

Control and prediction of the impact of pitch modification on synthetic speech quality

Esther Klabbers and Jan P. H. van Santen

Center for Spoken Language Understanding
OGI School of Science & Engineering at OHSU
20000 NW Walker Road, Beaverton, OR 97006, USA
{klabbers, vansanten}@ece.ogi.edu

Abstract

In order to use speech synthesis to generate highly expressive speech convincingly, the problem of poor prosody (both prediction and generation) needs to be overcome. In this paper we will show that with a simple annotation scheme using the notion of foot structure, we can more accurately predict the shape of local pitch contours. The assumption is that with a better selection mechanism we can reduce the amount of pitch modification required, thereby reducing speech degradation. In addition, we present a perceptual experiment that investigates the degradation introduced by pitch modification using the OGiresLPC algorithm. We correlated the weighted perceptual score with different pitch and delta pitch distances. The best combination of distance measures is able to explain 63% of the variance in the perceptual scores. Decreasing the pitch is shown to have a higher impact on perception than increasing the pitch.

1. Introduction

The research presented in this paper is carried out as part of the NSF project “Prosody generation for child-oriented speech.” The goal of this project is to improve the prosodic quality of state-of-the-art text-to-speech (TTS) synthesis so that it can be used in educational applications for children with or without learning disabilities. Highly expressive speech is essential to convey the intended meaning of the spoken sentences and to capture childrens’ attention.

Current corpus-based synthesizers are capable of generating intelligible, pleasant-sounding speech. The philosophy behind such systems is to search a large speech corpus for the largest speech units that match a certain context. This is done using a limited set of phonetic and prosodic selection criteria that minimize the amount of pitch modification required (or even eliminate it altogether). Although the pitch sounds smooth within the selected speech units, more often than not it fails to highlight the intended meaning, which hinders speech understanding. Moreover, the speech corpora are often read in a neutral speaking style to avoid large pitch discontinuities between speech units. For expressive speech synthesis, a much larger speech corpus with different speaking styles would be necessary, but even in the news-reading style for unlimited domains, data sparsity is already an issue [3].

The assumption is that with better selection criteria used both for corpus design and online search, the degradation of the speech quality due to pitch modification will be reduced.

This research was conducted with support from NSF grant 0205731.

The degradation is a function of the pitch modification algorithm and the difference between the original and target pitch. Ideally, one would want a prosody manipulation algorithm that can change not only pitch and duration but also spectral characteristics of the speech without major degradation of the speech quality. But in the absence of such an algorithm we must focus on other issues to improve the prosodic quality of current TTS systems. In Klabbers et al. [1], we started our investigation into better unit selection with an annotation scheme that predicts local pitch shape.

In Section 2 of this paper we will continue the investigation with different speech data and different distance measures. We will show that a very efficient annotation scheme using the notion of the *foot structure* is capable of predicting pitch shape better than other schemes.

In Section 3 we present a perceptual experiment, in which we investigate the degradation introduced by pitch modification. We correlated the weighted perceptual score with different pitch and delta pitch distances. The best combination of distance measures is able to explain 63% of the variance in the perceptual scores. Decreasing the pitch is shown to have a higher impact on perception than increasing the pitch.

2. Speech corpus analysis

2.1. Foot-based factors

Most corpus-based TTS systems available today use a small set of linguistic features to select units from different prosodic and phonetic contexts. The most commonly used features are lexical stress, accent and syllable finality. In Klabbers et al. a more sophisticated scheme was presented using the notion of the *left-headed foot*. A left-headed foot is defined as a sequence of one or more syllables, such that only the first syllable is accented (i.e., is the stressed syllable in an emphasized word). We call that syllable the *head* of the foot. A foot is always followed by either an accented syllable or a phrase boundary. The use of feet is based on these considerations. For a typical accent-lending up-down pitch movement, we observe a rise-fall within the accented syllable for monosyllabic feet, and a rise in the accented syllable followed by a fall in the subsequent unaccented syllables in polysyllabic feet [4]. This means that accented syllables should be differentiated in terms of whether they occur in mono- or polysyllabic feet, but no further distinctions are needed. For unaccented syllables, the key distinction is whether they occur right after the accented syllable or later in the foot, where the number of unstressed syllables following it is irrelevant.

