

# An implementation and evaluation of two diphone-based synthesizers for Turkish

Baris Bozkurt and Thierry Dutoit

Multitel-TCTS Lab,  
Faculté Polytechnique De Mons, Belgium  
bozkurt@tcts.fpms.ac.be

## Abstract

This paper presents two diphone-based Turkish text-to-speech systems. The first system is realized inside the MBROLA project, a freely available multilingual speech synthesizer and the second system is based on shape-invariant harmonic modeling. Both synthesizers use the same parametric representations of two diphone databases (male, female) obtained by processing speech data with a pitch-asynchronous, fixed frame length harmonic/noise analyzer. To obtain a pitch-synchronous representation from the original asynchronous representation for the harmonic synthesizer, harmonic phases are submitted to a phase shifting algorithm, which also estimates maximum harmonic frequencies for each frame based on the evolution of harmonic phases. The MBROLA based synthesizer has been implemented in a rudimentary TTS system inside EULER and the harmonic synthesizer captures files produced by the EULER system to perform synthesis. Informal listening tests are being performed for quality assessment.

## 1. Introduction

Speech is often synthesized by joining diphone units using a diphone-based concatenative speech synthesizer. At synthesis stage, the prosodic features of the selected database units need to be modified to produce synthetic speech with target prosody. Additionally, units need to be concatenated such that spectral discontinuities are lowered at unit boundaries without degrading their quality. The quality of diphone-based synthetic speech highly depends on the algorithms used to alter prosody and perform concatenation.

Various methods have been presented as a synthesis algorithm. TD-PSOLA [2] is one of the most popular approaches and it does not require a parametric model. This feature provides very high quality synthetic speech when the need of modifications on units is low (when units are extracted from a very large corpus, with selecting units with prosody close to target prosody) and the spectral discontinuities at selected unit boundaries is low (another important criteria for unit selection). But it is an important limitation when a small database (i.e. a diphone database) is used as the source of units. In such a system, important prosodic modifications or mismatches at segment boundaries problems remain. Harmonic frequencies in each frame are not altered during the pitch alteration process. This distorts the harmonic character of speech. Distortion during synthesis is high if the prosody alteration is high. Additionally, TD-PSOLA systems require very accurate pitch marking.

The MBROLA [3] algorithm overcomes some of the concatenation problems by re-synthesizing voiced parts of

diphones with constant phase at constant pitch. This enables a time-domain smoothing process at segment boundaries if the segments to be concatenated are voiced and stationary at their boundaries. However the problems due to the time-domain overlap-add process remain and the phase spectrum of speech units still gets distorted.

Harmonic models [4,5,6,7] provide the flexibility to work in a parametric domain during synthesis stage (which also brings an extra computational load to the system). This provides a proper base to try new ideas to overcome problems of previously stated algorithms. The second synthesizer is designed as a harmonic synthesizer explained in section 3.

Informal listening tests are being performed via a web page, which contains speech examples synthesized with both synthesizers using the same diphone database.

A simple mapping from letters to phonemes is used to achieve text to phonetics transcription for Turkish inside EULER [1], which is a multi-lingual TTS system. The intelligibility and segmental quality of the resulting synthetic speech from both systems is reported to be very high. The systems produce monotonous speech (constant pitch) due to the lack of a prosody unit. The prosody unit is in the development stage.

## 2. MBROLA Synthesis

MBROLA synthesis aims at combining the computational efficiency of time-domain synthesis with the flexibility of a harmonic model. To achieve this target, units are first submitted to harmonic/noise analysis with constant frame length and frame shift. Then voiced frames are re-synthesized with constant pitch and constant phase envelope (for the low frequency part of the speech spectrum) with a harmonic synthesizer;

$$s(t) = \sum_{k=1}^K A_k \cos(k\theta(t) + \phi_k) \quad (1)$$

$$\theta(t) = \int_0^t \omega_o(\tau) d\tau \quad (2)$$

where  $A_k$  and  $\phi_k$  are amplitude and initial phase values corresponding to harmonic number  $k$ . The re-synthesis procedure is applied to voiced frames only and unvoiced frames are directly copied.

