

# PROSODIC FACTORS FOR PREDICTING LOCAL PITCH SHAPE

*Esther Klabbers, Jan van Santen and Johan Wouters*

Center for Spoken Language Understanding  
OGI School of Science and Engineering at OHSU  
20000 NW Walker Road, Beaverton, OR 97006, USA

## ABSTRACT

In this paper, we investigate the predictive power of different prosodic factorization schemes with respect to pitch movement. We will use this to propose an extension of a standard diphone database with diphones that have been recorded in different prosodic contexts. The goal of this research is to reduce the amount of pitch modification required, thereby improving the segmental quality of the synthetic voice. We will present a factorization scheme based on the foot structure of utterances and show that this efficient scheme results in a fairly small number of additional diphones that need to be recorded.

## 1. INTRODUCTION

One major problem with diphone synthesis is that there is typically only one token for each diphone, generally coming from a stressed context. This results in hyperarticulated speech, and requires intrusive signal modification to enforce a target pitch contour on the concatenated diphones.

In the last decade we have seen the emergence of corpus-based synthesis systems to circumvent this problem [1, 3]. The philosophy behind the corpus-based approach is that one can search a large speech corpus, typically containing hours of speech, using a limited set of phonetic and prosodic selection criteria, to find the longest stretches of speech that match a certain context. This would minimize the amount of signal modification (or even eliminate it altogether). The resulting speech can sound very natural, but success is not always guaranteed. There are a number of drawbacks:

(1) the corpus contains hours of speech, making it unsuitable for use in embedded devices; (2) it poses a greater challenge on the speaker to speak with a constant style, voice quality, and energy, which also requires stricter monitoring by the recording staff; (3) the selection criteria must be simple enough to allow adequate coverage of all the contexts,

---

This research was conducted with support from NSF grants 0117911 "Making Dysarthric Speech Intelligible" and 0082718 "Modeling Degree of Articulation for Speech Synthesis", and from Intel and NTT. The authors wish to thank Vincent Pagel for his participation in this project.

but it must be predictive enough of phonetic/prosodic factors such as pitch to achieve speech with smooth and acceptable pitch contours.

In this paper we compare different prosodic factorization schemes for predicting the local pitch contour shape in a syllable. We compute different distances between pitch contours that capture the shapes rather than the absolute values of the contours. The scheme with the best predictive capability can be used either for selecting units in a corpus-based synthesis system, or for selecting and recording additional diphones in different prosodic contexts. Using the latter approach, one would end up with a much smaller corpus than needed for corpus-based synthesis, have full control over the prosodic contexts that are covered and maintain a clear articulation of the units.

The idea of incorporating more tokens for each diphone is not new. Drullman and Collier [4] presented a diphone synthesis system with accented and unaccented diphones. Their synthesis quality did not improve however, due to the fact that their speaker did not produce any vowel reduction in unaccented contexts. More recently, Barry et al. [2] have included diphones in four conditions: +/- stress and +/- accent and they obtained significant improvements as a result. Most corpus-based synthesis systems also use these factors along with a factor signaling phrase-finality. We hypothesize that these three factors are not adequate predictors of pitch contour shapes, and that a more sophisticated scheme is needed.

## 2. SPEECH CORPUS ANALYSIS

### 2.1. Prosodic factors

We advocate the use of the *left-headed foot* in our factorization scheme. 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 either by an accented syllable or by 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 [5]. 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 (with little difference made between whether there are further syllables in the same foot), or later in the foot.

In order to substantiate this choice we analyzed pitch movements in a speech corpus and compared them to four factorization schemes.

