RE-DEFINING INTONATION FROM SELECTED UNITS FOR NON-UNIFORM UNITS BASED SPEECH SYNTHESIS

Baris Bozkurt

MULTITEL A.S.B.L.

Initialis Scientific Park,
Copernic Avenue,
B-7000 Mons, Belgium
bozkurt@multitel.be

Thierry Dutoit

TCTS Lab, Faculté
Polytechnique De Mons

Initialis Scientific Park,
Copernic Avenue,
B-7000 Mons, Belgium
thierry.dutoit@fpms.ac.be

Vincent Pagel

MULTITEL A.S.B.L.

Initialis Scientific Park,
Copernic Avenue,
B-7000 Mons, Belgium
pagel@multitel.be

ABSTRACT

In this work, we propose a new algorithm for defining intonation curves from selected units in a non-uniform units-based text-to-speech synthesis system. Since the main trend in a non-uniform units-based system is to select the best and modify the least to achieve highly natural synthetic speech, the target intonation imposed on units is of great importance. We propose a 'shift-only' algorithm to re-define target intonation from selected units, which does not modify the general prosodic characteristics (micro-prosody, melodic movements) of units, while efficiently reducing F0 discontinuities at concatenation points. For the operation, a cost function is defined as a summation of discontinuities and shifts scaled by durations of the units. Minimizing this function for the shift variable, we optimize minimum shift and minimum discontinuity constraints.

1. INTRODUCTION

The recent improvements in the unit selection and corpus preparation technology showed that very high quality synthetic speech can be produced with a 'select the best, modify the least' paradigm [1,2,3,4,5,6]. An important factor to the naturalness of resultant synthetic speech is the target intonation curves imposed on units.

Though the unit selection module of such a system tries to guarantee F0 continuity at concatenation boundaries by including actual pitch values in target cost and pitch discontinuities in concatenation cost functions [6], some discontinuities exist if original pitch curves are directly concatenated to produce a target intonation. The amount of discontinuities depends highly on the selection criteria and also on the prosodic coverage of the corpus. Corpus preparation becomes quite complex when prosodic coverage is of interest. Imposing target intonation computed by an NLP module has its own problems and the natural character of units gets degraded depending on the distance between the actual and target pitch curves.

In this work, we propose the following approach to re-define the target pitch curve after the units have been selected. A cost function is defined as a summation of discontinuities after vertical shifts have been introduced and a penalty value for shifting operation. The penalty value is calculated as the sum of all shifts, scaled with the duration of units. Then this function is minimized for the shift variable to perform maximum discontinuity reductions with minimum shifts. With this operation some of the discontinuities can be removed without degrading the quality of concatenated speech units. The penalty value included in the function guarantees that smaller units are shifted most of the time. When the modification is audible, it is most often rated as an improvement. The amount of improvement in quality, however, depends on the actual pitch curves of selected units.

Informal listening has been performed with small phrases of synthetic speech. The quality of synthetic speech highly depends on the algorithms used to alter prosody and perform concatenation. In our tests, the popular TD-PSOLA algorithm [7] is used as the signal synthesizer, which does not perform any spectral smoothing at concatenation boundaries. Therefore our tests only address pitch discontinuities.

2. MOTIVATIONS

In a non-uniform units-based system, various length audio chunks are selected from a large speech corpus and concatenated to synthesize high quality natural speech. The concatenation algorithms are designed to modify the selected chunks and concatenate them such that the existing discontinuities (like energy, pitch, formant, voice quality discontinuities) at concatenation boundaries are reduced with smallest possible degradation in naturalness of chunks. One of the important issues of this operation is re-definition of intonation curves to reduce pitch discontinuities while keeping quality of audio chunks undistorted.

