdal}@research.att.com

Abstract

This paper investigates the practical limits of artificially increasing the prosodic richness of a unit selection database by transforming the prosodic realization of constituent sentences. The resulting high-quality transformed sentences are added to the database as new material.

We examine in detail one of the most challenging prosodic transformations, namely converting statements into yes/no questions. Such transformations can require very large prosodic modifications while at the same time there is a need to retain as much naturalness of the signal as possible.

Our data-driven approach relies on learning templates of pitch contours for different stress patterns of interrogative sentences from training data and later on applying these template pitch contours on unseen statements to generate the corresponding questions.

We examine experimentally how the modified signals contribute to the perceived synthesis quality of the resulting database when compared with baseline unmodified databases.

Index Terms: speech synthesis, RELP, prosody

1. Introduction

Unit selection synthesis [6] can generate very natural audio output but output quality may not be consistent, depending largely on the voice and size and audio quality of the database used. Best quality output is generally produced when synthesizing in-domain text.

In this paper we examine one method of addressing some of the limitations of unit selection synthesis by means of prosodically enriching the underlying speech database.

Descriptions of several techniques for enhancing the quality of unit selection databases have been published previously, each with its own merits. They tend to focus mostly on the segmental level. The possibility of substituting units generated by a formant synthesizer was examined in [4], [5].

There have also been efforts concerned with using data from other voices to boost the effective size of a database [2], [1]. By combining voices, the overall coverage is improved along with the resulting synthesis quality.

One limitation of unit selection synthesis is that generally no prosody modification is performed. A practical consequence is that at times a desired prosodic contour for a synthetic sentence will be unavailable and substituted by a less satisfactory sequence of units. The above mentioned techniques do not tackle this issue directly.

Some interpolation or averaging techniques have been developed, for example [9], [10] where a fusion approach is employed if specific appropriate units are not available.

Our approach in this paper is to create new prosodic patterns and to add them as extra data to the database. The extra data are based on utterances already in the database, but have been transformed using signal processing to have a different prosodic realization, e.g. as a question rather than as a declarative sentence. In particular, we systematically examine taking selected sentence or phrase elements from the unit selection database and modifying them prosodically, converting nuclear pitch accents from H* to L* and end tones from L-L% to H-H% in ToBI [11] formalism, then augmenting the database with the modified data.

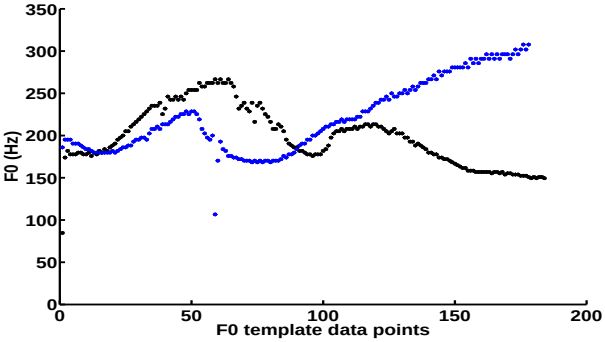


Figure 1: F0 values for a statement/question pair with stress pattern 10010.

We expect to increase the prosodic coverage of the database while maintaining quality. As part of this process we describe a robust data-driven technique for speaker-specific prosody prediction.

As an example, the prosody of the declarative sentence "Calling Michael Jordan." would typically have an high (H*) pitch accent on the first syllable of Jordan and a low end tone (L-L%), The question form as in "Was that Michael Jordan?" would typically have an low (L*) pitch accent on the first syllable of Jordan and a high (H-H%) end tone.

One challenge is the signal modification required for the task. As can be seen from Figure 1, the degree of modification needed is large (on the order of an octave phrase finally) and potentially can introduce artifacts into the signal, reducing naturalness. For this work, we used and tested Residual Excited Linear prediction (RELP) [7] and Pitch Synchronous Overlap and Add (PSOLA) [8], in combination with prosodic templates learned from data.

Additionally, there are challenges relating to data labeling accuracy, and to scaling up techniques to work effectively with large datasets that are not manually labeled.

