SPEECH SYNTHESIS AND PROSODY MODIFICATION USING SEGMENTATION AND MODELING OF THE EXCITATION SIGNAL

J.M. Gutiérrez Arriola, F.M. Giménez de los Galanes, M.H. Savoji, J.M. Pardo

Grupo de Tecnología del Habla, Departamento de Ingeniería Electrónica,
E.T.S.I. Telecomunicación, Universidad Politécnica de Madrid
Ciudad Universitaria. 28040, Madrid. Spain

ABSTRACT

In previous work we have presented a new method for improving the quality of LPC synthetic speech, where the excitation signal was modelled with a polynomial function followed by an adaptive filter. This scheme provides the properties of mathematical models which permits avoiding the problems related to prosody control [1], [2]. In order to reduce the storage needs, a segmentation technique was developed which grouped together several pitch periods based on spectral similarity. For every segment the same coefficient set (both the polynomial function and the post-processing filter) was used. These techniques were applied to a codification/decodification task were the resulting speech quality was promising [1], [2].

In this paper we present some results concerning prosodic modification, i.e. duration and fundamental frequency arbitrary changes which show the suitability of these methods for text-to-speech applications. We also present some results of the extension of the model to unvoiced segments of speech.

1. SOURCE MODEL DESCRIPTION

The original signal is first pitch marked with an algorithm very similar to that defined in [3]. Then, it is pitch synchronously analysed using the Durbin algorithm to calculate the prediction coefficients. The analysis window used is a two-period long Hamming window centred on every pitch mark. The original signal is filtered using these coefficients to obtain the LP excitation signal. This excitation signal is modelled for voiced parts using a parametric (polynomial) version of the original excitation signal. In this first version of the system the unvoiced segments were synthesised using a stochastic function as excitation.

1.1. Polynomial interpolation.

We use a 6th-order polynomial waveform model to represent the derivative of the glottal volume velocity waveform [4]. This derivative function is computed by direct integration of the residual and high pass filtering to zero-centre the resulting signal.

The polynomial function is obtained by curve fitting in a least square sense where a fine-tuning or readjustment is needed to exactly synchronise the pitch marks with the most negative sample.

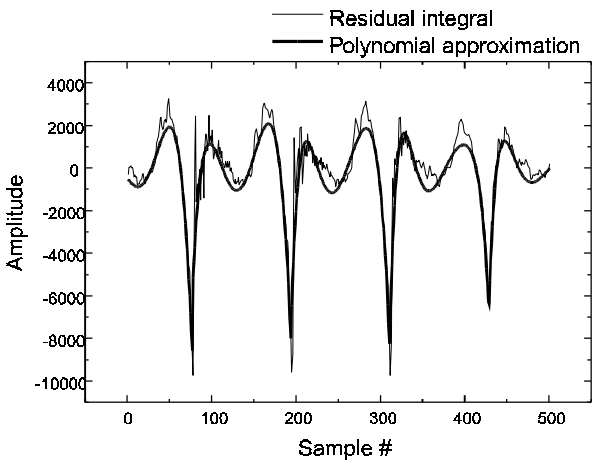

Figure 1. Polynomial approximation of the residual integral (6th order).

1.2. Equalisation of the synthesised waveform to the original speech by adaptive filtering

An optimum Wiener filter (FIR) is calculated and used on the synthesised speech to improve the final quality at the output.

This is equivalent to adaptive filtering because the optimum filter changes with each segment. The

order of the filter is fixed (between 30 and 50), and determined heuristically.

The LPC filtering and FIR filtering steps can be interchanged so the effect is that the FIR filter is modelling the stochastic part of the excitation, completing the polynomial source model. This last configuration is preferred because the LPC filter can smooth out the discontinuities originated by sudden changes in the equalisation coefficients.

Some alternatives to this method for modelling of the stochastic component have been proposed, e.g. [5], though they don't provide the flexibility needed for prosody modification. In the proposed method the model is unique for stable segments of speech, independently of the fundamental frequency.

1.3. Segmentation of the excitation signal

The original signal was segmented using a simple normalised measure of spectral change in the original waveform given by:

$$\frac{\sum_{\omega}(|S_1(\omega)| - |S_2(\omega)|)^2}{\sum_{\omega}|S_1(\omega)|^2} \geq threshold$$

For each segment, a single period is chosen from the middle region, (i.e., away from the transitions). The selected period is used to compute the coefficients of the excitation polynomial. All the samples in the segment are used to compute the equaliser parameters. The polynomial coefficients are repeated for every pitch period in the segment.

2. NEW UNVOICED SEGMENTS PROCESSING

Binary detection of voicing (detection of voicing boundaries) is responsible for an important number of errors in LPC synthesis, and methods that are blind to this property of the speech signal are highly preferred. This fact motivated us to try to use the same model that we employed for voiced segments for these fricative sounds, expecting that the post-processing filter could generate the strong stochastic component. For that purpose we extended the pitch marking over the unvoiced segments. The polynomial approximation is still performed on one "period" (selected from the middle region) and repeated all over the segment.

This replication creates a cuasi-periodic signal of very low energy (the polynomial approximation tends to the DC component) but the adaptive post-processing filter can easily mask it.

