Improved phase control strategies for MBE-PSOLA re-synthesis of TTS diphone databases.

B. Etxebarria, R. Gándara, J.C. Rodríguez, I. Hernáez
University of the Basque Country
ETSII/IT, Alda. Urquijo s/n E48013 Bilbao (Spain)
Tfn: +34 94 601 41 24
Fax: +34 94 601 4259
borja@bips.bi.ehu.es

Research financed by the University of the Basque Country. Ref. UPV 147.345-TA066/98

ABSTRACT

MBE re-synthesis has shown to be an invaluable tool to improve quality of PSOLA-alike synthesis systems, and simplify the diphone database designs, as the MBROLA algorithm has proved. This algorithm is based on a MBE-PSOLA pre-processing of the speech database; in this re-synthesis stage, tight phase and pitch control allows to generate a PSOLA segments database that will allow smooth segment concatenation and high quality speech. The downside is background buzziness on voiced segments of the generated speech. In this paper we will expose our approach to reduce this buzziness by means of improved phase control during the MBE re-synthesis of the database. This way, the real-time synthesis stage needs no additional computational power, though quality is noticeably improved.

INTRODUCTION

MBROLA [2,3] uses the PSOLA [4] algorithm with a pre-processed diphone/poliphone database. This database is obtained by re-synthesizing a natural speech diphone database: first the natural speech is coded using a Multi-Band Excitation (MBE) model [1], and then decoded, after applying some modifications [2,3], to produce the database used by the PSOLA algorithm [4]. The algorithm is applied pitch synchronously (pitch marks are automatically generated). 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. Spectral envelope mismatches can also be easily corrected by direct time interpolation between frames, because of the fixed phase relation imposed to the harmonics. The synthesis stage is this way extremely efficient, but a slightly metallic sound or buzziness can be perceived.

The buzziness present in MBROLA is mainly due to the phase constraints imposed to the harmonic spectrum during the MBE re-synthesis stage: to assert phase continuity on concatenated segments, all the database is re-synthesised using the same phase relation between harmonics. This phase relation is randomly selected: zero phase relation produces extremely metallic sound, while completely random produces more natural sounds, although waveform shape is completely changed and background buzziness is still perceived, even if we preserve the original phase on high frequencies [2].

New approaches to synthesis try to improve quality by using different coding strategies, but at the expense of increased computational complexity. Because of this we are trying to improve the quality of an MBROLA based synthesiser without increasing the computational complexity during the synthesis stage.

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. It should be possible to allow small phase shifts on the segment internal frames without noticeably increasing distortion.

- 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. 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 or bigger units instead.

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 for that phone. During the re-synthesis stage of the database, we will use this mean phase relation for every boundary instance of that phone. As a database segment will start and end with different phones, it will be re-synthesised with a phase in it's initial frame, and a different phase in the final frame. All the phases of the pitch synchronous frames between these two boundary frames should evolve smoothly from one to the other following a minimum phase interpolation 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 can be 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 all the frames were we apply the interpolation process use the same fixed 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 overcomed: as interpolation is necessary only in the first/last few frames of every segment (the number of frames depends on the phone), we keep fixed the initial phase relation during the first few frames, and similar in the last few frames. All the other frames between these should evolve from one set to the other one.

The two key points of our algorithm are the phase averaging and phase interpolation, as described now:

PHASE AVERAGING

We have checked two approaches to obtain a representative phase relation between harmonics for a given sound.

The first one is based on averaging phase values harmonic-by-harmonic [5]. This way we don't need performing a phase unwrapping algorithm, as the average is directly carried out using the wrapped phases, and vector-averaging them in the circular z-plane domain. All the frames to be averaged must have the same time reference inside the pitch period; as a rough approximation, we apply a linear delay to the phase response, so the phase of the first harmonic is zero.

The previous approach is valid only when there are small pitch deviations in all the MBE frames to be phase-averaged. Because the vocal tract response is not coupled with the excitation, if pitch values are too different, harmonic averaging actually mixes phases at very different frequencies, notably for high frequency harmonics.

