# SHAPE-INVARIANT PROSODIC MODIFICATION ALGORITHM FOR CONCATENATIVE TEXT-TO-SPEECH SYNTHESIS

Eduardo R. Banga, Carmen García-Mateo and Xavier Fernández-Salgado

Dpto. Tecnologías de las Comunicaciones. ETSI Telecomunicación.
Universidad de Vigo. E-36200. Vigo. SPAIN
e-mail: erbanga@tsc.uvigo.es carmen@tsc.uvigo.es
xsalgado@tsc.uvigo.es

## ABSTRACT

Concatenative text-to-speech systems require an algorithm that allows prosodic modifications of the speech units during the concatenation process. Nowadays, sinusoidal modeling seems to be a promising technique to achieve very flexible algorithms that provide high quality synthetic speech. The main difficulty of these type of algorithms is the treatment of the phase information, since an inadequate processing of this information gives rise to reverberation and audible artefacts. In this contribution we discuss the application of a shape-invariant sinusoidal model [1] to a text-to-speech system based on concatenation of speech units.

## 1. INTRODUCTION

Presently there exist several techniques for prosodic modification of the speech signal. Among them, we must highlight the TD-PSOLA algorithm [2] because of its simplicity and the quality of the resulting synthetic speech. However, since the TD-PSOLA works in the time domain, it is not able to modify the spectral characteristics of the speech as it would be desirable in some cases. For instance, it is not able to carry out large variations of the fundamental frequency required for voice conversion applications.

Over the last few years, there has been a growing interest on the sinusoidal models and their capabilities to overcome the limitations of TD-PSOLA [3]. In fact, sinusoidal models are the alternatives to TD-PSOLA, although more complex and computationally expensive. The major difficulty of sinusoidal models is the treatment of the phase information of the speech, since inadequate alterations result in synthetic speech that does not preserve the temporal structure of the original speech waveform. Shape-Invariance is referred to as the property that maintains most of the temporal structure of the speech in spite of pitch modifications.

A shape-invariant prosodic modification algorithm for continuous speech that delivers good synthetic speech quality is proposed in [1]. Nevertheless, in its application to the case of synthesis by concatenation it is necessary to introduce some modifications. This is because it is necessary to assure phase continuity when we concatenate two realizations of an allophone that were segmented from different words. In [4] an initial approximation to this problem was described. In this paper the method has been further refined.

The rest of this paper is organized as follows: in Section 2, we present the shape invariant sinusoidal model and its capabilities to modify the prosody of the speech; in Section 3 we describe the changes introduced in the previous model for concatenative synthesis; finally, in Section 4 we present some results.

## 2. THE SHAPE-INVARIANT MODEL

This method works on a frame by frame basis by modeling the speech signal as the response of a linear system, h(t), to an excitation signal e(t).The excitation signal is represented using a sinusoidal model and so does the speech signal, s(t), that is,

$$e(t) = \sum_{l=1}^{L} a_l(t) \cdot \cos[\Omega_l(t)] \quad (1)$$

$$s(t) = \sum_{l=1}^{L} A_l(t) \cdot \cos[\theta_l(t)] \quad (2)$$

where L represents the number of significant spectral peaks in the short-time spectrum of the speech signal and where $a_l(t)$, $A_l(t)$, $\Omega_l(t)$, and $\theta_l(t)$ denote the amplitudes and instantaneous phases of the sinusoidal components, respectively. The amplitudes and instantaneous frequencies of the excitation and the speech signal are related by:

$$A_l(t) = a_l(t) \cdot M_l(t) \quad (3)$$

$$\theta_l(t) = \Omega_l(t) + \psi_l(t) \quad (4)$$

where $M_l(t)$ and $\psi_l(t)$ represent the magnitude and phase of the transfer function of the linear system at the frequency of the l-th spectral peak. The excitation phase representation can be simplified by introducing the pitch pulse onset time, $t_o$, and assuming that the l-th peak frequency, $\omega_l$, is constant over the duration of a speech frame, that is

