SHAPE-INVARIANT PITCH-SYNCHRONOUS TEXT-TO-SPEECH CONVERSION

Eduardo R. Banga and Carmen García-Mateo

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

ABSTRACT

Text-to-speech (T-T-S) systems based on the concatenation of speech units need a prosodic modification algorithm to adjust the prosodic features of the stored speech units to the desired output values. In this paper, we discuss the application of a sinusoidal shape-invariant model to a T-T-S system for Spanish, paying special attention to the concatenation issues and phase treatment. The resulting speech waveform resembles the waveform of its contributory units, without sounding reverberant as in other sinusoidal implementations.

1. INTRODUCTION

Text-to-speech (T-T-S) systems based on the concatenation of speech units are widely employed nowadays, mainly because they are a good compromise between complexity, flexibility and performance. These systems try to overcome the lack of knowledge of the human phonation process by considering the speech sounds in many different phonetic contexts and then using a prosodic modification algorithm to adjust the prosodic features (pitch, duration and energy) of the stored speech units to the desired output values. Recent attempts to accomplish the prosodic modification task include PSOLA (Pitch Synchronous Overlap-Add) techniques [1]; its time-domain version (TD-PSOLA) is especially simple and produces good synthetic quality. However, TD-PSOLA performs poorly whenever the prosodic pattern requires a severe modification in the spectral content of the stored units. While the frequency domain version of PSOLA (FD-PSOLA) partially solves this problem, it introduces some spectral distortion because of the techniques employed to estimate the complex amplitudes of the new pitch harmonics.

As an alternative to PSOLA techniques, we proposed in [2] a prosodic modification method based on the IMBE (Multiband Excitation) vocoder [3]. (The IMBE vocoder is a harmonic speech coder which, after performing a frequency domain analysis, separates the voiced component and the unvoiced component, and applies different synthesis methods to each component.) An advantage of our algorithm [2] is that pitch changes affect only the voiced component, while the unvoiced component is left unchanged, so the local voiced/unvoiced property of the spectrum is not altered. Duration is also easily modified by changing the length of the synthesis frame. This model is simpler and more flexible than FD-PSOLA, since the source component need not be extracted and the voiced and the unvoiced components can be modified separately. Also, since it is based on a speech coder, it provides an efficient way to encode the speech units database. However, since in [2] we did not extract phase information from the original spectrum -and thus could not establish phase relations among pitch harmonics- the temporal structure of the original waveform was not maintained. The resulting phase distortion in the synthetic speech signal is perceived by listeners as a moderate amount of reverberation.

Just as phase information is thus crucial in speech coding to guarantee waveform matching, so is it in prosodic modification to ensure shape invariance. A shape-invariant prosodic modification algorithm which delivers good synthetic speech quality is proposed in [4], and is summarized in Section 2.

The purpose of this paper is to illustrate how the method presented in [4] can be applied to a T-T-S system. The rest of this paper is organized as follows: in Section 2, we outline the shape invariant sinusoidal model and its capabilities to modify the prosody of the analyzed speech. In Section 3, we show how to use this algorithm in the context of a T-T-S system and, in Section 4 we describe the characteristics of the experiments we have conducted. In Section 5 we explain how the quality of a T-T-S system based on the Multi-Band Sinusoidal coder can be improved by making use of phase information provided by the speech units to eliminate most of the reverberation. We close by providing some conclusions.