To overcome this, we can use an LPC model to strip-out the vocal tract response, performing the phase average over the MBE harmonics of the residual signal. We have found that the phase response of the residual signal also follows a pattern that is independent of the pitch period, as will be shown later. So, although in a lesser extent, the problem persists.

Our second approach to phase averaging is based on frequency-by-frequency averaging and copes with the problems of the previous approach. To perform frequency-by-frequency averaging, a continuous phase response is needed, so a phase unwrapping algorithm must be applied. We perform phase unwrapping over the MBE phase response of the LPC residual, and then add to it the LPC envelope phase response, obtaining this way the complete signal unwrapped phase response. The LPC filter is a minimum phase system that simply adds a modulation over the residual phase response increasing its phase dynamic range (phase dispersion) but preserves the mean phase response of the residual. Because of this it is safer to perform phase unwrapping over the residual signal, to simplify the phase unwrapping problem and to avoid as much as possible incorrect phase wrap-around detection. The LPC filter phase component is later easily added without phase wrapping ambiguities, as we have an analytic expression for the LPC polynomial.

The LPC residual is similar to a periodic delta train. So, taking the time reference on a delta maximum, the phase of the MBE analysed frame will be nearly zero for most of the harmonics. Actually, as the MBE analysis can be performed on any point inside the pitch period, the residual phase will present a linear phase. Our algorithm search for a delay d that minimises the phase variance: given the phases of the N harmonics on a MBE frame,

$$\{\phi_k,\ k=1,...N\}$$

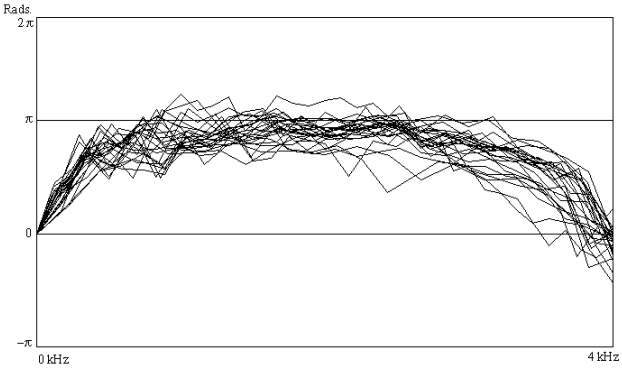

Fig. 1: MBE frames phase response, for the LPC residual of several different voiced sounds, after phase minimisation algorithm.

We apply to it a linear delay to obtain:

$$\phi^d_k = R^{2\pi}[\ \phi_k + 2\cdot\pi\cdot k\cdot d\ /\ P\ ]$$

Where P is the pitch period, and $R^{2\pi}[\phi]$ represents the principal value $(-\pi, \pi)$ of a phase value ϕ.

We compute the phase variance as:

$$P^d = \sum_{k=1}^{N} \left(\phi_k^d\right)^2$$

and search for the delay d that minimises this value, scanning the best one in the range $-P/2$ to $+P/2$ using small increments. This way we don't perform any explicit wraparound check.

Actually, for real-life LPC residual signals we have found that the phase response of the MBE harmonic components is not too linear-alike, but it presents a smooth deviation from the linear component. We have found that this deviation is independent of the pitch period, and very similar for all the frames composing the diphone database for a given speaker. Fig. 1 shows residual phase for several voiced MBE frames randomly selected from recordings of a given speaker, with extreme range for the pitch values (100 to 250 Hz). In these phase curves the linear delay has been removed using the previously described algorithm for phase minimisation, with a modification to consider the smooth deviation from zero phase: the phase variance to minimise is:

$$P_\varphi^{\ d} = \sum_{k=1}^{N} \left(R^{2\pi}\left[\phi_k^d - \varphi_k\right]\right)^2$$

Being φ_k. the phase response of an $H^{a,b}(z)$ system evaluated at the frequency of harmonic k:

$$H^{a,b}(z) = \frac{(z-a)^2}{(z-b)^2}$$

Where a and b are two real numbers in the range [-1,1] that allow us to model the shape of the smooth deviation. To minimise $P_\varphi^{\ d}$, these two constants are scanned in small steps in the range [-0.99, +0.99] for every delay d. As the phase deviation is similar in all the frames of the database, we don't need to perform a full scan for a and b for every frame to be phase-minimised. This reduces considerably the time to process the whole database.

