IMPROVING QUALITY OF MBROLA SYNTHESIS FOR NON-UNIFORM UNITS SYNTHESIS

Baris Bozkurt, Thierry Dutoit, Romain Prudon, Christophe D'Alessandro, Vincent Pagel
MULTITEL ASBL,
Initialis Sci. Park, B-7000 Mons, Belgium, {bozkurt,pagel}@multitel.be, dutoit@tcts.fpms.ac.be
LIMSI,
CNRS, Po Box 133 – F91403 Orsay, France, {prudon,cda}@limsi.fr

ABSTRACT

This paper describes the new version of the MBROLA algorithms (called TP-MBROLA standing for True-Period MBROLA)[1] for non-uniform units (NUU) synthesis. The database pre-processing of MBROLA has been modified such that short-time speech frames are not systematically re-synthesized at constant pitch and constant phase envelope. This operation highly reduces the coding-decoding effect on signal quality. For spectral smoothing, only the smoothing frames are re-synthesized at constant pitch and phase envelope and MBROLA smoothing is applied. Furthermore, these operations are performed on the fly, which brings some computational load to synthesis (though it is restricted to smoothing frames). The new version of MBROLA is tested on non-uniform units synthesis by synthesizing speech with units provided by LIMSI's unit selection system [2]. Formal listener tests have shown that TP-MBROLA synthesis quality is preferred compared to MBROLA and raw concatenation synthesis.

1. INTRODUCTION

A system involving high quality selection module and a high quality corpus can produce very high quality synthetic speech even without any signal processing module. However, depending on the material covered by the database, certain discontinuities still exist and there is still a need for signal processing algorithms. The first criterion a signal synthesis algorithm must satisfy to be a candidate for non-uniform unit based synthesis is being able to produce transparent copy synthesis (in other words speech degradation should not be perceivable when the original prosodic characteristics of the speech are not modified). Then the following criterions are: high quality prosody modification (if any), concatenation with smoothing capabilities, and easy voice development (the synthesizer should require little tuning, if any, for a new voice or language).

MBROLA (Multi-band re-synthesis overlap add) is a widely used synthesizer for diphone-based synthesis. For obvious reasons, the MBROLA synthetic speech quality is often graded as highly intelligible but computer like. One important limitation in diphone-based synthesis quality is the database limitation: a diphone database hardly represents all of variability of natural human speech. The association MBROLA-diphones works well (MBROLA currently supports 25 languages), because the quality of the voice produced by the algorithm somehow matches the level of naturalness achievable with a diphone-based synthesizer.

The plain MBROLA algorithm has previously been re-implemented such that it can be used in concatenation of non-uniform units. This was initially the basis of the NUMBROLA project [3], aimed at extending MBROLA to NUU voices through collaborations. Further tests showed that MBROLA, even when used in the context of NUU-based synthesis, fails to satisfy the first and most important criterion stated above. Copy synthesis is not transparent, especially for female speakers. The main reason for this degradation is that the short time speech frames are re-synthesized (with a harmonic model) at a constant f0 and phase envelope prior to synthesis, which is needed for time domain smoothing at concatenation points. The amount of degradation introduced by this modification is highly dependent on the actual f0 curve of the original recordings. Non-uniform unit based synthesis aimed speech corpora are far away from having constant f0 since a good coverage of general prosodic features of natural speech is needed (which is not the case for a diphone database).

To improve the quality of synthesis, database pre-processing in the NUMBROLA project has recently been modified such that the harmonic re-synthesizer re-synthesizes frames with their original pitch and phase envelope. This change improves the quality of copy-synthesis to a degree very close to transparency. The synthesizer is also modified to handle the new data (especially spectral smoothing). This new algorithm, named as TP-MBROLA (True Period MBROLA), will be shared within the NUMBROLA project.

Formal preference tests have been performed in the context of NUU-based synthesis for comparing the quality

of TP-MBROLA to that of MBROLA (in its NUU-based version), as well as to raw concatenation synthesis. Units provided by LIMSI's unit selection system [2] are concatenated to synthesize speech examples for the tests.