2. THE SHAPE-INVARIANT MODEL

The method presented in [4] is a sinusoidal scheme, but without explicitly distinguishing between voiced and unvoiced components. Both the excitation signal and the

0-7803-2431-5/95 $4.00 © 1995 IEEE

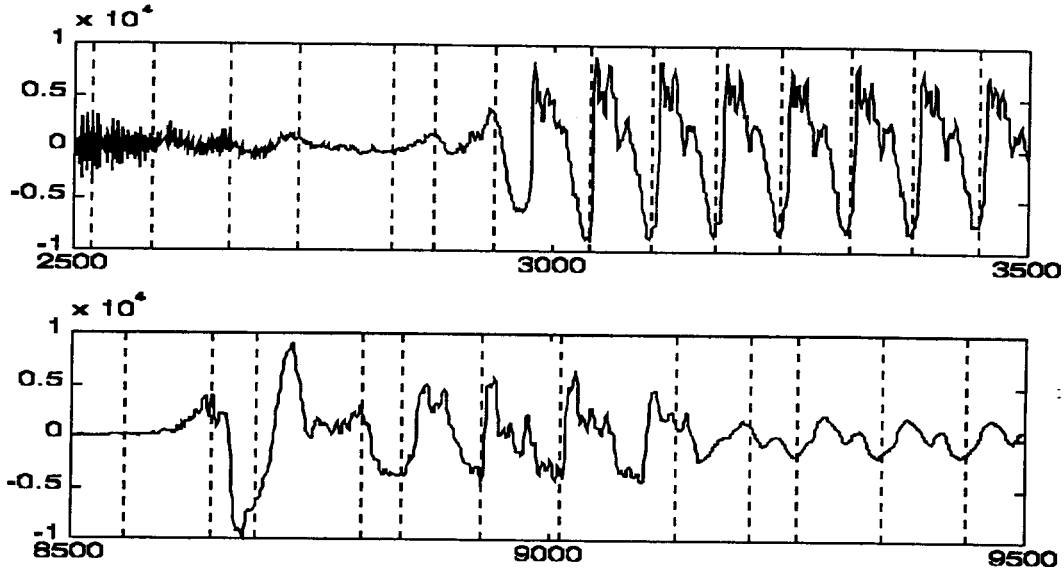

Figure 1: Pitch onset times of two speech segments

speech signal are modeled, frame by frame, as the response of a (time-varying) linear system, h(t), to an excitation signal e(t). Within each frame, the excitation signal 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 is the number of significant spectral peaks in the short-time spectrum of either signal, and $a_l(t)$, $A_l(t)$ and $\Omega_l(t)$, $\theta_l(t)$ denote the amplitudes and phases of the sinusoidal components, respectively. These 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 (in the current frame) at the frequency of the l-th spectral peak. If it is assumed that the l-th peak frequency, ω_l, is constant over the duration of a speech frame, then with the pitch pulse onset time denoted by t_o, the instantaneous phase of the excitation signal within the l-th frame can be expressed as:

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

and then the system phase can be estimated from (4) as

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

With this method duration and pitch can be changed, so that the prosody of the synthetic speech is altered. To modify duration, we must time-scale the excitation amplitudes as well as the magnitude and phase envelopes of the linear system. Pitch is modified by scaling the peak frequencies to the desired values, and then estimating the new magnitude and phase of the linear system. Meanwhile we must also ensure that the modified pitch pulse onset times are separated by the, new pitch period.

The synthetic speech obtained with the shape-invariant method basically maintains phase relations among the different sinusoidal contributions, so it does not sound reverberant. Moreover, the quality can even be improved, because we can take advantage of prior knowledge about the sounds we are processing: pitch modifications do not affect unvoiced sounds, and duration modifications affect some sounds more than others.

3. APPLICATION OF THE SHAPE-INVARIANT MODEL TO A TEXT-TO-SPEECH CONVERTER

When applying the method described above to a T-T-S system based on the concatenation of speech units, an additional problem arises: how to concatenate the selected units without introducing waveform discontinuities? The solution will be that every two speech segments must be joined within a fairly stationary region, and at a point where their waveforms are quite