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

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

Table 1 presents the different factorization schemes and the factor levels they distinguish. The Simple Scheme speaks for itself. The factor *stress* is binary and distinguishes stressed from unstressed syllables, the factor *accent* distinguishes accented from unaccented words. The *phrase-final syllable* factor distinguishes syllables in a medial, phrase-final and utterance-final position. In the Foot Scheme, the factor *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 *last accent* is 0. For unaccented syllables *last accent* gets a value of 1 if it follows the head, or 2 if it is one or more syllables removed from the head. The *next accent* factor is 0 when the current syllable is the last syllable in a foot, i.e., it is followed either by the head of a following foot or by 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. The number of factor levels for the Foot Scheme is 19 and not 24

because syllables in orphan feet never occur in phrase-final or utterance-final position.

There are two more complex factorization schemes, where the *last accent* factor is encoded slightly differently. It has a value of 0 if the previous syllable is accented and so on. This allows us to verify our assumption that the context preceding the head of the foot is irrelevant. Because the *last accent* factor now looks back and the *next accent* factor looks ahead, we lose information about the accent status of the current syllable. An extra *accent* factor therefore has to be included. The difference between Complex1 and Complex2 is that the second scheme has a longer window on both the *last accent* and *next accent* factor.

## 2.2. Speech corpus

The speech corpus used in our analysis was originally recorded for the purpose of training a prediction model for segmental duration. It contains 472 sentences and was spoken by a female speaker. It was segmented by hand and annotated with several factors including stress and accent. The accent status was indicated by three members of our corpus group. Because there was a fair amount of interrater variability, we counted a word as accented when two or more people marked it as such. The total number of syllables in the corpus is 8860, but we only included in our analysis syllables starting with a sonorant, to avoid any problems with segmental perturbations in the pitch and with absence of pitch in unvoiced segments. We ended up with a total of 1493 syllables. They consisted of a sonorant, a vowel and zero or more trailing sonorants. The pitch values were measured at 0.05 ms intervals using ESPS Waves+. A smoothing algorithm was applied to remove outliers. The pitch contours were then upsampled such that each syllable contour contained an equal number of pitch values to remove the effect of syllable duration.

## 2.3. Measures

We computed pairwise distances between each pitch contour and every other pitch contour in the corpus. The distance measures that we used reflect the following assumptions that we make about pitch contour modification:

(1) pitch modifications (within a certain range) that leave the pitch contour shape intact have a smaller impact on the speech quality than modifications that change the shape; (2) pitch modifications that change the slope direction of the pitch contour have a larger impact on the speech quality than modifications that don't.

**RMSE:** The Root-Mean Square Error is a commonly used measure to compute distances between two pitch contours. In our case, we wish to separate differences in time alignment from differences in pitch values. The RMSE is therefore computed between a pitch contour  $x$  and a pitch

contour  $\hat{x}$  that has been predicted from the other pitch contour  $y$  by estimating parameters  $a$  and  $b$  using a least-squares fit between  $\hat{x} = ay + b$  and  $x$ . In this way, the overall shape of both contours is compared, allowing for changes in the absolute pitch values and in the pitch range. Any mismatches in peak alignment result in a higher RMSE. In order to make the distance symmetrical, the RMSE is also computed for pitch contour  $y$  and its derived version  $\hat{y}$  and the overall RMSE is an average of both RMSE's.

**Delta Distance:** This returns the largest distance between the delta of a template pitch contour (representing the average of the two contours under investigation) and the deltas of those two pitch contours. It takes into account the slope of the pitch contours and the direction of change.

## 2.4. Results

Table 2 provides the within-cell means for each factorization scheme. A factorization scheme performs better when the mean distance within its cells are lower than those in another scheme. For the Simple and the Foot Schemes all possible factor combinations are found in the corpus. For the Complex1 Scheme only 30 out of the possible 54 combinations were found, and for the Complex2 Scheme only 48 out of the possible 96 combinations. The RMSE and Delta Distance decrease with increasing complexity of the factorization scheme. The exception is the Delta Distance which is lower for the Foot Scheme than for the Complex1 Scheme.