An important dimension of degradation introduced by alteration of an intonation curve is the distance between the actual and target melodic movements. Therefore, we have designed our smoothing algorithm based on constant pitch shifting on blocks of speech for smoothing pitch discontinuities.
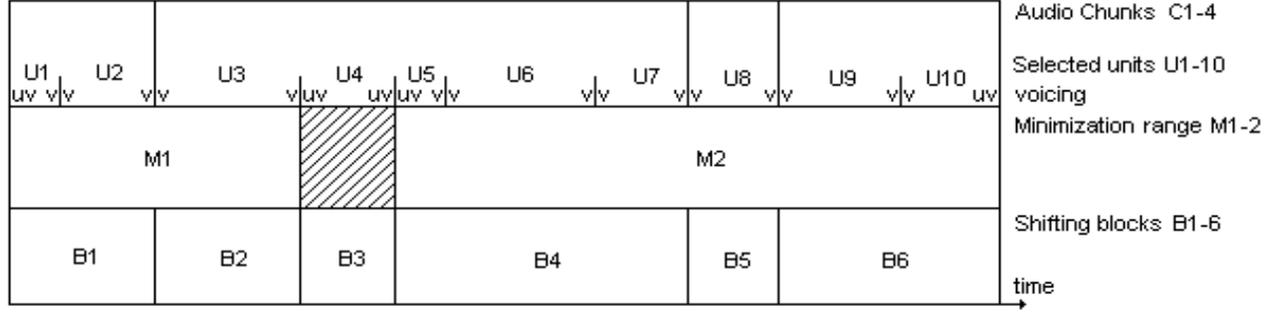
Figure 1: *Representation of sequences of selected units, audio chunks and related minimization ranges and shifting blocks.*

In our shift-only approach, it is advantageous to apply smoothing on small speech segments rather than on whole sentences due to increased computational load and difficulty in controlling dynamics of sentence level intonation movements. Thus, the first step of the smoothing operation is to determine the boundaries of smoothing areas on the sequences of audio chunks. We assume that pitch discontinuities at unvoiced phonemes are of least importance. Therefore, unvoiced unit boundaries are taken as borders for each separate smoothing operation. We call these smoothing areas minimization ranges (Fig1).

Shifting blocks are formed by grouping consecutive units in each minimization range. Then for each minimization range, the vertical pitch shifts to be applied on smoothing blocks are calculated by a minimization operation explained in the next section.

In our system, the voiced/unvoiced labelling of boundary frames is performed on the basis of a V/UV energy ratio computed during MBE complex least squares analysis [Dut96].

During synthesis, no pitch modifications will be applied to unvoiced chunks. Therefore, if some shift values are calculated for unvoiced frames (due to being bounded by voiced frames inside a shifting block), they are treated as dummy values.

3. CALCULATING SHIFTS TO BE INTRODUCED

The following function (H) has been defined as a measure of combined discontinuity and shift penalties for a sequence of K units;

$$H = \sum_{n=1}^{K-1} abs((T0_n^e + s_n) - (T0_{n+1}^b + s_{n+1})) + k * \sum_{n=1}^{K} abs(s_n * d_n)$$

$$d_n = \frac{duration_n}{\sum_{i=1}^{n} duration_i}$$

where $T0_n^e$ is the final pitch period value estimate of the nth unit and $T0_n^b$ is the initial pitch period value estimate, which are both obtained after median filtering few points at boundaries of the unit. s_n is the shift to be applied to nth unit, k is a weighting factor for determining relative importance of shift penalty in the function and d_n is its duration scaled by the total duration of the speech segment.

Minimizing this function is equivalent to simultaneously minimizing each sum, which leads to the following system of equations for the operation:

$$(T0_n^e + s_n) - (T0_{n+1}^b + s_{n+1}) = 0$$
$$n = 1,2,3....K-1$$

$$k * s_n * d_n = 0$$
$$n = 1,2,3...K$$

and the matrix representation of the formulations is as follows:

$$\begin{bmatrix} 1 & -1 & 0 & ... & ... & 0 \\ 0 & 1 & -1 & 0 & ... & 0 \\ ... & ... & ... & ... & ... & ... \\ 0 & ... & ... & 0 & 1 & -1 \\ k*d_1 & 0 & ... & ... & ... & 0 \\ ... & ... & ... & ... & ... & ... \\ 0 & ... & ... & ... & 0 & k*d_K \end{bmatrix} * \begin{bmatrix} s_1 \\ s_2 \\ ... \\ ... \\ ... \\ s_K \end{bmatrix} = \begin{bmatrix} -(T0_1^e - T0_2^b) \\ ... \\ ... \\ -(T0_{K-1}^e - T0_K^b) \\ 0 \\ ... \\ 0 \end{bmatrix}$$

$$A_{(2K-1)*K} * \vec{s}_{K*1} = \vec{d}_{(2K-1)*1}$$

For $K>1$, the number of equations is higher than the number of unknowns and a classical Linear Least Squares solution is appropriate (for $K=1$, the shift is simply zero) :

$$\vec{s} = (A^T * A)^{-1}(A^T * \vec{d})$$

The weighting factor k is manually set by trial and error. This factor serves as a measure of penalization of the shifts to be applied. It can be set according to the post-processing applied after shifting. This is discussed in the next section.

