ISCA Archive
http://www.isca-speech.org/archive

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

IMPROVING QUALITY IN A SPEECH SYNTHESIZER BASED ON THE MBROLA ALGORITHM

B. Etxebarria, I. Hernáez, I. Madariaga, E. Navas, J.C. Rodríguez, R. Gándara
University of the Basque Country
ETSII/IT, Alda. Urquijo s/n - E48013 Bilbao - Spain
borja@bips.bi.ehu.es
http://bips.bi.ehu.es

ABSTRACT

Speech synthesis based on the Multiband Resynthesis OverLap-Add (MBROLA) algorithm produces high quality speech without requiring too much effort to design the diphone database, and using a low computational power. The main drawback of this algorithm is the slightly metallic sound or buzziness that can be perceived on voiced segments. We are working on a speech synthesizer based on the MBROLA algorithm, and trying to improve its speech quality by means of on an enhanced phase control strategy.

Keywords: MBE, MBROLA, speech synthesis.

1. INTRODUCTION

MBROLA speech synthesis uses the PSOLA algorithm [2, 4], applied over a pre-processed speech segments database. This database is obtained by re-synthesizing a natural speech diphone database: first natural speech is coded using a Multiband Excitation (MBE) model[1], and then decoded with certain modification rules to produce the database used by the PSOLA algorithm [2]. The algorithm is applied pitch synchronously using completely automatic pitch mark generation (pitch marks can be placed anywhere inside the pitch period, so no glotal closure analysis is needed). This re-synthesis algorithm uses a fixed pitch value to avoid pitch mismatches in the PSOLA synthesis stage. It also avoids phase mismatches by using a fixed phase relation between harmonics in every pitch synchronous frame. This process is applied only over voiced frames. The original spectral envelope in each segment is preserved, so envelope mismatches must be corrected at synthesis time; this can be easily performed by direct time-interpolation between frames, due to the fixed phase relation imposed to the harmonics [2]. These strategies for pitch, phase, and envelope continuity reduces the time employed on the design of the speech units (diphones) database. Moreover the synthesis stage is extremely efficient, reducing to a simple OLA algorithm.