Mean	Simple	Foot	Complex1	Complex2
Levels	12	19	30	48
RMSE	13.1	12.8	12.7	11.9
Delta Distance	11.9	10.9	11.3	10.4

**Table 2.** Average within-cell means for the different factorization schemes.

It is clear that the Foot Scheme performs better than the Simple Scheme, although the improvement is not large. The RMSE values between the Foot Scheme and the Complex1 Scheme are almost equal, but the Delta Distance is better for the Foot Scheme. However, when the number of levels in the complex scheme is expanded, it performs better than the Foot Scheme. These results are encouraging, especially when you consider the fact that the recordings were not controlled for pitch movement and contain odd phrasing at times. We want to keep the number of factor levels as small as possible, to reduce the number of prosodic contexts in which the diphones have to be recorded. In that respect, we should rule out the Complex2 Scheme because it has too many factor levels, and the Simple Scheme because it is not performing as well as the Foot Scheme. We can further compare the adequacies of the Foot and the Complex1

Scheme by looking at two specific hypotheses.

**Hypothesis 1:** The distinction between medial, phrase-final and utterance-final feet is important for predicting pitch contour shapes.

To test this hypothesis we looked at the heads of polysyllabic feet. In the Foot Scheme these would end up in three different groups depending on whether they are medial, phrase-final or utterance-final feet. In the Complex1 Scheme they are all classified as phrase-medial syllables. We performed a t-test on the peak height, peak location and slope between syllables in medial and utterance-final feet. The peak height is defined as the maximum value in the contour. The peak location is the frame at which the highest pitch is found divided by the total number of frames. The slope is determined by a linear fit. Significant effects were found for peak height ( $t = 4.75, df = 44.8, p < 0.001$ ) and slope ( $t = 3.81, df = 45.9, p < 0.001$ ), but not for peak location ( $t = 1.33, df = 42.2, p = 0.19$ ). The mean peak height was higher in medial syllables than in utterance-final syllables (329.5 vs. 281.7 Hz) and the slope was larger for medial syllables than for utterance-final syllables (1.6 vs. 0.6 Hz/n<sup>1</sup>).

**Hypothesis 2:** The position of the previous accented syllable is irrelevant if the current syllable is the head of the foot.

In the complex scheme we separated the accent status of the current syllable from the distance to the previously accented syllable to test whether that is an important distinction to make. Looking at syllables that form the head of a foot, we found significant effects for peak location ( $t = -5.30, df = 296, p < 0.001$ ) and slope ( $t = -8.25, df = 267.3, p < 0.001$ ), but only between syllables that were immediately preceded by an accented, stressed syllable and those that were not. This is the case when a monosyllabic foot precedes the current one. It gives a so-called *stress clash* and causes the peak location to be earlier in the syllable (45 vs. 67%) and the slope to be drastically different (-0.1 vs. 1.6 Hz/n). Peak height was not significantly different ( $t = -0.26, df = 298.8, p = 0.79$ ).

## 3. TEXT CORPUS ANALYSIS

### 3.1. Text corpus

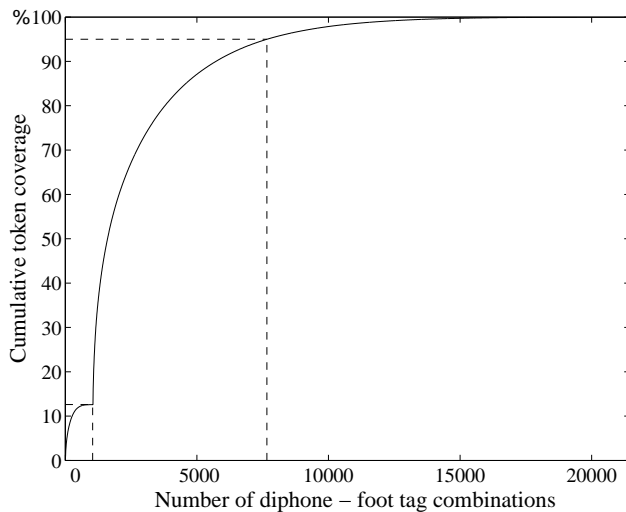
Using the current Foot Scheme, we analyzed a large text corpus to gain an insight into the frequency distribution of diphones with different foot tag combinations. The text