After performing the phase minimisation over the harmonics of the residual signal, the phase response of the vocal tract (LPC filter) is added to obtain the complete phase response, that goes on preserving the minimum delay characteristic. Once that every frame to

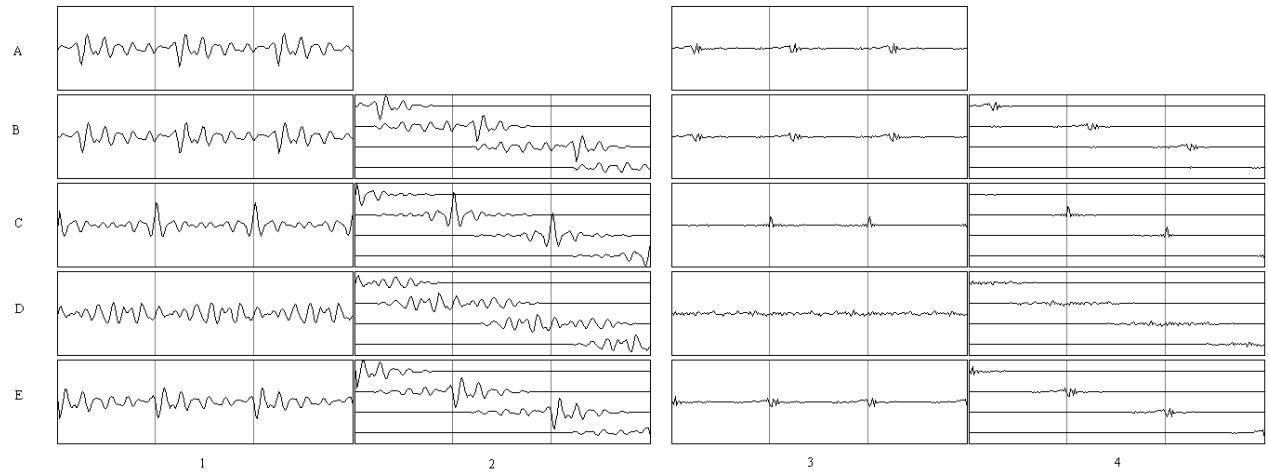

Fig. 2 Some MBE resynthesis examples (see text)

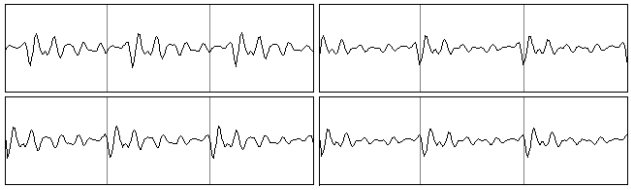

Fig. 3 Two different instances (left/right) for the same sound. Top: original signals. Bottom: MBE resynthesized using minimum phase.

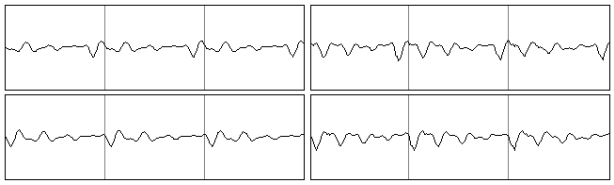

Fig. 4 Two different instances (left/right) for the same sound. Top: original signals. Bottom: MBE resynthesized using minimum phase.

be phase-averaged is normalised this way, phase at frequencies different to those of the MBE harmonics can be computed by direct interpolation between harmonic phases. We then perform the frequency-by-frequency phase average by averaging these interpolated phase responses at every frequency, using the MBE amplitude response of each frame as a weighting function.