2. Methodology and Dataset

We first describe how new prosody was generated from existing utterances using a data-driven approach. To achieve this we used a specially constructed dataset. The speech database was recorded from a female speaker of American English under controlled conditions, and is part of a larger set of recordings

designed for various synthesis experiments. The audio files are 16kHz, 16 bit audio. The prosody dataset is composed of approximately 2100 sentence pairs of the form "Calling Robert Kerr." and "Was that Robert Kerr?". Each pair uses a different combination of first name and last name. One from each pair has a declarative intonation and one a yes-no interrogative.

We randomly split the data into a training set with 1600 example pairs and a test set with 500 example pairs. We extracted only the first and last name portion of the examples using their transcriptions. All the examples are categorized based on their syllable stress pattern. Syllables with primary stress are marked *1* and all other syllables are marked *0*. Ignoring word boundaries, this leads to 10 stress-pattern classes for the data, with the distribution among classes shown in Figure 2. For example, the most common pattern, 1010 could represents names like *Richard Johnson* or *Jane Andreesen*. Next, we trained target

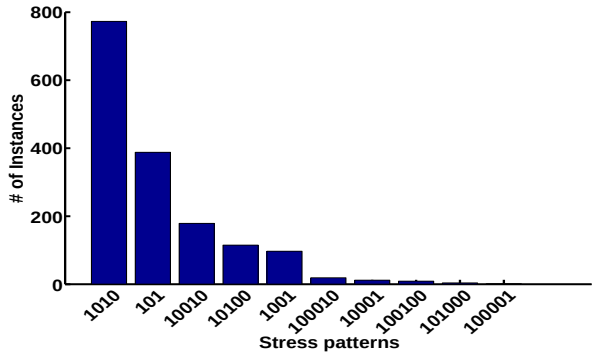


Figure 2: Frequency count of the 10 stress pattern classes in the training data.

prosody templates as described in detail below. The templates were used to generate hypotheses that can then be compared with the reference data both objectively and subjectively.

We examine multiple methods to convert the declarative pronunciation form of names to their interrogative form. In all approaches, we attempt to achieve this goal initially through pitch modification. In the first approach, we employ the PSOLA algorithm[8]. In the second approach, we decompose the speech signal into residual and LPC coefficients using one implementation of the RELP algorithm [7]. Finally, we use RELP-PSOLA [3], where PSOLA operates on the residual signal, in order to reduce the amount of distortion that PSOLA introduces to the speech signal, especially when there is a large difference between the pitch in the original and the target speech signal.

2.1. Pitch Template Computation

The first step in all approaches was to estimate a template pitch contour for the interrogative form. We observed that the shape of the pitch contour largely depends on the stress pattern of the name. Different names with similar stress patterns have similar pitch contours. Taking account of this, we categorized the interrogative training examples according to their stress pattern and estimated an average pitch contour for each category. Figure 3 displays pitch contour templates for the 10 stress patterns. In order to generate the pitch contour templates, we performed the following procedure for every stress category:

- Generate pitch marks for all interrogative training examples using the RELP algorithm and form a pitch vector from the pitch duration values.
- Rank all pitch vectors based on their length and choose the median pitch vector as the class's reference.
- Apply dynamic time warping (DTW) on all pitch vectors in order to align them with the reference pitch vector.
- Compute the mean of the aligned pitch vectors. The mean vector is still not a smooth representation of a pitch template contour due to occasional errors in the pitch marks and the performance of the DTW algorithm.
- Perform one-dimensional median filtering on the mean pitch vector to generate a smoother pitch contour template.

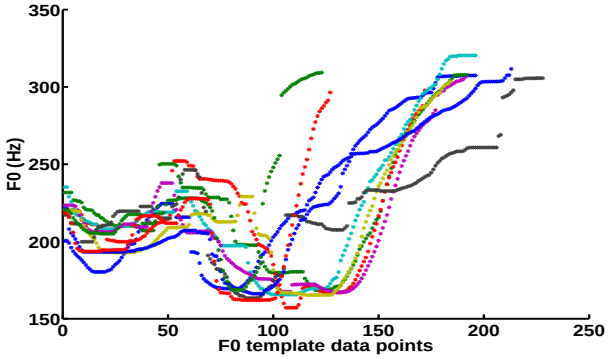


Figure 3: Interrogative pitch contour templates for different stress patterns

In this way, an interrogative pitch contour for every stress-pattern is generated. Next, we employ this contour to modify the pitch values of declarative sentences according to their stress patterns.