This eliminates pitch mismatch and some of the phase mismatch during concatenation. Another important advantage of phase reset is that spectral envelope interpolation becomes equivalent to direct temporal interpolation. This enables time domain smoothing at segment boundaries. During synthesis, the smoothing operation is applied to stationary voiced frames by distributing the difference of boundary frames

linearly to the neighbor stationary voiced frames on the left and right units.

To allow constant pitch re-synthesis while preserving the original vocal tract properties, harmonic amplitudes are recalculated by re-sampling their envelope spectrum at a constant pitch frequency. Re-synthesis at constant pitch brings a remarkable advantage to the Mbrola system: pitch marks are automatically set. This drastically reduces the time needed for database preparation because there is no need for manual pitch mark editing. The database preparation process is reduced to preparing a list of diphones, a list of logatoms, and then recording and segmenting the corresponding speech corpus.

Voiced stable states in each segment are automatically detected from the V/UV energy ratios. V/UV decision slightly suffers, especially at some fricative-voiced boundaries, from using thresholds, which are manually set for processing speech from different speakers. However this artifact is quite rare. V/UV decision is refined by detecting voiced and unvoiced stable states /transients from the V/UV energy ratios. There are currently 46 public Mbrola databases, which provide high quality synthetic speech, although most of them were analyzed and re-synthesized with the same coefficients for V/UV thresholds.

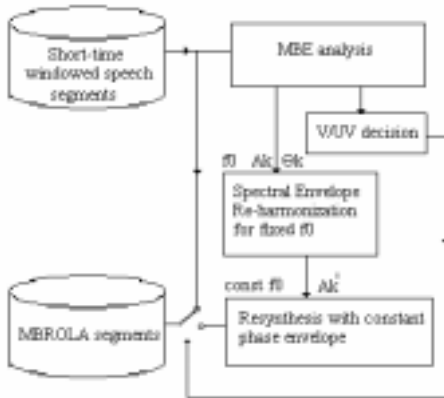


Figure 1. MBROLA database processing [2]

The Mbrola algorithm produces high quality speech but various problems remain unsolved. Among them:

- Phase distortion is introduced during re-synthesis
- Phase discontinuity arises due to the underlying time domain shift-add algorithm to obtain target pitch values, which is due to additional time shifts (deletions) inserted between pitch synchronously extracted frames.
- V/UV decision problems cannot be avoided especially for semi-voiced parts. MBROLA heavily relies on this decision; unvoiced frames might thus get erroneously re-synthesized as voiced, in which case the constant low-frequency phases imposed create artificial harmonicity. Conversely, misinterpreting a frame as unvoiced prohibits pitch modification by the MBROLA synthesis algorithm (but this is a usual drawback even with other synthesizers).
- Linear smoothing in the time domain produces artificial sounds (although not much hearable if the

left and right units are close). Natural transitions between speech sounds do not really correspond to the formant fade-in/fade-out produced by this kind of interpolation.

- Vocal tract spectrum is re-sampled with constant, fixed  $f_0$  but not with target  $f_0$  values.
- Fixed frame size pitch-asynchronous analysis lacks accuracy if recorded speech is not close to constant pitch.

### 3. Harmonic Synthesis

The second system is based on harmonic synthesis. The harmonic synthesizer uses the same parametric representation (harmonic amplitudes and phases) as the one used in the MBROLA synthesizer (actually it even uses the same harmonic analyzer) but it synthesizes speech in a completely harmonic way.

The harmonic model we use is very similar to the HNM model [8] with the following differences:

- The database (of harmonic parameters) is obtained by processing pitch-asynchronous fixed frame size harmonic/noise analysis results to obtain pitch-synchronous representations (there is no need for accurate pitch marking).
- Synthesis stage is pitch asynchronous. Duration alteration is achieved by scaling the frame lengths during synthesis and phase continuity is achieved by forcing perfect continuity for the first harmonic and shifting higher order harmonic phases linearly.
- Maximum voiced frequency estimation is based on evolution of harmonic phases.
- The unvoiced parts of speech (higher order harmonics) are synthesized with randomized phases.