PHASE INTERPOLATION

To evolve from an initial phase set to a final one along the internal MBE pitch-synchronous frames that form a diphone we have checked two approaches.

The first one uses linear interpolation between the two boundary sets, bearing in mind phase wraparound to perform minimum phase variations. This way, the phase relation on internal frames is just an intermediate value between the initial and the final frame set, and it has no relation with that frame's original phase set. Due to this, waveform shape is usually not preserved in those areas, and although using this algorithm we obtain smooth transitions and good speech quality, buzziness is still slightly noticeable in some segments.

Our second approach is based on differential phase responses, and produces better results. First of all, all the frames on the database are linearly shifted to present the minimum phase response, as exposed previously. This way, all the frames in the database follow a similar time-domain criterion for placing the pitch marks, although preserving the original phase relation between harmonics. Then, to obtain the average phase responses in the boundaries of a diphone, we compute the difference between the actual phase response and the averaged one on both boundaries. These differential responses are added to the respective boundary, obtaining this way the average phase. All the frames between the boundaries are corrected using a differential phase response obtained by interpolating between the two boundary differential responses. As both the boundary averaged phase values as the original phase values follow the minimum phase criterion, the differential phase is generally small, and the waveform shape is better preserved all along the diphone.

RESULTS

To illustrate how the phase minimisation and phase interpolation algorithm works, we present some graphical results.

Fig. 2 presents an MBE re-synthesis for a short voiced speech segment. Column 1 presents the waveform and column 2 shows the corresponding OLA units. Columns 3 and 4 present in the same way the MBE re-synthesis process for the same segment, but applied to the LPC residual signal, where phase modification effects can be better perceived. The time position where the MBE re-synthesis has been carried out is marked with vertical lines (pitch marks). For all the four columns, the graphs on the top line (A) present the original waveform. Line B is the MBE re-synthesised waveform without applying any changes, which is almost identical to the original signal, with almost perfect sound quality. Lines C and D are shown for comparison purposes, and present respectively the MBE re-synthesised speech resetting to zero the phase of the MBE harmonics (line C) and resetting them to a fixed randomly selected phase value (line D). Finally, on line E the MBE re-synthesis has been carried out applying the phase minimisation algorithm described in this paper; as can be seen on the residual signal, phase minimisation has the effect of bringing the excitation maxims near the centre of the synthesis frame. As a result, the speech waveform also presents maxims near that point. This effect is better perceived on fig. 3, where two segments of speech, containing the same voiced sound but from different contexts are compared: despite that both original signals use different pitch marks time references, the MBE re-synthesised segments, present similar pitch marks as a result of applying the minimum phase criterion. Fig. 4 is another example similar to figure 3, but for a different voiced sound.

Figure 5 shows the results after applying phase interpolation to a speech segment. At the diphone boundaries, the diphone will present the averaged phase response computed for each of the two kind of sounds using other units of the database. Fig. 5.a is the original signal. Fig. 5.b presents the result after applying the first of the two methods described for phase averaging (harmonic-by-harmonic) and the first method proposed