$$\Omega_l(t) = (t - t_o) \cdot \omega_l \quad (5)$$

so the system phase can be estimated as

$$\psi_l(t) = \theta_l(t) - (t - t_o) \cdot \omega_l \quad (6)$$

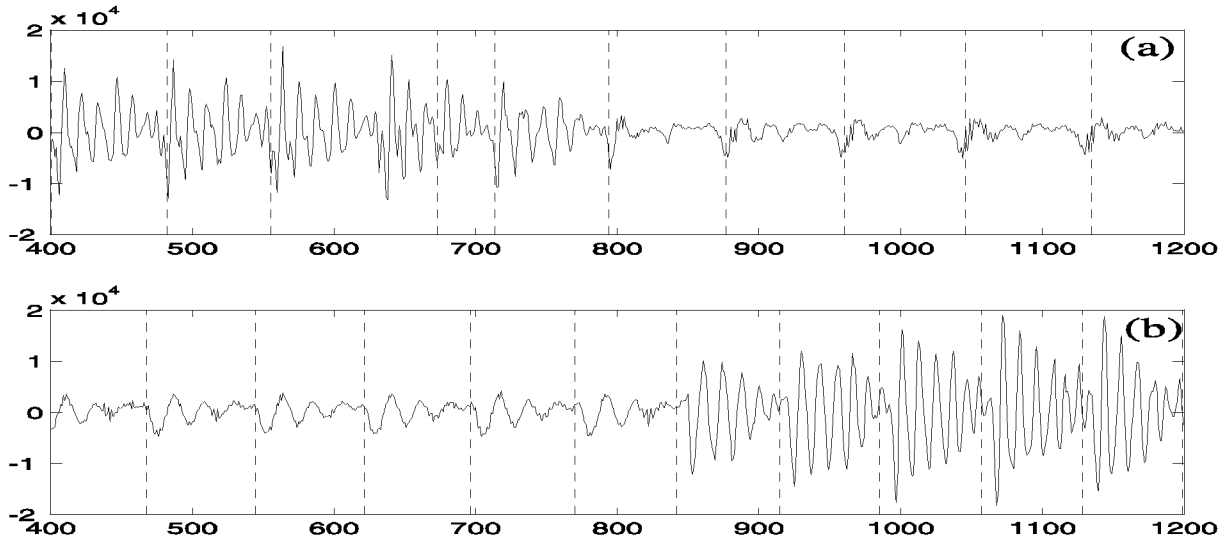


**Figure 1:** *Variability of the estimation of the pitch pulse onset times*

The synthetic speech obtained with the shape-invariant method basically maintains phase relations among the different sinusoidal contributions, so it does not sound reverberant. The quality is improved further when we have a prior knowledge about the sounds we are processing, like in text-to-speech applications. For example, we know that pitch modifications do not affect unvoiced sounds and that duration modifications affect some sounds more than others.

## 3. THE PITCH-SYNCHRONOUS SHAPE-INVARIANT MODEL

Let us assume that in any stationary segment of an allophone the configuration and temporal evolution of the system that form vocal tract and glottis are similar in different realisations. In accordance with the Shape-Invariant model the instantaneous phase of the l-th component is given by

$$\theta_l(t) = \psi_l(t) + \omega_l(t - t_0) \tag{7}$$

For $t=t_0$ we obtain

$$\theta_l(t_0) = \psi_l(t_0) \tag{8}$$

that is, at the pitch pulse onset time the instantaneous phase is equal to the system phase. We can also assume that the system phase is slowly time-variant, so system phases at consecutive pitch pulse onset times will be quite similar. This consideration can be extended to the case in which we are trying to concatenate two segments of a same allophone that belong to units obtained from different words. We could think that the pitch pulse onset times are good concatenation points since at those times the instantaneous phase is equal to the system phase and we can consider the system phase slowly variant. Obviously, this assumption will be valid mainly in the central periods of the allophone, where the coarticulation effect is minimized. Also it will depend on the variability of the speaker's voice, that is, on the similarity among the waveforms of the different realizations of the allophones.

However, in practice the pitch pulse onset time is not the most appropriate place to concatenate speech units as we illustrate with the following examples. In figure 1 we can observe the different onset times obtained for the Spanish diphones /al/ and /la/ . Apart from a clear error in the upper graph as a consequence that the error function used for the estimation presents numerous local minima, we can observe that the position of the different onset times (relative to a period) present a certain variability even for the very same allophone. Using this method duration modifications are obtained by time scaling the excitation amplitudes and the magnitude and the phase envelope of the linear system. Pitch modifications can be done by scaling the peak frequencies to the desired values, estimating the magnitude and the phase of the linear system for those new frequencies and taking into account that the new pitch pulse onset times are separated by the new pitch period.

In the case of the phoneme /l/ we can notice that in the upper plot the onset times are located approximately at the local minimum of each pitch period, while in the bottom plot they have been displaced significantly from this point. This effect can be even more noticeable in other allophones. Due to the variability in their situation, in case of using the pitch pulse onset times like concatenation points, alterations of the instantaneous phases appear. As a consequence, the periodic structure of the speech signal is destroyed.In order to avoid these effects we decided to use a set of pitch marks placed pitch synchronously on voiced portions and at a constant rate on unvoiced segments. On a stationary segment the pitch marks, $t_m$ , are located