2. TP-MBROLA SYNTHESIS

TP-MBROLA is a new version of the MBROLA algorithm, where most important modifications are in database pre-processing (the new database pre-processing procedure is presented in figure 1.). MBROLA database pre-processing was specially designed for diphone databases, in which the recorded is speech is close to flat pitch for most of the database units, therefore fixed frame size analysis is appropriate. However speech corpora designed for non-uniform units synthesis are quite colorful in terms of pitch. Therefore, pitch asynchronous fixed frame size analysis is problematic. For this reason, the database pre-processing of TP-MBROLA is made synchronous with first harmonic's phase for each pitch period, which will also synchronize OLA operation during synthesis (as suggested by [5] in the context of removing linear phase mismatches between concatenated speech frames). This synchronization is achieved by calculating time locations where the phase of the first harmonic is zero using the MBE analyzer [4]. This assumes that the high-energy portions of the speech signals are close to high-energy portions of the first harmonics and that the phase information may provide necessary information for synchronization of overlap add operation. Then, the harmonic analyzer is used once more to estimate harmonic amplitudes and phases (this time pitch synchronously by centering analysis frames at these pitch synchronous time locations).

The voiced/unvoiced labeling of frames is performed on the basis of a V/UV energy ratio computed during MBE complex least squares analysis [6] and dummy pitch marks are placed at a constant rate on unvoiced frames.

In the original MBROLA database pre-processing, short-time speech frames were re-synthesized at a constant pitch by re-sampling the amplitude spectrum and imposing constant phase envelope for all frames. This is known to create some buzziness in synthetic speech. The amount of buzziness introduced with this operation is dependent on the amount of prosody modification that will be imposed during synthesis (especially duration stretching) and also on the speaker (not all of the MBROLA databases have the same buzziness). For NUU synthesis using TP-MBROLA, a second important modification is made on database pre-processing: short-time frames are re-synthesized at their original pitch values and phase envelopes. Pitch values are rounded to integer period values (data points). With this property, the frames can be looped as much as needed without discontinuity.

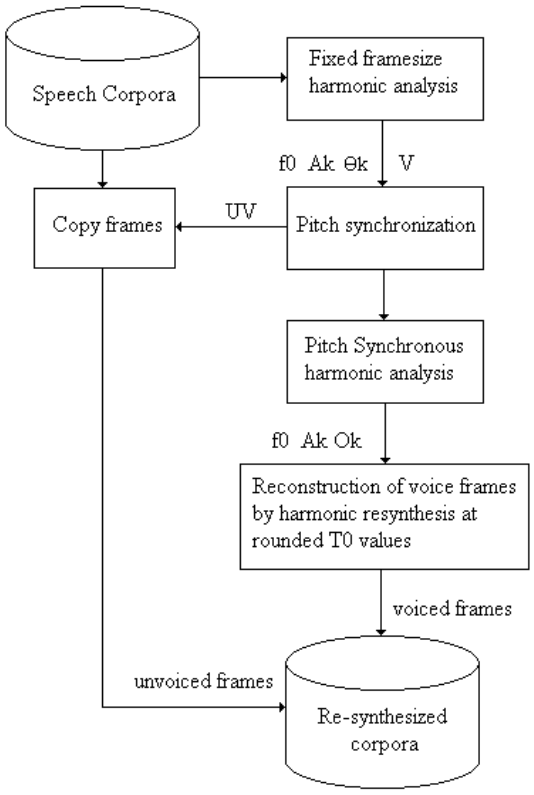

Figure 1: *TP-MBROLA database pre-processing*

In TP-MBROLA (as in MBROLA), synthesis of speech signal and prosody modification is achieved with a simple OLA algorithm where short time speech frames (OLA frames) are copied/deleted and/or shifted to impose the target prosody. Each OLA frame is obtained by copying the same re-synthetic frame (one period) twice and windowing the resultant two period signal with a hanning window. Due to harmonicity of the re-synthetisized frames, no discontinuity is introduced by copying. Spectral smoothing is applied at voiced concatenation boundaries. The smoothing frames are re-synthesized at their average pitch value with the phase envelope of the left most smoothing frame of the left unit. Then the MBROLA time-domain smoothing is applied, which is a fade in/out distribution of the difference between boundary frames to right and left smoothing frames.

4. SMOOTHING F0 DISCONTINUITIES

Though the unit selection module of a non-uniform units synthesis system tries to guarantee F0 continuity at concatenation boundaries by various ways (i.e. including actual pitch values in target cost and pitch discontinuities in concatenation cost functions), 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 and pitch continuity in synthesis are of interest. Imposing target intonation computed by an NLP module solves this coverage problem but it brings in other difficulties as the natural character of units gets degraded depending on the distance between the actual and target pitch curves.