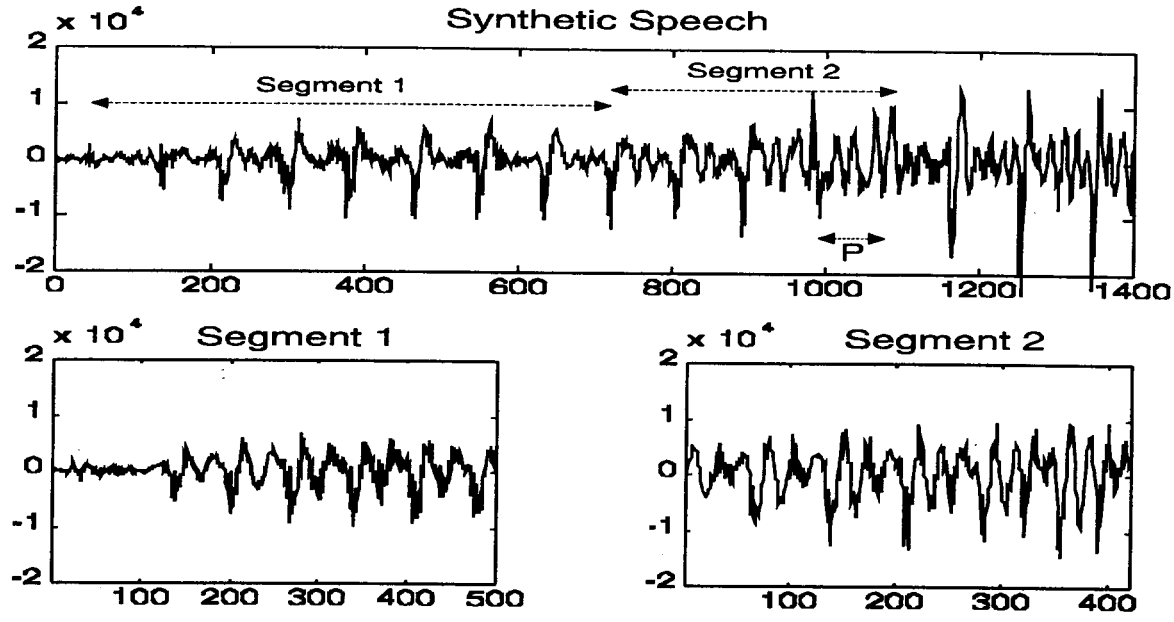

Figure 2: Synthetic speech obtained by concatenation of two speech segments

similar. Otherwise, phase discontinuities would appear damaging the periodic structure of the speech and causing audible artifacts. Therefore, one important task is to figure out the best place where to join the different speech units.

Figure 1 shows two different speech segments with superposed vertical dashed lines to mark pitch onset times [4]. Since in a voiced speech segment, the distance between two onset times is the current pitch period, a reasonable choice as junction point would be the onset time. Nevertheless, to avoid phase jitters which give rise to a rough synthetic speech, the onset time must be estimated with considerably accuracy and, generally, it is not an integer number of samples. Moreover, since the onset time is obtained by minimizing an error function that has many local minima, some deviations from the expected value were observed, especially in transition regions where the stationary assumption fails. To solve this problem, we use a set of pitch marks as reference points. For voiced speech, pitch marks were located at the maximum of the different pitch periods, while for unvoiced speech pitch marks were arbitrarily equispaced at 10 ms intervals. The analysis windows were then centered on the pitch marks; since these were usually near the onset time, this resulted in a pitch synchronous analysis.

4. EXPERIMENTAL RESULTS WITH THE SHAPE-INVARIANT PITCH-SYNCHRONOUS MODEL

We applied the shape-invariant method to our Spanish text-to-speech converter based on diphones concatenation. 968 diphones were extracted from nonsense words pronounced by a male talker. The speech units were low-pass filtered and sampled at 8 KHz. A set of pitch marks were determined for the different speech units as described in the previous section; this was aided by the knowledge that the normal pitch range of the talker was known to be around 120 Hz.

Next, the speech units database was analyzed pitch-synchronously to obtain the set of about 35 peak frequencies, their amplitudes, the onset times (relative to the center of the window) with a precision of 0.1 time samples and a parameter β (0 or 1) to derive the sign of the synthetic signal.