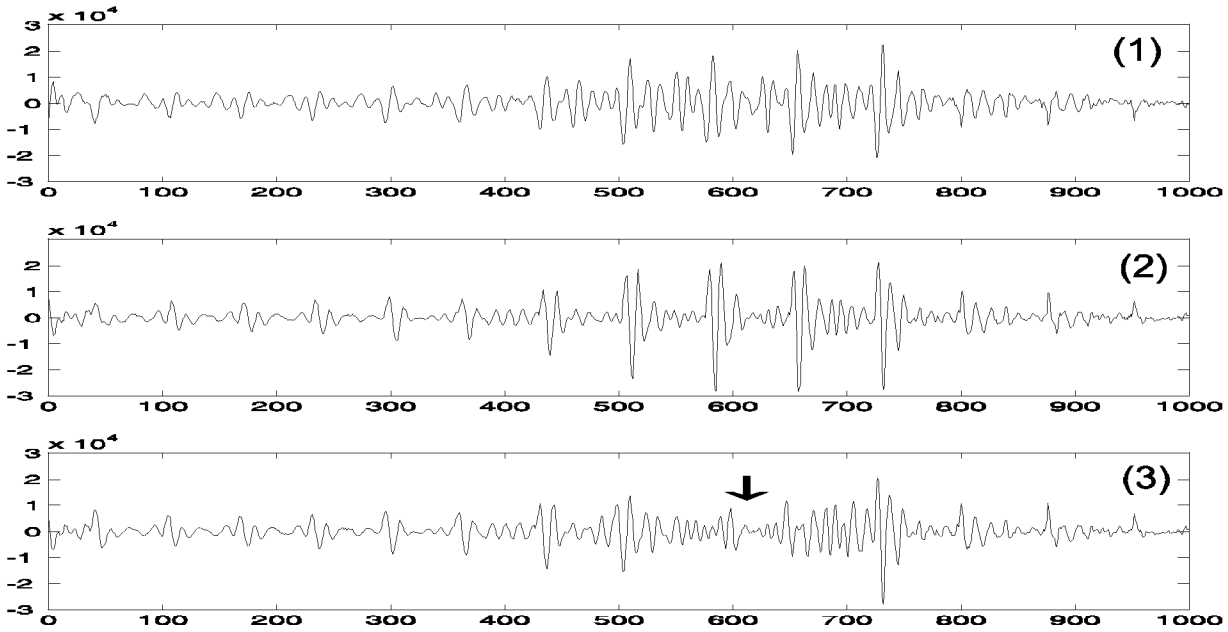


**Figure 2:** *Example of results of experiments 1,2 and 3*

at a constant distance, $t_d$ , from the authentic pitch pulse onset time, that is, the instant of glottal closure, $T_0$. Therefore,

$$t_m = T_0 + t_d \tag{9}$$

and the instantaneous phase at $t=t_m$ is

$$\begin{aligned} \theta_l(t_m) &= \psi_l(t_m) + (t_m - T_o) \bullet \omega_l \\ &= \psi_l(t_m) + t_d \bullet \omega_l \end{aligned} \tag{10}$$

so it is equal to the system phase plus a linear phase component. Assuming local stationarity, the difference between the glottal closure instant and $t_d$ is maintained along successive periods. The linear phase component is equivalent to a time shift which, from a perceptual point of view, is irrelevant.

From the above result we also conclude that any set of time marks placed pitch synchronously can be used as pitch marks, independently of their situation in the pitch period. In addition, if during the analysis stage, the analysis window is centred at the pitch marks, the phases of the spectral peaks are equal to the system phases (except for the linear phase term). Then, if we use a pitch-synchronous analysis it is not necessary to estimate the pitch pulse onset time. We can consider that the onset time is any constant (zero, for example).

## 4. APPLICATION TO CONCATENATIVE SYNTHESIS

In this section we discuss the application of the previous model to a text-to-speech system based on speech units concatenation. The first step is to determine the set of pitch marks of the speech units database. In order to do this, we used a pitch determination algorithm combined with a voiced/unvoiced classifier. During voiced segments pitch marks are placed at local maximum (in absolute value) for every pitch period while during unvoiced segments they are placed every 10 ms. Pitch marks are visually inspected and occasional errors corrected.

The next step is a pitch synchronous analysis of the database of speech units. Every speech frame is parameterized by frequencies, magnitudes and phases of the spectral peaks. As a consequence of the pitch-synchronous analysis, the phases of the spectral peaks are a good estimation of the system phase.

The main disadvantage of this method is the large number of parameters that have to be stored. Assuming that the peak frequencies are harmonically related this number can be significantly reduced. If more efficient storage is required we need to introduce some additional simplifications to the model. For example, if we assume that the glottis and the vocal tract form a minimum phase system, then the systems phase can be obtained from the magnitude of the spectral peaks. Nevertheless, as we will see later, this simplification modifies the speech waveform.