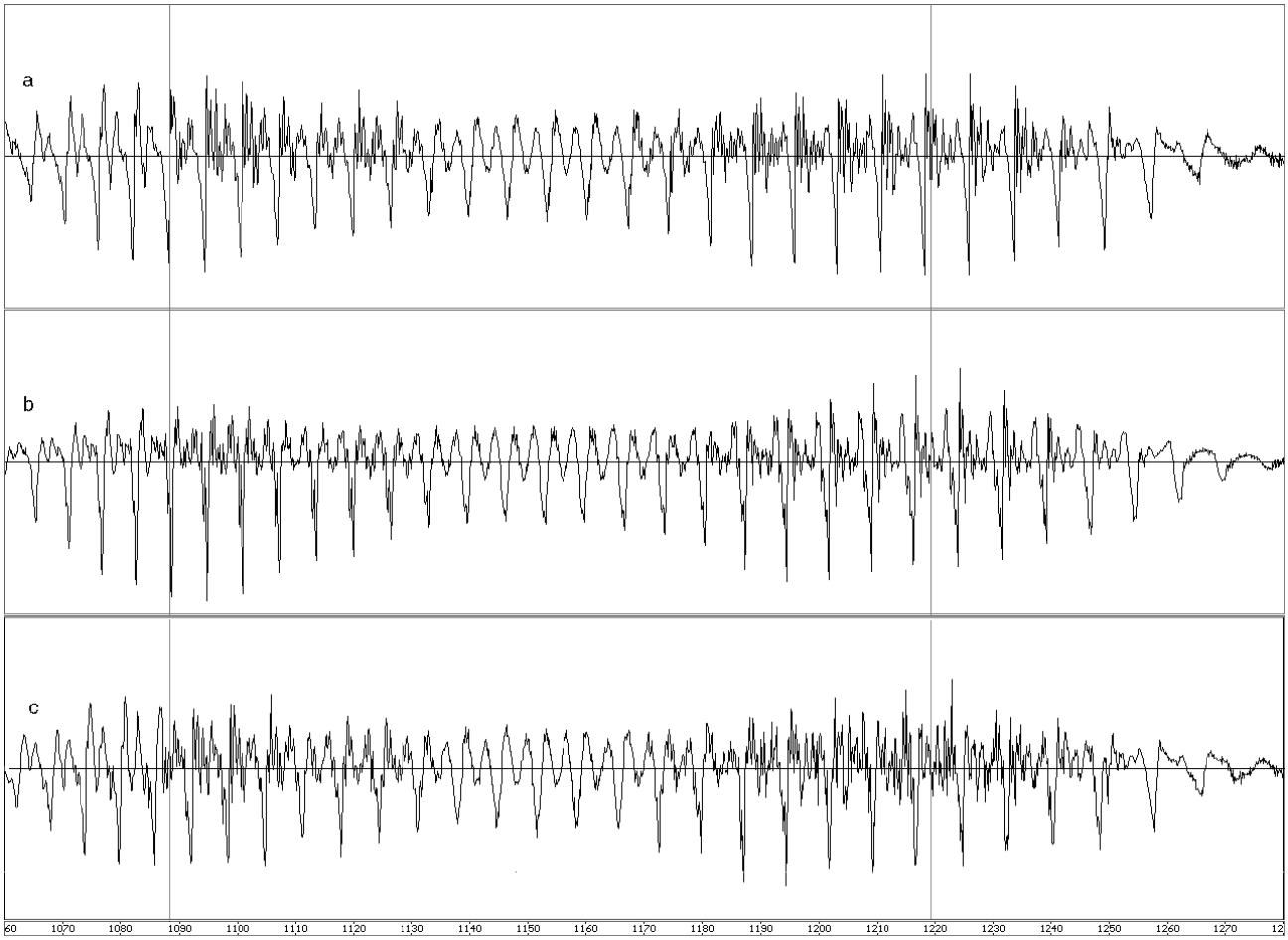

Fig. 5 (a) Original waveform for triphone o-n-a (between the boundary lines). **(b)** MBE transcoded waveform, using the averaged phase sets for 'o' and 'a' at the boundary points, method 1. **(c)** MBE transcoded waveform, using the averaged phase sets for 'o' and 'a' at the boundary points, method 2.

for interpolation (direct average interpolation). Fig. 5.c presents the resultant diphone when using the second averaging method (frequency-by-frequency) and the differential phase response interpolation. It can be seen on fig. 5.b that the central part of the diphone is very different to the original waveform, as the phase relation has nothing to do with the original one. Buzziness is still noticeable. Fig 5.c, on the other side, preserves waveform much better, not only on the central part, but also, near the boundary of the diphone. Buzzines is normally not perceptible, although sometimes we still have some strong buzziness, so the algorithm must still be improved to achieve high quality sound in the whole database.

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.

[5] B. Etxebarria, I. Hernáez, et al., *"Improving quality in a Speech Synthesizer based on the MBROLA algorithm"*, EUROSPEECH'99