These differences are further explained below.

#### 3.1 Obtaining Pitch-Synchronous Representation From Pitch-Asynchronous Representation

A few operations are to be performed before the system is ready for the synthesis stage. A phase shift algorithm, which applies a linear shift to harmonic phases, is used to obtain pitch-synchronous representation from pitch-asynchronous representation. The operation is simple; the first harmonic phase is shifted to a constant value by  $\Delta\phi(1)$ , then the shift to be applied to the  $n$ th harmonic is;

$$\Delta\phi(n) = n * \Delta\phi(1) \quad (3)$$

This process is applied to all of the frames such that the first harmonic phase is systematically shifted to the same initial value. We obtain a pitch-synchronous representation, since the operation corresponds to shifting the analysis frame in time to the position of a pitch mark, which is set relative to the first harmonic. A similar technique has previously been presented in the context of finding the center of gravity of speech waveforms [9].

This pitch-synchronous representation may also serve as a phase visualization tool, which provides evolutions of harmonic phases (relative to that of the first harmonic as a function of time) and may be useful in research focused on phase continuity problems. In Figure 2 we present an example output of such relative phase plots.

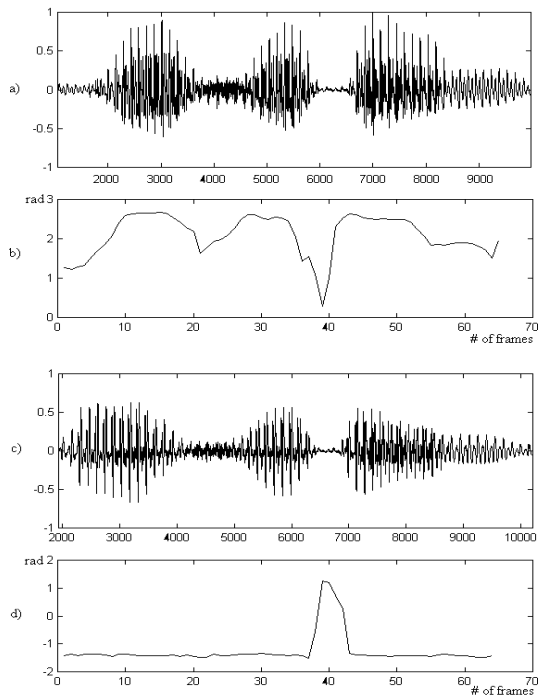


Figure 2. Phase evolution of second harmonic in original and synthetic speech (both having the same approximately flat pitch curve); a) original recording for Turkish word /gezegen/; b) relative phase plot for the second harmonic of the original recording, c) speech synthesized with MBROLA; d) relative phase plot for the second harmonic of MBROLA speech.

As a second step, we analyze these phase plots to estimate maximum voiced frequency. We assume that a harmonic phase plot will be smooth due to quasi-stationary characteristic of voiced speech, whereas an unvoiced harmonic phase (i.e. the phase of a spectral peak which has been mistaken for a harmonic) plot will not be smooth in time. The estimation process first labels each harmonic by comparing the phase derivative to a manually set threshold. When the change in phase is bigger than the threshold, the harmonic is labeled as unvoiced. These labels are further post-processed and maximum voiced frequency is then estimated for each frame by examining the results of this harmonic-by-harmonic V/UV decision. In the figure 3, we present an example output of the maximum voiced frequency estimation, obtained by post-processing harmonic-by-harmonic V/UV decision (figure 4). In figure 5, harmonic evolutions of a few harmonics which is used in maximum voiced frequency estimation is shown.