Table 1 presents the different factorization schemes and the

Simple	Foot
stress {0,1}	syll. to last accent {0,1,2}
accent {0,1}	syll. to next accent {0,1,(2)}
phrase-fin. syll. {0,1,2}	phrase-fin. foot {0,1,2}
Nr. of levels: 12	Nr. of levels: 19
Complex1	Complex2
accent {0,1}	accent {0,1}
syll. to last accent {0,1,2}	syll. to last accent {0,1,2,3}
syll. to next accent {0,1,2}	syll. to next accent {0,1,2,3}
phrase-fin. foot {0,1,2}	phrase-fin. foot {0,1,2}
Nr. of levels: 54	Nr. of levels: 96

Table 1: Factors and factor levels in each factorization scheme.

factor levels they distinguish. The Simple Scheme speaks for itself. The factors *stress* and *accent* are binary. The *phrase-final syllable* factor distinguishes syllables in a medial (0), phrase-final (1) and utterance-final (2) position. In the Foot Scheme, the factor *syll. to last accent* refers to the number of syllables the current syllable is removed from the previous accented syllable. We hypothesize that for the head of the foot the preceding context is not relevant for the shape of the pitch contour. Hence, when the current syllable is accented (and stressed), the value of is 0. For unaccented syllables the factor gets a value of 1 if it follows the head, or 2 if it is one or more syllables removed from the head. The *syll. to next accent* factor is 0 when the current syllable is the last syllable in a foot, i.e., it is followed by either the head of a following foot or a phrase boundary. The factor is 1 when it is at least one syllable removed from the end of the foot. The value 2 is reserved for phrase-initial syllables that are either stressed and unaccented, or unstressed and accented. We call these *orphan* feet. The *phrase-final foot* factor distinguishes feet in medial, phrase-final and utterance-final position.

In the Complex1 and Complex2 schemes the *syll. to last accent* factor is encoded slightly differently. It has a value of 0 if the previous syllable is accented, 1 if it has one syllable between the accent and the current syllable, and 2 if there are two syllables. This allows us to verify our assumption that the context preceding the head of the foot is irrelevant. Because we lose information in this annotation scheme about the accent status of the current syllable, an extra *accent* factor has to be included. The difference between Complex1 and Complex2 is that the second scheme has a longer window on both the *syll. to last accent* and *syll. to next accent* factor.

2.2. Speech corpora

Duration corpus: The speech corpus used in [1] was originally recorded for the purpose of training a prediction model for segmental duration. It contains 472 sentences spoken by a female speaker. It was segmented by hand and annotated with several factors including stress and accent. In our analysis we only included all-sonorant CV (consonant-vowel) and CVC syllables, resulting in 1467 syllables.

Foot corpus I: This corpus was recorded specifically for testing the effect of the position in the foot structure on the local pitch contour. It contains 285 sentences, spoken by a female speaker with a highly expressive voice. In each sentence the target is an all-sonorant CVC syllable. It occurs in 19 foot contexts, with 5 different vowels in 3 repetitions. The syllables are *moon*, *mean*, *main*, *mine*, and *moan*. Table 2 shows the type of sentences that were recorded in this corpus.

This corpus showed substantial consistency in pitch con-

tours. However, the set-up of the material and the fact that our speaker has a very lively speaking style, leading to a corpus where most accents were realized by a contrastive pitch contour, where the pitch movement on the head of the foot has a much smaller effect on subsequent syllables than with a typical pitch accent.

Foot corpus II: To obtain additional material, we recorded a second corpus, this time instructing our speaker to speak more neutrally and having the sentences be less contrastive. The recordings turned out to have very flat pitch contours, and unfortunately the speaker was obviously uncomfortable speaking in this mode.

2.3. Distance measures

We computed pairwise distances between each pitch contour and every other pitch contour in the three corpora. The pitch values were measured at 5-ms intervals using ESPS Waves+[5]. The pitch values in each contour were interpolated such that each phoneme in the syllable was represented by 50 data points. In the case where one syllable was of type CV and the other of type CVC, the data values in the last consonant were ignored. In the previous paper we used two distances, the Root-Mean Square Error (RMSE) and the maximum delta distance, to measure homogeneity within categories. The maximum delta distance returns the largest distance between the delta of a template pitch contour (representing the average of two pitch contours) and the two pitch contours. We decided to use different measures to reflect differences in absolute pitch and directional change. We define D_p as the sum of squares of the differences in log pitch values between pitch contours i and j . It corresponds to the area between two pitch curves.