The drawback of using a fixed phase relation between

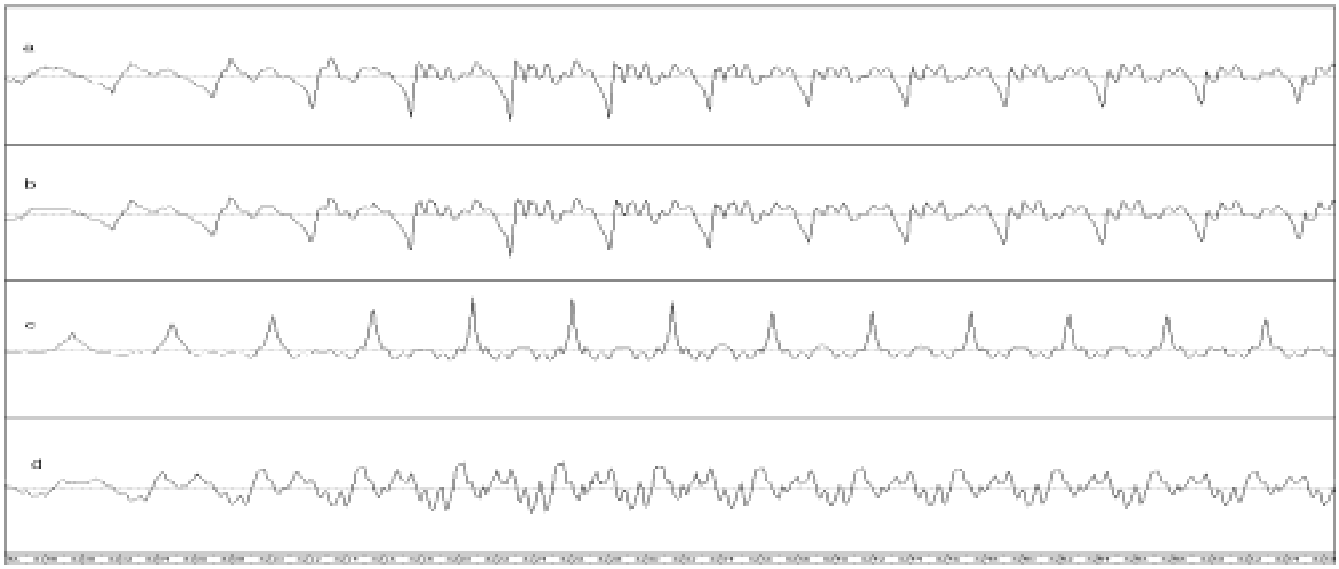

Figure 1: Different phase reset strategies in MBE-PSOLA transcoding of a voiced segment. **(a)** Original waveform. **(b)** MBE transcoded waveform, preserving original phases in every pitch syncronous frame. **(c)** MBE transcoded using a fixed zero phase for every harmonic in each frame. **(d)** MBE transcoded, with a fixed random phase set in every frame.

10.21437/Eurospeech.1999-584

harmonics is that it adds an undesired distortion to the transcoded MBE signals, perceived as a slightly metallic sound or buzziness on voiced segments. The distortion can be observed in figure 1. To avoid side effects other than those produced by phase shifts, no pitch or time scale modification has been applied to the signal in this example. Figure 1b presents the MBE-PSOLA transcoded signal preserving the original phase relations; there is almost no noticeable differences between transcoded and original signal (fig. 1a). If every harmonic's phase is reset to zero (or to a linear phase relation) in each pitch period, the transcoded signal shows a dominant metallic sound, being the waveform extremely artificial (fig. 1c). When the phases are reset to a fixed randomly selected phase set in each pitch period, we get a much more natural sound and the waveform is not so artificial (fig. 1d), but buzziness is still noticeable.

As phase distortion is the main reason for the buzzy sound on MBROLA, we propose the use of an enhanced phase control model in order to improve the quality of the synthesized speech without increasing the synthesis stage computational requirements and preserving the database design simplicity.

2. PHASE CONTROL MODEL

Our approach is based on the following facts:

- To avoid buzziness, the original phases in each pitch period should be preserved. This conflicts with the fixed-phase requirements needed during the OLA synthesis stage to avoid phase mismatches in the boundaries of different speech segments (diphones).

- In a diphone database, a segment that is right-limited by a given phone, needs matching just with those other segments from the database that are left-limited by the same phone.

- Actually, the fixed-phase relation only needs to be asserted on segment boundaries to keep on using OLA with no phase-mismatch distortions. Assuming that small phase shifts between harmonics from one pitch period to another one are harmless, we can evolve from different phase sets on left and right segment boundaries.

- We assume that the phase relation between harmonics in a given phone at a boundary (initial or final stable point of a diphone or polyphone) follows a certain pattern for all the instances of that phone in the database.

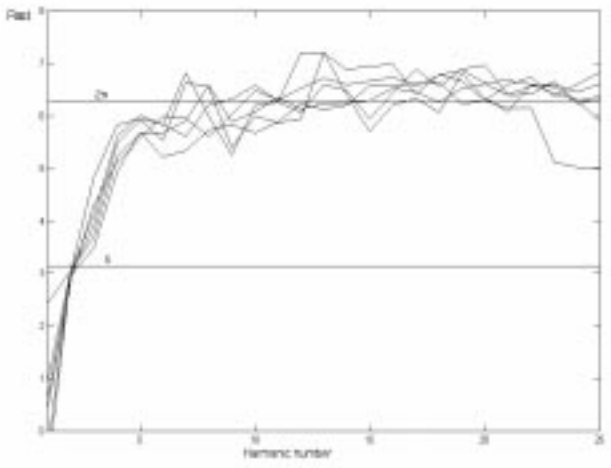

Figure 2: Unwrapped phase relation for several instances of a given vowel.

Based on the previous facts, instead of using the same fixed phase relation between harmonics in the whole database, we use a different phase relationship for each kind of boundary sound. In this way, we try to reduce phase distortion on the database, preserving the advantages of MBROLA synthesis.