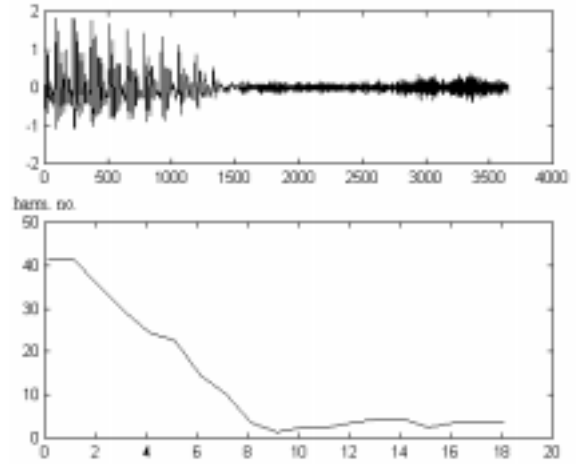


Figure 3. a) original recording for diphone /a-s/, b) Maximum voiced harmonic number estimation results.

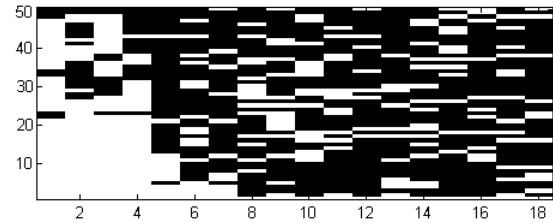


Figure 4. Harmonic-by-harmonic V/UV decisions for each harmonic (for all frames of diphone /a-s/), based on phase evolutions.

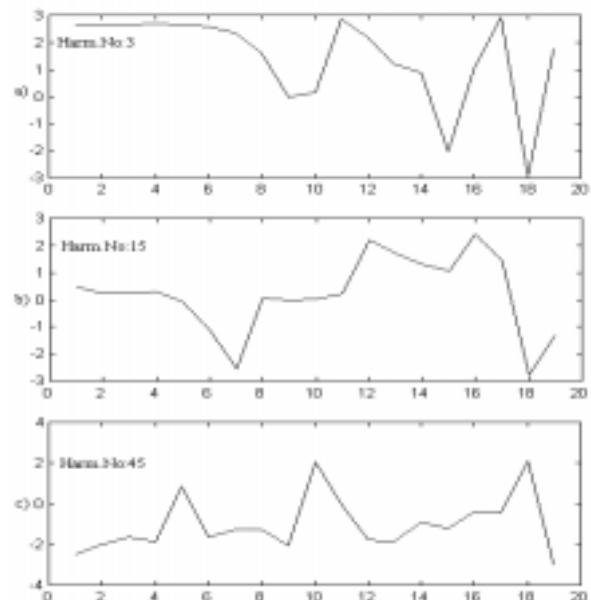


Figure 5. Relative phase plots for the diphone /a-s/

As the last step of database preparation, equalization is done by scaling harmonic amplitude values such that the sum of square of harmonic amplitudes are the same at all diphone boundaries of a phoneme. Boundary scaling factors are calculated by comparing actual values by average values and

linear envelopes are applied to (harmonic amplitudes of) diphones.

### 3.2 The synthesis stage

During synthesis, the envelope spectrum is re-sampled for target pitch values as explained in [10], separately for phase and amplitude envelopes (phase envelope being unwrapped before the re-sampling operation). With this procedure pitch alteration is successfully performed without distorting the vocal tract envelope.

Duration modification is easily performed by scaling the window lengths and shifts by the duration factor calculated as a ratio of target speech duration and source speech duration. This way, duration alteration steps are no more quantized in number of frames as in the case with TD-PSOLA, MBROLA, HNM synthesizers. Phase continuity is achieved by forcing the first harmonic to be perfectly continuous and shifting the voiced harmonics with the same phase shift procedure. The phase change of the first harmonic is simply calculated by

$$\Delta\phi(1) = (\text{shift} / \text{period}) * 2\pi \quad (5)$$

The phases of harmonics having a frequency lower than the local maximum voiced frequency are submitted to linear shifting to obtain phase continuity in the shape-invariant OLA process; phases of harmonics at higher frequencies are randomized to obtain noise like speech. As a result, the first harmonic is perfectly continuous and the phase evolutions of higher voiced harmonics follow their original characteristic (time stretched depending on applied duration alteration ratio) while phases of unvoiced harmonics are randomized.

To achieve spectral continuity at segment boundaries, a smoothing process is applied to the voiced harmonics of stationary frames. Smoothing is performed by a symmetric moving average filtering (degree of smoothing can be specified by the user who defines the number of times filtering will be performed) on harmonic amplitude evolutions, at concatenation boundaries, inside stationary voiced frames.