2.2. Declarative to Interrogative Conversion based on PSOLA

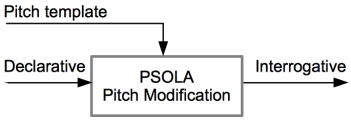


Figure 4: PSOLA conversion.

The block diagram of the first pitch modification approach based on PSOLA is depicted in Figure 4. The interrogative pitch contour is represented as a vector of pitch values. In order to use it with the PSOLA algorithm, we resampled this vector through interpolation in such a way that the summation of all pitch values in the final pitch vector is approximately equal to the length of the test example. Next, we aligned the first pitch mark of the template with the first pitch mark of the test example, and used PSOLA to modify the pitch of the example. Sometimes there is a large change in pitch value, particularly for the final syllables, for which PSOLA is not an ideal choice. Motivated by this we also examined RELP-based approaches.

2.3. Declarative to Interrogative Conversion based on RELP

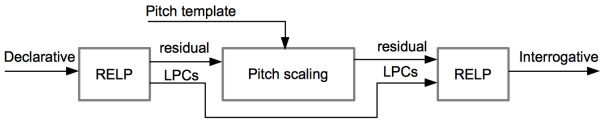


Figure 5: RELP conversion.

In the second approach, we decomposed the speech signal into residual and LPC coefficients and perform a simple modification of pitch marks through resampling the residual. The block diagram of this method is shown in Figure 5. We applied RELP to perform the decomposition, and then extracted the pitch marks from the residual signal. We resampled the pitch template vector in order to have the same number of pitch marks as the residual signal, and computed the ratios between template pitch values and test example pitch values for every adjacent pitch mark. Finally, we resampled the residual signal using this vector of ratio factors and reconstructed the hypothesis speech signal with the modified residual signal and original LPC coefficients. Since the higher pitch values (towards the end of question form) affect the length of the speech output and hence the local speech rate, we compensate for this phenomenon through resampling the speech and keeping the duration of the output speech approximately equal to the original signal.

2.4. Declarative to Interrogative Conversion based on RELP and PSOLA

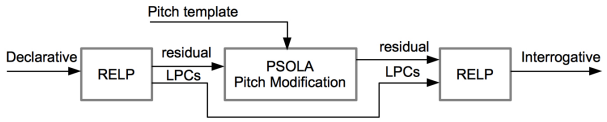


Figure 6: RELP and PSOLA

In the third approach, we combined the PSOLA- and RELP-based techniques. Instead of applying PSOLA to the test example using a template pitch contour, we decomposed the speech signal into residual and LPC envelope using RELP and applied the pitch contour on the residual signal using PSOLA. Next, we reconstructed the signal with RELP using the modified residual and the original LPC coefficients. This approach is similar to the previous one with the advantage that here, the pitch modification of the residual signal is achieved through a more sophisticated algorithm rather than a simple resampling technique. Also, the number of pitch marks does not necessarily need to be equal in the original and modified residual. If PSOLA repeats a frame, we simply use the same LPC coefficients and if it drops the frame, we ignore the LPC coefficients for that frame. Another advantage of this approach is that since PSOLA is used only to modify the residual signal, the amount of distortion introduced to the target signal is potentially less than the PSOLA-only approach.

The algorithms described above are relatively sensitive to the segmental time-alignment accuracy of the database. The phoneme boundaries for the source and target data are a result of running forced alignment recognition on the speech data and are relatively (but certainly not completely) accurate. Boundaries that are inaccurate have the potential to affect the quality of the imposed pitch curves. In practice, while this was a concern and was monitored, it did not seem to strongly impact synthesis quality for the examples we chose. However it could be a concern where large portions of a database are being processed.