In applications where the speech units database must be stored more efficiently, we can apply the restriction that the peak frequencies must be harmonics of the fundamental frequency (the inverse of the distance between consecutive pitch marks). This assumption slightly degrades the quality of the synthetic speech.

Prosodic information is obtained from the linguistic-prosodic module for Spanish described in [5]. The synthesis is performed as described in [4] and is not pitch synchronous. The new pitch onset times are placed according to the new fundamental frequency. During unvoiced sounds the onset time (relative to the center of the synthesis frame) may be set to zero.

An example of the synthesis is shown in Figure 2. Segment 1 and Segment 2 (lower half) are concatenated after scaling to generate the synthetic speech waveform. As is seen, the shape of the segment waveforms is basically maintained even though the pitch and duration have been modified. Notice also that the sign of a pitch cycle, denoted by P in Figure 2, has been changed via the parameter β during analysis. Yet, such a change in the sign is not perceptually important if it does not cause a sudden jump in the instantaneous phase, a condition that is ensured by the phase smoothing algorithm employed at the synthesis stage.

Informal listening tests have shown that the synthetic speech signal does not have the typical reverberant quality of earlier methods, especially for moderated pitch scale factors.

5. IMPROVED PROSODIC MODIFICATIONS BASED ON A IMBE MODEL

As it was mentioned in the Introduction, in [2] we proposed the use of a IMBE model as an effective way to separate the speech in a voiced and an unvoiced component. The synthetic unvoiced component was generated as the inverse Fourier transform of a noisy spectrum, and it was not affected by pitch changes. The voiced component was obtained as the sum of the contributions of the pitch harmonics. The amplitudes of the new pitch harmonics were obtained by linear interpolation from the original ones, while the phases were determined as a linear function of the fundamental frequency, assuming that the pitch is constant during a frame. Phase continuity was assured by using a second or maximally-smooth third degree polynomial. This method did not maintain the phase relations among the pitch harmonics and so some reverberation is perceived.

In order to improve this model, the harmonic phases can be estimated using the onset time method and the reverberation will be eliminated. The synthesis stage of the IMBE model is computationally less complex than the shape-invariant model because the unvoiced component is generated by means of a unique IFFT instead of as a sum of sinusoids. However, some fricative sounds obtained by the IMBE model sound somewhat less natural than the same sounds obtained by the shape-invariant model.

6. CONCLUSIONS

We have discussed how the sinusoidal shape-invariant method can be used in a T-T-S based on speech units concatenation. Some variations of this model oriented to applications with memory restrictions were also presented. The most important advantages of this model are its flexibility and the quality of the synthetic speech obtained. However, much work needs still to be done to exploit all the flexibility this model offers.

7. ACKNOWLEDGMENTS

This work was partially supported by PRONTIC under the project TIC92-0800-C05-02 and "Oficina PLANBA" under project "TEMA"..

We wish to acknowledge Francisco García and Leandro Rodríguez because they did much of the hard work. We also acknowledge to our colleagues and good friends Eduardo López and Luis Hernández for their many suggestions and useful advises. Finally, thanks to Andrew Sekey from UCSB for his editorial help.

REFERENCES

[1] 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):453-67.

[2] Banga, E. R. and García-Mateo C. (1993b), "New frequency domain prosodic modification techniques", 3rd European Conference on Speech Communication and Technology.(EUROSPEECH'93). Berlin. Germany. 1993. vol 2, p. 987-90.

[3] "INMARSAT-M voice coding system description," DRAFT version 1.3, 1991.

[4] 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.

[5] López-Gonzalo, E. and Hernández-Gómez L. (1994), "Data-Driven Joint F_0 and Duration Modeling in Text to Speech Conversion for Spanish", Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'94) Adelaide. Australia. 1994. vol. 1, p. 589-92.