In the figure below, an example of original $T0$ curves of units and new intonation curve obtained by the shifting operation are presented (s1 and s6 are unvoiced chunks).
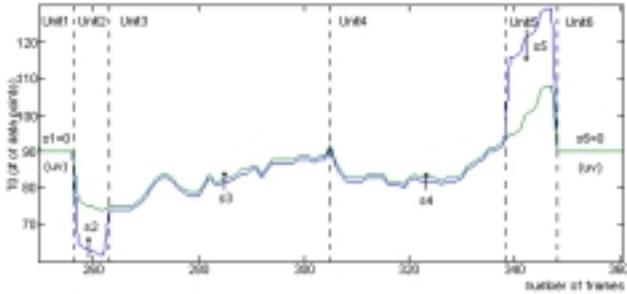
Figure 1: *T0 curve plots before and after shifting operation (shifts are indicated on original curves)*

The shifts calculated depend on the actual T0 curves of the selected units (therefore, on the unit selector and on the speech corpus) and not all of the discontinuities can be removed with the proposed method. The following figure presents an example chunk with a high degree of *T0* discontinuity between units. The shifting operation reduces the discontinuity but some disturbing discontinuities still remain in synthetic speech.
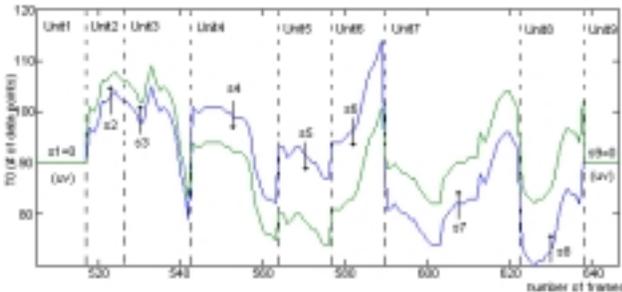

Figure 2: *F0 curve plots for a chunk with severe pitch discontinuities (some of the discontinuities are reduced with the shifting algorithm)*

Clearly, importing this idea in the unit selector itself might still increase the profit of the shifting operation: sequences of units for which F0 shift can be effectively applied should be preferred to sequences for which important discontinuities cannot be removed. This will be subject of further research in our group.

4. POST-PROCESSING

The proposed algorithm removes many discontinuities in the F0 curves, but depending on the selected units, a drawback exists; some unexpected pitch movements can be produced (for example if three units have the same rise characteristics and they are shifted for guaranteeing pitch continuity a long rising intonation may be created). For avoiding this problem the weighting factor k can be increased such that shift penalties are dominant in the cost function. This however will limit not only large shifts but also the small shifts. For this reason, we relaxed the weighting factor to 0.6 (which is manually set by trial and error) and the shifts to be applied are filtered by limiting them within thresholds (set as frequency ratios).

As the next step, smoothing is applied at boundaries where pitch discontinuities still exist after the shifting operation. From many possible smoothing functions, we use linear distribution of pitch discontinuity to left and right units (the region of interpolation is set as one third of each unit). The problem is similar to that of spectral smoothing, therefore other smoothing algorithms may be successfully applied (as the control of formant dynamics in [9]).