Figure 7 displays pitch values for two reference examples and their relative pitch-modified hypotheses generated by the second RELP-based approach. As can be seen, the predicted pitch contours are very similar to their reference counterparts. Voiced and unvoiced data are treated differently in terms of pitch marks, so accurate matching of voiced to voiced frames and unvoiced to unvoiced frames in source and target were carefully monitored.

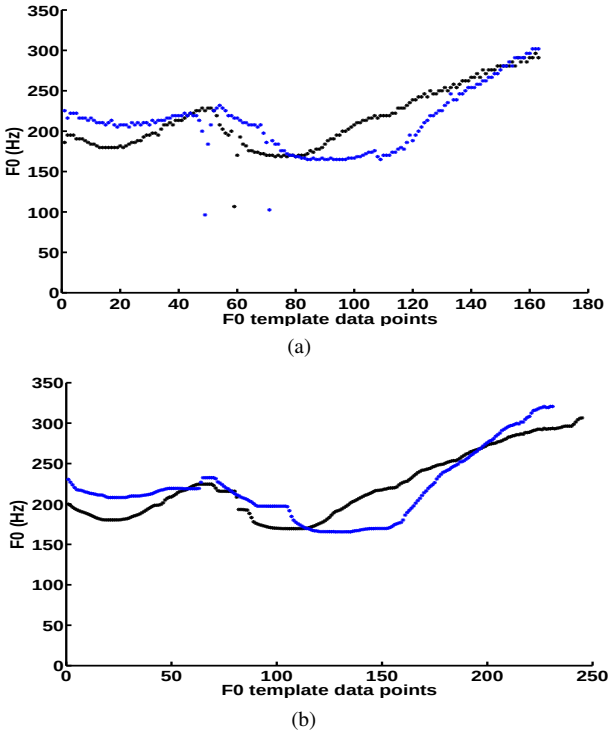


Figure 7: F0 values of (a) reference natural speech data and (b) RELP-based pitch-modified data for two sentences.

3. Experiments

The main goal of this work is to extend the database of a speech synthesizer through adding prosodically-modified units (based on existing units) to the database. These extra units can be selected and concatenated like any others in the database, potentially leading to a more natural pitch contour in the synthesized speech. We designed an experiment to test this idea. Before we carried out the main experiment described below, we performed an informal evaluation of the three approaches (PSOLA, and the two RELP variations) by listening to approximately 100 examples from each approach. We concluded that RELP-based approaches have similar performance while the the PSOLA-based approach has a slightly worse performance. We decided to include only the second RELP-based approach in the subjective experiment.

The next step was to prepare voices for the synthesis experiment. Here we describe the three synthesized database compositions. The three databases have in common 45 minutes of recordings designed to be diphone-rich. The voices differ with respect to the type of name-specific data they contain. The *high baseline* contains 55 minutes of natural recordings of complete interrogative sentences (carrier phrases and names). The *low baseline* has no interrogative sentences. The *RELP-based* voice contains 40 minutes of the RELP-based data: declarative names-only sentences (no carrier phrase) that have been converted to interrogative form using signal processing.

Once the voices were built, our standard unit selection synthesizer was used to generate stimuli for testing. We chose a set of 20 example sentences of the form "Was that James Baxter?" and synthesized them with the three different voices, giving a total of 60 stimuli. It is important to note that in choosing the sentences we used a best-case approach and preferred samples where the signal-processed audio was generated with the correct prosody by the unit selection. Figure 8 demonstrates the unit

selection patterns of test examples generated using the RELP-based voice. As expected, the units for the carrier phrase ("Was

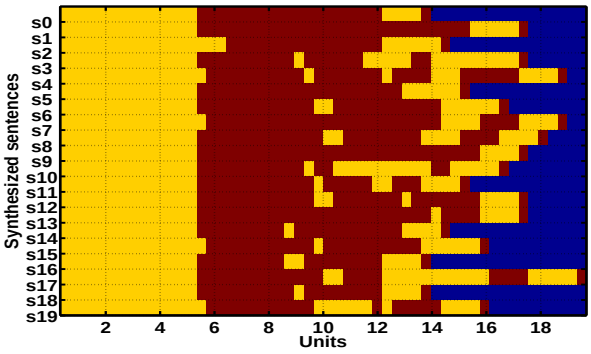


Figure 8: Units selected by the synthesizer for 20 test examples are colored based on their origin. Yellow represents "general data" and red "RELP-based pitch-modified data". Blue represents the silence that follows the end of each sentence.