In concatenative speech synthesis, the amount of degradation in synthetic speech quality is highly correlated with the amount of prosody modification applied on speech signals. For getting the best synthesis quality with TP-MBROLA, we thus prefer to preserve the original intonation curves and to apply some pitch smoothing only when needed. Therefore we have integrated our shift-only pitch smoothing algorithm.

For reducing pitch discontinuities without much degradation of signal quality, we have previously proposed a shift-only smoothing algorithm [7], which has been integrated with TP-MBROLA to form the signal synthesis block of the NUU synthesis system. By this operation, smoothing is applied while preserving micro-prosody. In many examples, pitch smoothing is observed to be more effective in improving the overall quality of synthetic speech than spectral smoothing at concatenation boundaries. The algorithm estimates optimum uniform pitch shifts to be applied on units by cost minimization. A cost function is defined as a summation of discontinuities after vertical shifts have been introduced and a penalty value is added for preventing too many shifting operations. 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. An example is presented at figure 2.

5. EVALUATION TEST

It seemed necessary to check the quality of TP-MBROLA within the context of a selection-concatenation TTS system. A formal pair preference test has thus been designed and conducted.

5.1. Selection-concatenation TTS system

Both the MBROLA and TP-MBROLA algorithms have been integrated to the LIMSI Selection/Concatenation TTS system, a corpus-based system for the French language described in [2]. This TTS system is based on the optimal selection and concatenation of natural speech segments found in large databases (like in [8]), without any synthesis rules either at the segmental or suprasegmental levels. The reader is referred to [2] for details on the original system: only the signal processing features will be discussed here.

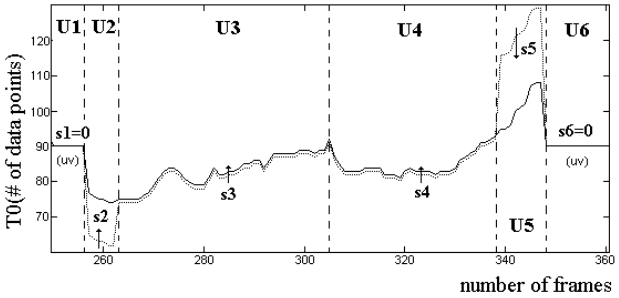

Figure 2: *T0 curve plots before and after shifting operation (shifts are indicated on original curves)*

The system does make use of only minor prosodic modification procedures. The best fitted prosodic units (in terms of pitch and durations) are searched for in the database, and they are concatenated as such, without pitch modifications. Too short durations and too high pitches are pruned from the data, on the basis of average values of pitch and durations in the base (about 7% of the data are pruned). Duration modifications are obtained within the limits of the actual segments, by varying the proportion of signal actually used. The segments boundaries are known, and it is then possible to vary the percentage of signal used (typically 48% of the segments on each side of a diphone boundary are used). There is no energy modification. The segments are then concatenated using a simple splicing procedure for smoothing the signal at concatenation points (10 ms Hamming window smoothing). The databases employed in the present test last about 2 hours for each speaker. The speech signal is enriched with linguistic, prosodic and acoustic information, as described in [2].

The speech quality obtained has been checked using a subjective overall quality test and a comparison test with LIMSI's diphone-based TTS system. The selection-concatenation system demonstrated a significant improvement in voice pleasantness.

5.2. Test Procedure

A male and a female speaker have been chosen for the present test. Synthetic stimuli have been computed, based on 10 sentences (from the newspaper "Le Monde") ranging between 60 and 120 phones (i.e. 3s and 10s). The "raw" condition corresponds to the original synthetic utterances produced by the TTS system. The "Mb" (resp. "TpMb") condition corresponds to the same utterances, but processed by MBROLA (resp. TP-MBROLA) for

signal smoothing and concatenation. Therefore 60 synthetic stimuli were available for testing.