## 4. Discussion

Informal listening tests are being performed (via the web) and the results will be presented on the day of presentation. Example outputs are available on the informal listening test page (<http://tcts.fpms.ac.be/~bozkurt/testturkish.htm>). A few listening (researchers experienced in speech) tests have been performed and the qualities of the two systems are graded to provide very highly intelligible synthetic speech. The early comparisons of the two systems show that the MBROLA system provides better quality in unvoiced segments of speech (which are directly copied from real speech) and the synthesis speed is very high, since computational cost is very low at synthesis stage. The advantages of the harmonic synthesizer stem from its parametric character (considering further improvements) and the maximum frequency concept successfully replaces the need for a voiced/unvoiced decision, which is hard in the context of speaker independency. On the other hand the quality of unvoiced speech segments and the speed of synthesis are lower in our harmonic synthesizer. Further improvements of our harmonic synthesizer will be on better phase continuity and better unvoiced speech synthesis. A prosody generation unit is also in development stage.

## 5. References

- [1] Dutoit, T., Bagein, M., Malfre, F., Pagel, V., Ruelle, A., Tounsi, N., and D. Wynsberghe, "EULER : an Open, Generic, Multi-lingual and Multi-Platform Text-To-Speech System" , *Proc. LREC'00, Athens, May 2000*, p. 563-566.
- [2] Moulines, E. and F.Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones", *Speech Commun., Vol.9, Dec.1990*, p 453-497.
- [3] Dutoit, T. and H.Leich, "Text-to-speech synthesis based on a MBE re-synthesis of segments database", *Speech Commun., Vol.13, 1993*, p 435-440.
- [4] McAulay, R.J., and T.F.Quatieri, "Speech analysis/synthesis based on a sinusoidal representation", *IEEE Trans. Acous., Speech, Signal Processing, Vol.34, Aug.1986*, p 744-754.
- [5] Marques, J., and L. Almeida, "Frequency-varying sinusoidal modeling of speech", *IEEE Trans. Acous., Speech, Signal Processing, Vol.37, 1989*, p 763-765.
- [6] Macon, M.W., "Speech synthesis based on sinusoidal modelling", PhD. Dissertation, Georgia Inst. Technol., Atlanta, Oct. 1996.
- [7] Stylianou, Y., "Harmonic plus noise models for speech, combined with statistical methods, for speech and speaker modification", PhD. Dissertation, Ecole Nationale Supérieure des Télécommunications, Paris, France, Jan. 1996.
- [8] Stylianou, Y., Dutoit T. and J. Schroeter "Diphone concatenation using a harmonic plus noise model of speech," *Proc. of Eurospeech, 1997*, pp.613-616.
- [9] Stylianou, Y., "Removing phase mismatches in concatenative speech synthesis", *Proc. 3rd ESCA Speech Synthesis Workshop, Nov. 1998*, p 267-272.
- [10] Stylianou, Y., Laroche, J., and E. Moulines "High quality speech modification based on a harmonic + noise model", *Proc. Eurospeech, 1995*, p 451-454.
- [11] Macon, M.W., and M. A. Clements, "Speech concatenation and synthesis using an overlap-add sinusoidal model," *Proc. of Int. Conf. on Acoustics, Speech, and Signal Processing, 1996*, pp. 361-364.
- [12] Macon, M.W., and M. A. Clements, "Speech concatenation and synthesis using an overlap-add sinusoidal model," *Proc. of Eurospeech, 1999*, pp. 2327-2330.
- [13] Banga, E. R., and C. Garcia-Mateo "Shape invariant pitch-synchronous text-to-speech conversion", *Proc. of Int. Conf. on Acoustics, Speech, and Signal Processing, 1995*, pp. 361-364.
- [14] Dutoit, T. 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.
- [15] Stylianou, Y., "Applying the harmonic plus noise model in concatenative speech synthesis", *IEEE Trans. Acous., Speech, Signal Processing, Vol.9, Jan.2001*, p 21-29.