that") are all extracted from the general data while a large portion of units for the name pronunciations originate in the RELP-based data. However, the synthesizer seems to select many general data units towards the end of the sentence except for a few final units for which the pitch is in its highest values. This is mainly due to the synthesis algorithm and the tradeoffs between respecting prosody requests and achieving appropriate transitions at unit boundaries.

The 60 stimuli were presented to listeners in the form of a web based test. The stimuli were presented in a different random order for each listener. Listeners were asked to judge each of the audio samples on a scale of 1 to 5 where the following descriptive terms were used 1 (Bad); 2 (Poor); 3 (Fair); 4 (Good); 5 (Excellent). There were a total of 17 listeners. Listeners were only asked to rate the sentences and were not given any specific instructions about naturalness or about focusing on prosody. There were two supplementary questions to determine (a) whether or not the listener was a native speaker of English, and (b) whether the listener heard the audio via headphones or via loudspeaker.

4. Results and Discussion

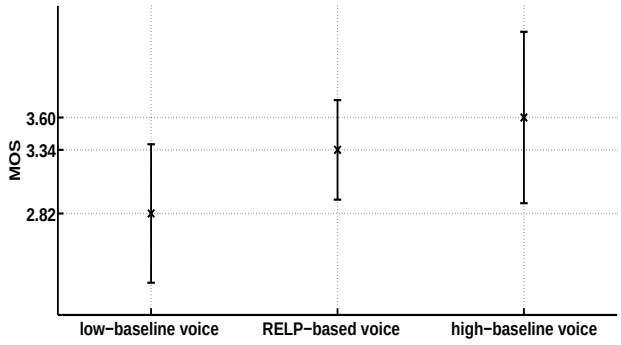
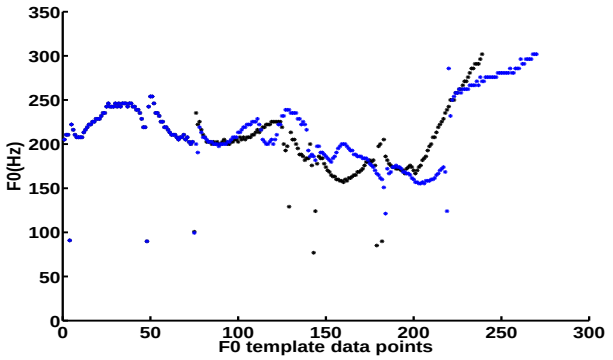
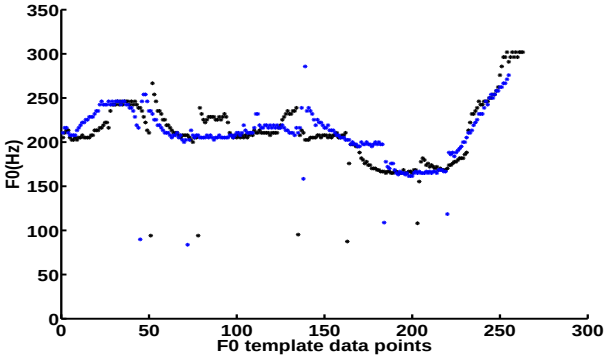


Figure 9: MOS with standard deviations for the three voices used in the web-based test.

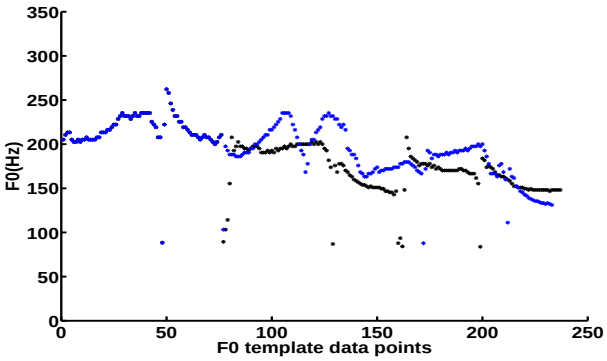
The results of the web-based test are depicted in Figure 9. The highest Mean Opinion Score (MOS) for the test, 3.60, is achieved by the high baseline system, while the low baseline score is 2.82. The RELP-based interrogative names database score of 3.34 is intermediate between the baseline conditions. The differences between all conditions are statistically significant ($p < 0.005$).