$$D_p = \sum (\log_{10}(F_{0i}) - \log_{10}(F_{0j}))^2 \quad (1)$$

Let D_{wp} be D_p weighted with the energy in the syllable. This is to test our hypothesis that pitch modification in higher energy regions (i.e. vowels) leads to higher perceived degradation than lower energy regions (i.e. nasals).

$$D_{wp} = \frac{\sum E(\log_{10}(F_{0i}) - \log_{10}(F_{0j}))^2}{\sum E} \quad (2)$$

The delta distance $D_{\Delta p}$ is the sum of squares of the differences in the first derivative of the log pitch values, which relates to the shape of the pitch contour change. Our hypothesis is that we can change the pitch substantially without audible distortion provided the pitch shapes are fairly similar.

$$D_{\Delta p} = \sum (\Delta \log_{10}(F_{0i}) - \Delta \log_{10}(F_{0j}))^2 \quad (3)$$

The weighted delta distance $D_{w\Delta p}$ is the delta distance weighted with the sum of the energy.

$$D_{w\Delta p} = \sum E(\Delta \log_{10}(F_{0i}) - \Delta \log_{10}(F_{0j}))^2 \quad (4)$$

where $E = \sqrt{E_i \times E_j}$.

2.4. Results

Table 3 provides the within-cell means for each factorization scheme. A factorization scheme performs better when the mean distance within its cells is lower than those in another scheme. It can be observed that for all three corpora, the Foot annotation scheme performs better than the Simple annotation scheme.

He lives on [_F MOON LANE]	He lives in [_F MOON], [_F Oregon]	I saw the [_F MOON]
He lives in [_F MOON river] [_F Oregon]	He lives in [_F MOONery], [_F Oregon]	I saw the [_F MOONery]
He lives in [_F BLUE moon] [_F Oregon]	He lives in [_F BLUE moon], [_F Oregon]	I saw the [_F BLUE moon]
I saw the [_F BLUE moonery a] [_F GAIN]	He lives in [_F BLUE moonery] [_F Oregon]	I saw the [_F BLUE moonery]
He lives in [_F SILver moon] [_F RIVER]	He lives in [_F SILver moon] [_F Oregon]	I saw the [_F SILver moon]
I saw the [_F SILver moonery a] [_F GAIN]	He lives in [_F SILver moonery] [_F Oregon]	I saw the [_F SILver moonery]
The moon [_F RIVER place]		

Table 2: Sentences recorded in Foot Corpus I. The brackets show foot boundaries.

	Simple	Foot	Complex1	Complex2
Duration corpus				
Levels	6	19	34	34
D_p	7.11	6.81	7.46	7.46
D_{wp}	0.065	0.064	0.069	0.069
$D_{\Delta p}$	0.028	0.015	0.018	0.018
$D_{w\Delta p}$	26.62	15.74	17.01	17.01
Foot corpus I				
Levels	6	19	17	17
D_p	7.11	2.20	2.40	2.40
D_{wp}	0.050	0.015	0.016	0.016
D_{Δ}	0.024	0.016	0.016	0.016
$D_{w\Delta p}$	18.85	11.59	12.11	12.11
Foot corpus II				
Levels	6	19	16	21
D_p	3.98	3.42	3.92	4.15
D_{wp}	0.026	0.024	0.028	0.028
$D_{\Delta p}$	0.017	0.015	0.013	0.011
$D_{w\Delta p}$	4.22	3.78	3.67	3.25

Table 3: Average within-cell means for the different factorization schemes.

Also it is generally better than the Complex annotation scheme.

Inspecting the individual pitch contours, especially in Foot corpus I, we observed that some levels in the foot annotation scheme could be further collapsed. For the head of the foot, it does make a difference whether there are unstressed syllables following it or not. For unstressed syllables, it only makes a difference whether they are immediately preceded by the head of the foot or not. And for all syllables it is important to take into account whether they occur in a phrase-medial foot, a phrase-final foot with continuation rise, or an utterance-final foot. This leads to a new factorization scheme with 12 levels. A recomputation of the within-cell means for Foot Corpus I confirms that the within-cell means for the simplified Foot scheme are still substantially lower than those for the Simple scheme (D_p is 3.13, D_{wp} is 0.021, $D_{\Delta p}$ is 0.019 and $D_{w\Delta p}$ is 15.29).