For every database segment where a given phone appears as a boundary (left or right), we perform a MBE analysis to get the phase relationship between harmonics at the boundary point. All the phase responses computed this way are combined to obtain an averaged phase response. As shown on figure 2, the spectral phase characteristics in the stationary part of a phone (where the diphone boundary is usually defined) is somehow similar in most of the instances of that phone in the database. This does not happen in the segments where the phone is heavily influenced by adjacent sounds, but usually these points are not selected as segment boundaries, using triphones instead. Experimentally we have checked that the mean value obtained by averaging the individual phase responses is a good choice to represent them all.

When performing phase averaging we must bear in mind that each phase relation instance can be computed on an arbitrary position inside a pitch period (at a different time delay or phase reference), and also that we actually manage wrapped phase values $[-\pi,\pi]$. In order to have a common phase reference, we always subtract a linear phase so that the fundamental harmonic resets to a zero phase. To avoid performing a phase unwrapping algorithm, the average is directly carried out harmonic by harmonic using the wrapped phases, and vector-averaging them in the circular z-plane domain: for the N instances of a given segment boundary, the averaged phase of the k-th harmonic is obtained from:

In the few cases in which A is below a small threshold, there is too much phase dispersion to select a representative averaged phase and we choose zero phase for that harmonic.

During the re-synthesis stage of the database, we use the estimated mean phase relation for every boundary instance of a given phone. As a database segment will generally begin and finish with different phones, it will be re-synthesized with a phase in its initial frame, and a different one in the final frame. All the phases of the pitch synchronous frames between these two boundary frames will evolve smoothly from one to the other: the phase of an harmonic is linearly interpolated between the initial and the final one, clockwise or anticlockwise to consider the effect of phase wrapping, following this way a minimum phase variation criterion.

In plosives and noise-like diphone boundaries, abrupt phase changes are harmless, so in these units no linear interpolation is needed and the phase relations of the voiced boundary are kept fixed all along the unit.

One of the advantages of MBROLA is that the spectral envelope mismatches can be easily corrected by simple interpolation in the time domain. This is possible only if every frame where we apply the interpolation process uses the same fixed pitch and phase relation between harmonics. As described above, our approach does not have a fixed phase relation inside a segment, as the phase has to evolve from the initial phase relation to the final one. The problem can be solved: as interpolation is necessary only in the first/last few frames of every segment (the number of frames depends on the phone), we keep the initial phase relation fixed during the first few frames, and in a similar way in the last few frames. All the other frames between these, evolve as explained above.

3. EXPERIMENTAL RESULTS

Informal audio tests have shown that the proposed algorithm for phase control produces higher quality than the random phase approach.

An example of the performance of the proposed algorithm is shown in figure 3. Figure 3a shows the original waveform and spectrum for a voiced segment 'o-n-a'. From the whole diphone/polyphone database we obtained an averaged phase relation for the boundary point of both 'o' and 'a' phonemes. Then the 'o-n-a' segment was resynthesized using these phase relations for initial and final MBE frames.