(a)



(b)



(c)

Figure 10: F0 values for two examples of synthesized speech. (a) High baseline voice. (b) RELP-based voice. (c) Low baseline voice.

Pitch contours of two test examples synthesized with three voices: high baseline, low baseline, and RELP-based are shown in Figure 10. Both high baseline and RELP-based voices successfully show a rising pitch contour towards the end of the sentence, while the low baseline voice lacks the yes-no question-form pitch contour. This is expected as no interrogative units exist in the database of the low baseline voice.

The results indicate that there was a definite overall benefit from including the prosodically-modified data in the database, but that it is not, as configured currently, a perfect substitute for including unmodified recordings. It is however encouraging that the value for the signal-modified database is closer to the high baseline than to the low baseline.

5. Conclusions and Future Work

In this paper we have shown that it is possible to use signal processing techniques to modify speech signals considerably, even far outside the typical +/- 20% range that is recommended in the literature, in order to create prosodically distinct variants of sentences. We achieved our goal by means of prosodic templates derived from a training set. We describe using these techniques to create questions from statements.

The prosodically modified data was added to our unit selection database and used at synthesis time. Effectively we have increased the size, and most importantly, the prosodic coverage of the database. The techniques are tested here on name data since we have natural target speech available, but can be applied in principle to any form of prosody modification where some parallel prosodic data is available for training.

The experimental results indicate that enriching a database prosodically, even with material that has been subjected to signal processing manipulations, can benefit synthesis quality significantly.

Future work will extend the ways in which we can generate new prosody contours for material to be added to the unit selection database.

6. References

[1] A. Conkie, and A. Syrdal, "Composite TTS Voices", 7th Speech Synthesis Workshop 2012, Kyoto, Japan.

[2] E. M. Eide, and M. A. Picheny, "Towards pooled-speaker concatenative text-to-speech", Proc. of ICASSP 2006.

[3] B. Gold, N. Morgan, D. Ellis, "Speech and Audio Signal Processing: Processing and Perception of Speech and Music".

[4] S. R. Hertz, I. C. Spencer, and R. Goldhor, "When can speech segments serve as surrogates?", Presentation at From Sound to Sense: 50+ Years of discoveries in Speech Communication 2004.

[5] S. R. Hertz, I. C. Spencer, and R. Goldhor, "Perceptual consequences of nasal surrogates in English: Implications for speech synthesis", 147th meeting of the Acoustical Society of America, 2004.

[6] A. Hunt, and A. Black, "Unit selection in a concatenative speech synthesis system using a large speech database", Proc. of ICASSP 1996, pp. 373-376.

[7] D. T. Magill, and C. K. Un, "Speech Residual Encoding by Adaptive Delta Modulation with Hybrid Companding", Proc. of National Electronics Conference 1974, pp. 403-408.

[8] E. Moulines, and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones", Speech Communication 9, pp. 453-467.

[9] T. Okubo, R. Mochizuki, T. Kobayashi, 2006. "Hybrid voice conversion of unit selection and generation using prosody dependent HMM", IEICE Trans. Inf. Syst. E89-D (11), 27752782.

[10] M. Tamura, N. Braunschweiler, T. Kagoshima and M. Akamine, "Unit Selection Speech Synthesis Using Multiple Speech Units at Non-adjacent Segments for Prosody and Waveform Generation", IEEE Proc. ICASSP2010, March 2010.

[11] K.E.A. Silverman, M.E. Beckman, J.F. Pitrelli, M. Ostendorf, C.W. Wightman, P. Price, J.B. Pierrehumbert, and J. Hirschberg, "TOBI: a standard for labeling English prosody", in Proc. ICSLP, 1992.