5. LISTENING TESTS

Informal listening tests (preference tests) are performed with small phrases of synthetic speech produced with target pitch curves obtained by
i) concatenating actual pitch curves of the selected units,
ii) concatenating shifted pitch curves of the selected units.
The selection algorithm included the following criterions for selection: context, duration and average F0 matching for target cost and F0 continuity for concatenation cost. A French female speech corpus of 60 minutes is used as the speech database.
A number of long sentences were synthesized and phrases where some shifting is applied are chosen as test examples. No post-processing is applied, to be able to judge just the shift algorithm, and the k factor was set to 0.6.
Speech is synthesized with the popular TD-PSOLA algorithm and there was no spectral smoothing involved at concatenation boundaries. The pitch marks are estimated by calculating time locations where the phase of the first harmonic is zero by a harmonic analyzer [14] with the assumption that the high-energy portions of the speech signals will be close to high-energy portions of the first few harmonics and the phase information may provide necessary information for synchronization of overlap add operation [10].
20 listeners (with no experience in listening to text-to-speech synthesis) were asked to report their choices through a web based testing interface (AB test) for 14 pairs of small synthetic speech phrases. They were asked to listen to examples as much as they need to be able to make a choice according to the naturalness of examples and they were not allowed to state equivalence. The overall preference of the shifted-smoothed examples was 75%. Most of the listeners have reported that for some examples the quality differences were obvious and for some examples they could not figure any difference.

6. DISCUSSION

In this paper, a new algorithm for re-defining target pitch curves is presented. The listening tests showed that some of the pitch discontinuities at concatenation boundaries could be reduced without degrading the intrinsic quality of units, thereby making synthetic phrases more natural. The degree of quality improvement depends highly on the actual intonation characteristics of the selected units. Further research will include a joint implementation of the presented algorithm within a unit selection system.
The algorithm could also be used for generating intonation contours for diphone-based synthesis where corpus based methods are used to select intonation chunks and obtain a natural intonation curve by concatenating these chunks [11,12].

7. REFERENCES

[1] M. Balestri, A. Paechiotti, S. Quazza, P. L. Salza, S. Sandri "Choose the best to modify the least: a new generation concatenative synthesis system", *Proc. of EUROSPEECH, Budapest, Hungary, Sept. 1999.*

[2] G. Coorman, J. Fackrell, P. Rutten, B. Van Coile "Segment selection in the L&H Realspeak laboratory TTS system", *Proc. of ICSLP, 2000.*

[3] K. Fujisawa, and N. Campbell "Prosody based unit-selection for Japanese speech synthesis", *Proc. of 3rd ESCA/COCOSDA Workshop on Speech Synthesis, Jenolan Caves, NSW, Australia, Nov. 1998.*

[4] B. Möbius "Corpus-based speech synthesis: methods and challenges" *Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung (Univ. Stuttgart), AIMS 6 (4), 87-116, 2000.*

[5] M. Beutnagel, A. Conkie and J. Schroeter, Y. Stylianou, and A. Syrdal "The AT&T NextGen TTS system", *Proc. of the Joint Meeting of ASA, EAA and DAGA , Berlin, Germany, 1999.*

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

[7] E. Moulines and F.Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones", *Speech Commun., Vol.9, Dec.1990, p 453-497.*

[8] T. Dutoit and B.Gosselin, "On the use of a hybrid harmonic/stochastic model for TTS synthesis-by-concatenation", *Speech Commun., Vol.19, 1996, p 119-143.*

[9] J. Wouters and M.W.Macon, "Control of spectral dynamics in concatenative speech synthesis", *IEEE Trans. on Speech and Audio Proc., Vol.9(1), Jan.2001, p 30-38.*

[10] Y. Stylianou, "Removing phase mismatches in concatenative speech synthesis", *Proc. 3rd ESCA Speech Synthesis Workshop, Nov. 1998, p 267-272.*

[11] T. Saito and M. Sakamoto "Generating F0 contours by statistical manipulation of natural F0 shapes", *Proc.of Eurospeech Scandinavia, 2001, p 1171-1174.*

[12] A.I.C. Monaghan "Extracting microprosodic information from diphones, a simple way to model segmental effects on prosody for synthetic speech", *Proc.of ICSLP, Banff,Canada, Nov. 1992, p 1159-1162.*

[13] A.G. Korn and T.M.Korn, *Mathematical Handbook for Scientists and Engineers*, McGraw-Hill, 1968.

[14] D.W. Griffin, Multi-band excitation vocoder, PhD Dissertation, MIT, 1987.

[15] W.N. Campbell and A.W. Black, Prosody and the selection of source units for concatenative synthesis. In Jan van Santen, Richard W. Sproat, Joseph P.Olive, and Julia Hirschberg, editors, *Progress in Speech Synthesis. Springer, New York, 1997, p 279-292.*