In conclusion, the Foot annotation scheme provides a very simple mechanism for selecting units with a certain local pitch shape. So far we have assumed that with better selection in corpus design and search the necessary pitch modification will be of an innocuous nature. To date, there has not been any investigation into the amount of pitch modification that is permissible without degrading the perceived speech quality. In the next section we present a perceptual experiment designed to investigate that issue.

3. Perceptual experiment

In this paper we restrict ourselves to pitch modification using OGiresLPC [2]. Other modification algorithms are likely to lead to other results, but OGiresLPC is one of the few algorithms that is widely available as part of Festival and is freely available for non-commercial use (<http://cslu.cse.ogi.edu/tts/>). The results of the experiment are correlated to different distance measures to shed more light on the amount and nature of pitch modification that can be achieved without degrading the perceived naturalness of the speech.

3.1. Material

The material is taken from Foot corpus I in which all-sonorant CVC syllables occur in different prosodic contexts. This corpus contains highly expressive speech from a female speaker. Her average F_0 lies around 200 Hz, but highly emphasized words can go as high as 600 Hz. The sentences used in the perceptual experiment were constructed as follows. First, they were spliced into a carrier phrase part and a target word part. The target word was always one of the sonorant CVC syllables mentioned above. Then they were recombined. In Version A the target word is replaced by a repetition of the same word in the same prosodic context. In Version B the target word was replaced by the same word in a different prosodic context. Both sentences were then resynthesized in Festival with OGiresLPC, transplanting the original pitch and duration onto the new target word. The sentence parts were then concatenated using Snack[6]. Discontinuities at splicing boundaries were avoided by inserting a small 20-ms pause between carrier phrase and target word, and fade-in and -out of the waveforms. Finally, the global energy of the inserted target words was adjusted manually to match that of the control version. This was necessary as some target words were taken from utterance-final position where they had a much lower energy than in utterance-medial position. This had no measurable effect on the speech quality.

3.2. Method

The task of the listeners was to judge which one in a pair of sentences (version A and B) had the better voice quality. They could indicate their preference on a 7-point CMOS scale, ranging from -3 (A is much worse than B) to 3 (A is much better than B). Ten subjects listened to 90 sentence pairs using a web-based interface and high quality headphones. The test was preceded by a training set of 3 sentence pairs. The sentences were randomized for each subject and the order (AB versus BA) was balanced.

3.3. Results

The Pearson's correlation shows that agreement amongst listeners varies between 0.2 and 0.5. We computed a weighted final score per sentence, by performing a z-score normalization of

the subjects' ratings with the overall mean and standard deviation and weighing that with the eigenvector of the correlation matrix between the subjects. Thus we can correct for variation in subjects' ratings and differences in scaling. Then we performed linear regression analysis, using the different pitch distances to predict the weighted average scores. Table 4 displays the statistics returned by the linear regression analysis using the different distance measures as terms. At first we found the highest variance explained by a combination of the pitch distance D_p and the delta distance D_d . Surprisingly, the weighted pitch distance D_{wp} and the weighted delta distance D_{wd} did not yield higher correlations. This suggested that the pitch distortion was observed both in the vowel and in the surrounding nasals. But when the material is more varied and contains other voiced sounds such as voiced fricatives, the weighted distances might give better correlations.

We suspected that the direction of pitch change might have an effect on the perceived quality. We therefore computed two new distance measures D_{dp} and $D_{d\Delta p}$. The distances between log pitch values are split into two sets, one set where the pitch is increased D_p^+ and one set where it is decreased D_p^- . The same holds for the distances between the delta log pitch values.

$$D_{dp} = \alpha \times D_p^+ + (1 - \alpha) \times D_p^- \quad (5)$$

$$D_{d\Delta p} = \beta \times D_{d\Delta p}^+ + (1 - \beta) \times D_{d\Delta p}^- \quad (6)$$

In Figure 1 we have plotted the value of R^2 obtained in the linear regression analysis for different values of α and β . The highest correlation is found at an α value of 0.25 and a β value of 0.625. To translate this back to the measure, this means that the highest correspondence between the subjects' scores and the distance measures is found when the negative pitch modifications are weighed 4 times as much as positive pitch modification. In other words, it costs more to decrease the pitch than to increase it. It was already known that TD-PSOLA has more difficulty with decreasing pitch than increasing it. The same will hold for OGiresLPC albeit to a lesser extent as the waveform is passed through the LPC filter first. For the delta pitch distance, positive changes are weighed more than negative ones. As can be seen, the combination of D_{dp} and $D_{d\Delta p}$ leads to a major improvement, explaining 63% of the variance. Using the raw pitch values instead of the log pitch decreased the performance of the model. The duration difference between the target and original syllables did not contribute significantly.