Monophonic stimuli were played through headphones in a sound insulated booth according to a paired preference test paradigm. 21 subjects, all native speakers of the French language participated in the experiments. (8 from France and 13 from Belgium). The test was divided into 4 parts: raw vs TpMb (male), raw vs TpMb (female), Mb vs TpMb (male) and Mb vs TpMb (female). It lasted between 15 and 20 mn on average. The synthetics speech examples used in testing can be found at http://www.tcts.fpms.ac.be/synthesis/numbrola/tests/LIMSI/examples_LIMSI.htm

5.3. Test results

The test results are reported in Figure 3. The TpMb stimuli have been preferred to the raw data for more than 80% of the stimuli for the female speaker (about 80% for the male speaker). All these results are highly significant. All the subjects preferred the TP-MBROLA version for all the sentences on average. The results for the female speaker are slightly higher for TpMb. Therefore the TpMb smoothing procedure is highly effective for improving the speech quality of the selection-concatenation system.

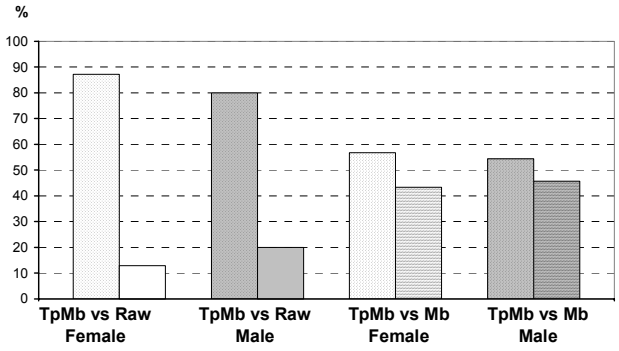

Figure 3: *Results of the preference test (% of preference for each pair, for all subjects and sentences).*

The TpMb procedure rated also better than the Mb procedure for both speakers. The difference is significant for the female speaker (z=1.96, confidence level of 0.05), and weakly significant for the male speaker (z=1.25, confidence level of 0.15). The results depend a lot on the sentence examined, especially on the frequency and degree of discontinuities. Many subjects reported that these parts of the tests were rather difficult, and that in many cases the stimuli seemed almost the same, But for some sentences (where units are rather continuous) TpMb is clearly preferred, although for some other sentences(where discontinuities are more frequent) Mb is clearly preferred. It seems that both methods have different sound quality and also different robustness to pitch and concatenation problems.

6. DISCUSSIONS AND CONCLUSION

The TP-MBROLA algorithm has been tested in the context of large database selection synthesis. TP-MBROLA copy-synthesis quality is close to transparency, which is a remarkable advantage especially for NUU synthesis.

TP-MBROLA synthetic speech is preferred to raw concatenation synthesis and the results are highly significant. This is mainly due to phase synchronous concatenation and f0 smoothing using a shift-only smoothing algorithm.

TP-MBROLA is also preferred to MBROLA for synthesis but the results are only just above statistical significance. Preference seems also dependant on the voice type (better for the female voice than for the male voice) and on the type of synthesis discontinuities in the sentences. MBROLA synthetic speech sounds more homogenous and uniform, and is therefore preferred for sentences where many discontinuities exist. For unit sequences with fewer discontinuities TP-MBROLA outperforms MBROLA due to its improved transparency in copy-synthesis.

7. REFERENCES

[1] Dutoit, T. and H.Leich, "Text-to-speech synthesis based on a MBE re-synthesis of segments database", *Speech Commun.,* Vol.13, p 435-440,1993.

[2] Prudon, R. and C. d'Alessandro. "A selection/concatenation TTS synthesis system : Databases development, system design, comparative evaluation. " *Proc. 4th ISCA Speech Synthesis Workshop*, Pitlochry, Scotland, pp. 137-142, 2001.

[3] Bozkurt, B., M. Bagein, and T., Dutoit, " From MBROLA to NU-MBROLA ", *Proc. 4th ISCA Speech Synthesis Workshop*, Pitlochry, Scotland, pp. 127-129, 2001.

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

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

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

[7] Bozkurt, B., T., Dutoit, and V., Pagel , " Re-defining intonation from selected units for non-uniform units based speech synthesis ", *Proc. SPS-2002 IEEE Benelux Signal Processing Symposium*, Leuven, p 141-144, March 2002.

[8] Beutnagel, M., A. Conkie, J. Schroeter, Y. Stylianou, and A. Syrdal. "The AT&T next-gen TTS system", *Proc. Joint Mtg. ASA*, EAA and DEGA, Berlin, 1999.

[9] Beutnagel, M., A. Conkie, and A. Syrdal. "Diphone synthesis using Unit Selection", *Proc. 3rd ESCA/COCOSDA Workshop on Speech Synthesis*, Jenolan Caves, Australia, Nov 1998.