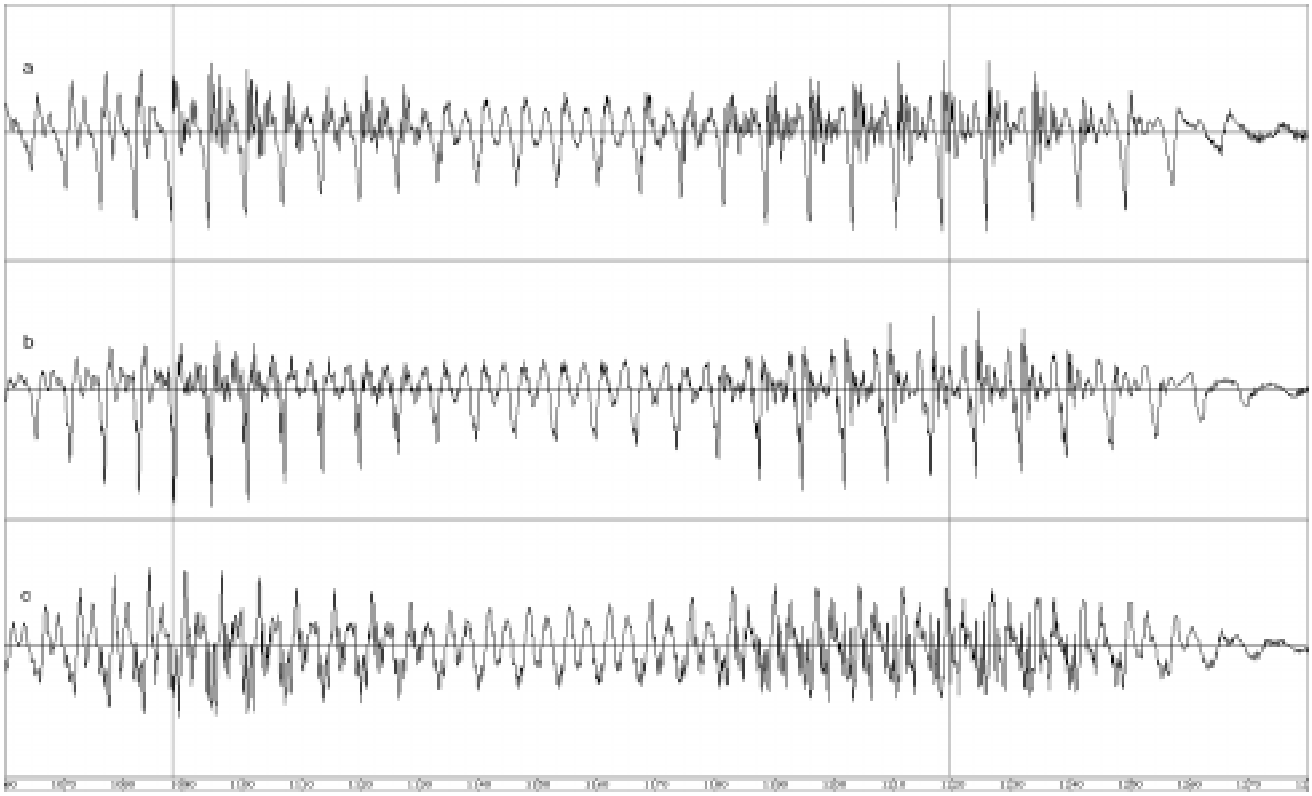

Figure 3: (a) Original waveform for triphone o-n-a (between the boundary lines). **(b)** MBE transcoded waveform, using the averaged phase sets for 'o' and 'a' at the boundary points. **(c)** MBE encoded waveform using a fixed random phase set for every pitch syncronous frame.