4. Conclusion

We presented a foot-based scheme that helps to minimize pitch mismatch in corpus design and search. With a small amount of parameters we can improve prediction of the local pitch shape of a syllable. Additionally, we showed that the perceived pitch mismatch can to a large extent be predicted by combining two objective distance measures that describe differences between log pitch values and their first derivative penalizing decreasing pitch modifications more than increasing modifications. Ideally, we would want a prosody modification algorithm that is not only capable of changing duration and pitch, but also spectral characteristics, thus leading to a less intrusive method.

5. References

[1] E. Klabbbers, J. van Santen and J. Wouters, "Prosodic factors for predicting local pitch shape", In *Proceedings 2002 IEEE Workshop on Speech Synthesis, Santa Monica, CA, 2002*.

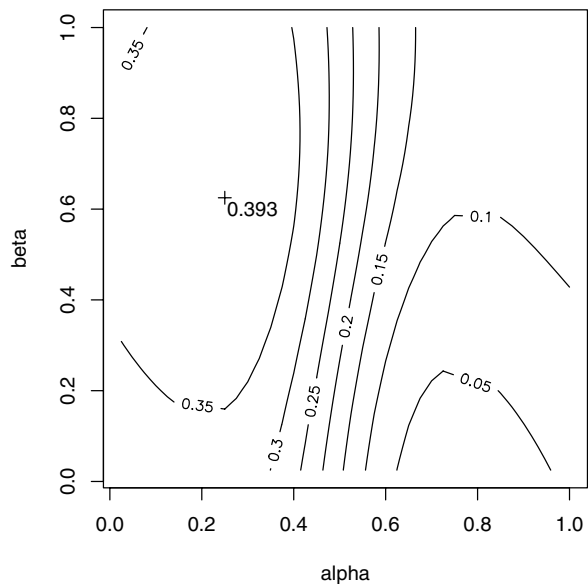


Figure 1: Contour plot of R^2 for different values of α and β , weights attached to direction of pitch change.

Measures	R^2	R	F/t-value	p-value
D_p	0.117	0.341	F(1,88) = 11.5	0.001
$D_{\Delta p}$	0.087	0.295	F(1,88) = 8.37	0.005
$D_p + D_{\Delta p}$	0.246	0.496	F(2,87) = 14.2	4.614e-6
D_p			t(-4.286)	4.67e-5
$D_{\Delta p}$			t(-3.881)	0.0002
D_{wp}	0.074	0.271	F(1,88) = 6.97	0.0098
$D_{w\Delta p}$	0.064	0.254	F(1,88) = 6.05	0.0158
$D_{wp} + D_{w\Delta p}$	0.187	0.432	F(2,87) = 9.98	0.0001
D_{wp}			t(-3.617)	0.0005
$D_{w\Delta p}$			t(-3.480)	0.0008
D_{dp}	0.256	0.506	F(1,88) = 30.23	3.71e-7
$D_{d\Delta p}$	0.103	0.320	F(1,88) = 10.07	0.002
$D_{dp} + D_{d\Delta p}$	0.39	0.627	F(2,87) = 28.16	3.72e-10
D_{dp}			t(-6.450)	6.07e-9
$D_{d\Delta p}$			t(-4.435)	2.68e-5

Table 4: Results of a linear regression analysis for the different distance measures.

[2] M. Macon, A. Cronk, J. Wouters and A. Kain, "OGiresLPC: Diphone synthesizer using residual-excited linear prediction", *Technical Report CSE-97-007, Dept. of Computer Science, Oregon Graduate Institute of Science and Technology, Portland, OR, Sep 1997*.

[3] B Möbius, "Rare events and closed domains: two delicate concepts in speech synthesis", *Proceedings of the 4th ISCA Tutorial and Research Workshop on Speech Synthesis, Blair Atholl, UK, 41-46, 2001*.

[4] J. van Santen and J. Hirschberg, "Segmental effects on timing and height of pitch contours", In *Proceedings IC-SLP'94, Yokohama, Japan, p719-722, 1994*.

[5] D. Talkin, *ESPS.*, Entropic Research Lab Inc., 1993.

[6] The Snack Sound Toolkit, <http://www.speech.kth.se/snack/>.