<sup>1</sup>n refers to the number of measurement frames in the contour, which we set to 50

corpus is a collection of 359,576 sentences from newspaper texts, the bible and a number of contemporary American novels. We computed foot factor levels for each diphone using Festival [6] and ended up with 16,926,727 diphones. The type count (i.e., the number of distinct tokens) was 22,865. We only compute foot factors for vowels. For consonants we make a distinction between onset and coda. For simplicity's sake, we disregard the consonant's position in the syllable for now and assume that we need not record multiple versions of consonant-consonant diphones. When we exclude these diphones from the database we end up with 9,367,407 tokens and 21,458 types of consonant-vowel (CV), vowel-consonant (VC) or vowel-vowel (VV) diphones with attached foot tags.

Our standard diphone database contains 3353 diphones, with 1836 tokens of the CV, VC or VV type and 1285 types. In fact, we found that only 1043 of those actually occurred in the large text corpus. Those units not found were generally rare VV diphones with a rare foot tag combination. Most of them have been recorded in a phrase-medial context where they are the head of a two-syllable foot.

### 3.2. Results



**Fig. 1.** Number of diphone - foot tag types to record against token coverage.

Figure 1 shows the cumulative token coverage against the number of diphone - foot tag type contexts that need to be added to the database. The first 12.6% are the types that already occur in our standard database. With the basic diphone database and an extension of 6020 extra diphones we cover 95% of all diphone - foot tag tokens in the text corpus, increasing the size of the diphone database by a factor three. This is many times smaller than the average corpus used in corpus-based synthesis.

## 4. CONCLUSION

In this paper we have presented several factorization schemes for predicting local pitch contour shapes. We have seen that the foot-based factorization scheme performs slightly better than the Simple Scheme and has some nice properties that are not captured by the more complex schemes. For instance, pitch contours are markedly different in heads of utterance-final feet than of medial feet. The Foot Scheme can be used to extend a diphone database with diphones that have been recorded in different prosodic contexts such that the amount of pitch modification is minimized. An additional advantage is that these diphones will show different degrees of vowel reduction dependent on the context in which they have been recorded. Analysis of a large text corpus has shown that in order to cover 95% of the diphone - foot tag combinations in that corpus, the diphone database needs to be expanded by only a factor three.

The next step in our research is to actually record these additional diphones. This allows us to run perceptual experiments to validate our hypotheses and see if we obtain an improved synthesis quality.

## 5. REFERENCES

- [1] M. Balestri and A. Pacchiotti and S. Quazza and P.L. Salza and S. Sandri. "Choose the best to modify the least: A new generation concatenative synthesis system", In *Proceedings EUROSPEECH'99, Budapest, Hungary*, p2291-2294, 1999.
- [2] W. Barry, C. Nielsen and O. Andersen. "Must diphone synthesis be so unnatural?", In *Proceedings EUROSPEECH'01, Aalborg, Denmark*, p975-978, 2001.
- [3] M. Beutnagel, A. Conkie, J. Schroeter, Y. Stylianou and A. Syrdal. "The AT&T NextGen TTS System", In *Proceedings of the Joint Meeting of ASA, EAA and DAGA, Berlin, Germany*, 1999.
- [4] R. Drullman and R. Collier. "On the combined use of accented and unaccented diphones in speech synthesis", *Journal of the Acoustical Society of America* (90), p1766-1775, 1991.
- [5] J. van Santen and J. Hirschberg. "Segmental effects on timing and height of pitch contours", In *Proceedings ICSLP'94, Yokohama, Japan*, p719-722, 1994.
- [6] P. Taylor, A. Black and R. Caley. "The architecture of the Festival speech synthesis system", In *Proceedings of the 3rd ESCA/COCOSDA Workshop on Speech Synthesis, Jenolan Caves, Australia*, p147-152, 1998