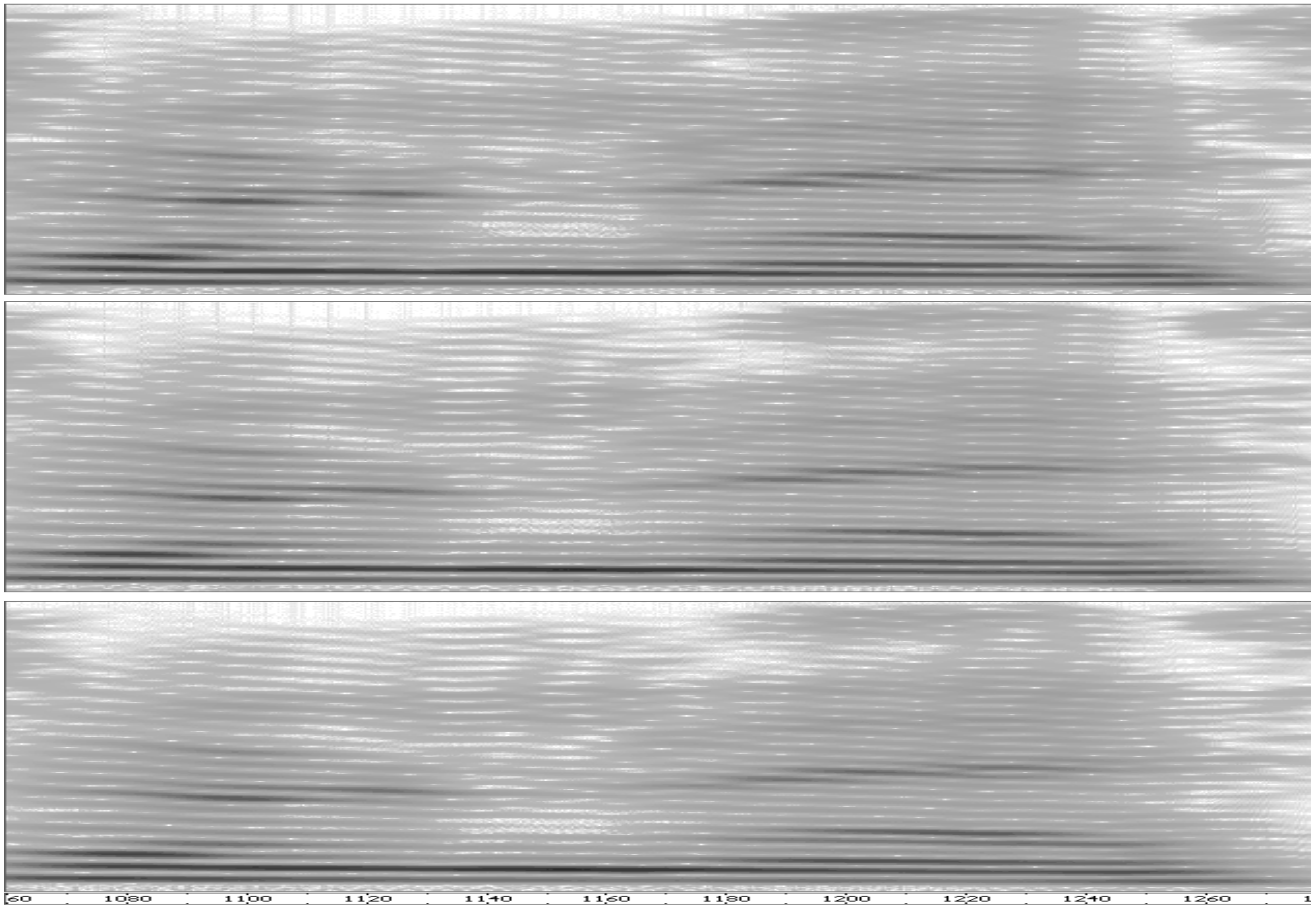

Figure 3 (continued): spectrogram (up to 4kHz) for the three signals of fig. 3a, 3b and 3c.

The result is shown in figure 3b, where it can be seen that the waveform shape is almost preserved near the left and right boundaries of the segment. For comparison purposes, on figure 3c the same segment resynthesized using a fixed random set of phases can be seen.

The fact that the waveform shape in fig. 3b is so similar to the original one proves that the phase averaging approach is not bad despite of all the assumptions we have made. Although in this example (fig. 3b) the waveform shape on the central area of the segment looks quite similar to the original waveform, this is not so in general as the phase relationships have nothing to do with the original phases, being just a linearly interpolated set between the initial set for 'o' and the final one for 'a' .

4. FUTURE WORKS

Although the quality in the resynthesis has been improved, there is still some residual audible buzziness. The two key points of our algorithm are the phase averaging and phase interpolation. Our current approach based on direct harmonic by harmonic phase averaging assumes that the pitch has very small variations in the frame instances to average. We are studying another approach based on frequency by frequency averaging that needs a robust phase unwrapping algorithm that should allow higher pitch variance. Other algorithms for phase interpolation are going to be tested in order to improve the waveform shape preservation on the central part of diphones/polyphones.

5. REFERENCES

[1] D. Griffin, J. Lim, “Multiband Excitation Vocoder”, IEEE Trans. on Acoustics, Speech and Signal Processing, Vol. 36, n.8, Aug. 1988

[2] T. Dutoit, H. Leich, "MBR-PSOLA: Text to Speech synthesis based on a MBE re-synthesis of the segments database", Speech Communication 13, 1993

[3] T. Dutoit, H. Leich, "A comparison of Four candidate Algorithms in the context of High Quality Text to Speech Synthesis". ICASSP'94

[4] E. Moulines, J. Laroche, “Non-parametric Techniques for pitch-scale and time-scale modification of Speech”, Speech Communication 16 (1995) 175-205, Elsevier.