Pitch and duration changes are made using the original Shape-Invariant method as the third step in the concatenation process. Amplitude discontinuities between speech units are avoided by the frame-to-frame linear interpolation and a previous energy normalization of the speech units.

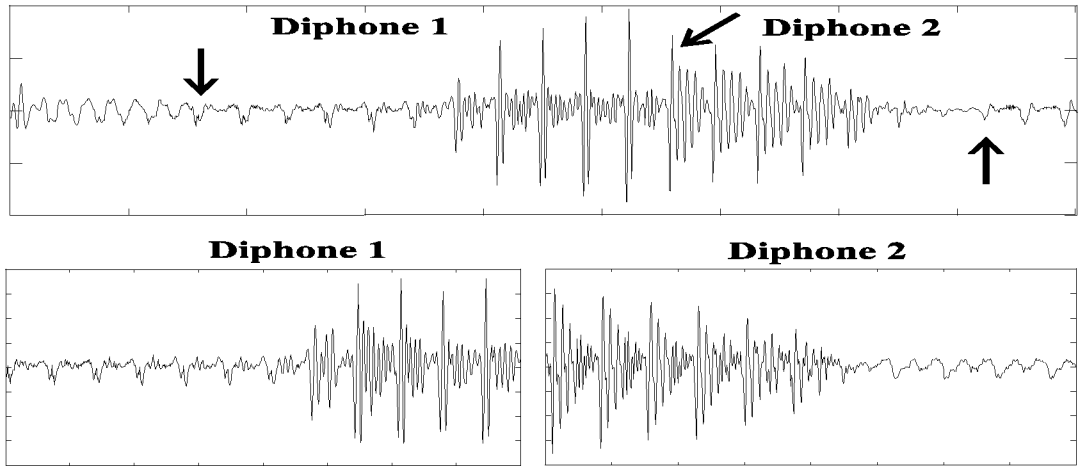


**Figure 3:** *Concatenation of two diphones*

## 5. RESULTS

In order to evaluate the performance of the Pitch-Synchronous Shape-Invariant Model, we applied this method to a Spanish text-to-speech system based ondiphones concatenation. The speech units database consists of 968 units sampled at 8 KHz obtained from a male speaker with a mean fundamental frequency about 120 Hz. We conducted three experiments:

- **Experiment 1:** A pitch-synchronous analysis of the speech units database was made. Frequencies, magnitudes and phases of the spectral peaks are computed for every frame. We employed the phases of the spectral peaks as an estimation of the system phase.
- **Experiment 2:** Similar to experiment 1, but now the system phase was assumed to be minimum phase and it was estimated from the spectral magnitudes. Again, as a consequence of the pitch-synchronous analysis the system phase is equal to the set of phases of the spectral peaks.
- **Experiment 3:** The pitch-synchronous analysis is maintained, but this time we used the original estimation of the pitch pulse onset time. In this case the system phase was not supposed to be equal to the phases of the spectral peaks. In this experiment we try to observe the influence of the variability of the onset time estimation on the synthetic speech signal. This experiment is equivalent to considering the onset time as the concatenation point between speech units.

In figure 2 we can observe the results of the three experiments on a synthetic speech segment. In the results of experiment 1 the periodic structure of the speech is preserved and the waveform of the diphones is basically maintained (even when the realizations of the common allophone are not quite similar, as illustrated in figure 3). The minimum phase approximation, taken in experiment 2, modifies the original speech waveform although it maintains the periodic structure. Finally, in experiment 3 we notice the alterations in the periodic structure of the speech, a consequence of using the onset time as concatenation point. These alterations are perceived as "hoarseness" in the synthetic speech.

## 6. CONCLUSIONS

In this contribution, we have discussed how the sinusoidal shape-invariant method can be applied to speech synthesis by concatenation. The most important advantages of this model are its flexibility and the quality of the synthetic speech. This method is also easily implemented on a hybrid model that separates voiced an unvoiced components of the speech signal [4].

## 7. ACKNOWLEDGEMENTS

This work has been partially supported by the Spanish CICYT under the project TIC96-0956-C04-02

## 8. REFERENCES

[1] Quatieri, T.F.and McAulay, R.J. (1992), "Shape invariant time-scale and pitch modification of speech", IEEE Transactions on Signal Processing, March 1992, vol.40, (no.3):497-510.

[2] Moulines, E. and Charpentier, F. (1990),"Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones", Speech Communication, Dec. 1990, vol.9, (no.5-6):453-467.

[3] Stylianou, Y and Laroche J. and Moulines, E. (1995), "High-Quality Speech Modification based on a Harmonic+Noise Model", EUROSPEECH'95. Madrid. Spain. p. 451-454

[4] Banga, E. R. and García-Mateo C. (1995) ,"Shape-Invariant Pitch-Synchronous Text-to-Speech Conversion ", Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'95) Detroit. U.S.A. 1995.