The segmentation procedure was slightly modified because no single threshold can be found that satisfies a stability criterion for voiced segments and also groups together the periods that belong to an unvoiced sound. The modification consisted of the following steps:

- The signal was segmented using the original threshold.
- If some consecutive one-period length segments were found, the threshold was iteratively increased until they are grouped together.
- Isolated periods were merged to the most similar adjacent segment.

When we applied the aforementioned model to the unvoiced segments, a lack of frication was observed (as shown in figure 2).

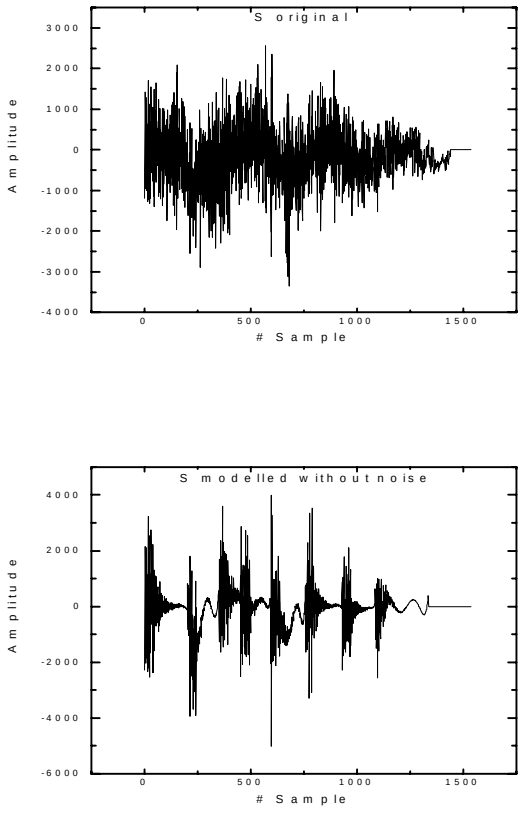

Figure 2.The top figure corresponds to the original s while the bottom one is the s modeled without noise

In order to solve this problem we added a gaussian white noise source to the model including a noise gain in the parameter set. This new variable was determined as follows:

- As we calculated the polynomial approximation we estimated the mean squared error between the first derivative of the glottal flow and its approximation.
- If this error was higher than an empirical threshold we considered that it was basically noisy and we associate a non-zero noise gain with this segment. This gain is proportional to the error.

New results are shown in figure 3.

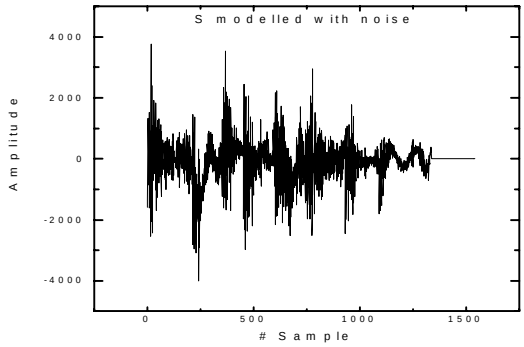

Figure 3. S modeled with noise

The final quality of the unvoiced synthetic segments is similar to that obtained by the old approximation but this method is preferred, as we pointed before, because no voiced/unvoiced detection is needed and we have a general model for voiced, unvoiced and transition segments. We point out the improvement on the transitions because of the mixture of periodicity and noise. Therefore there is no lack of modelling for voiced fricatives, which is important in a general model.

3. PROSODY MODIFICATION.

In order for this method to be useful in the context of text-to-speech applications, a flexible prosody modification scheme should be developed that allows fundamental frequency modification as well as segmental duration control.

We present some results concerning pitch modification within our model:

- A new set of pitch marks is generated that follows the synthesis intonational pattern and a correspondence between the analysis and synthesis axes is computed that maps every pitch mark from an axis to one pitch mark on the other. This mapping implicitly creates segmentation on the synthesis axis.
- For every segment in the synthesis axis there is a polynomial function describing the glottal pulses. This functional description is used to generate all the synthetic periods in the segment by time stretching/compressing the glottal pulse linearly.
- The coefficients for the adaptive filter that models the stochastic component of the signal are copied directly from the equivalent segment in the analysis axis and kept constant within the segment.

4. TEST AND RESULTS

Significant data compression is achieved with this model. While the PCM diphones have a size of 2Mb that coded with the model have a size of 570kb and celp-coded diphones occupy 315kb.

Though prosody modification using this model remains to be formally tested, we present the results of a bottom-line test. The sentences used were generated using TD-PSOLA on a set of dyphones coded using two different methods: a 8KHz-bandwidth CELP coder (as described in [8]) and the basic method described in this paper.

The quality obtained is clearly superior to that of standard LPC methods, at least for the voiced sounds.

The test proceeded as follows: ten sentences were synthesised using both the CELP coded and the new model coded dyphones. Each pair of sentences was presented in a random order and the subject asked to evaluate which one was better or if they were the same. The results for ten listeners are shown in the following table.

As we can observe from these results, there is a slight tendency in favour of the CELP-coded version, though it is not statistically relevant for this number of listeners.

	Indistinct	Our model	CELP
1	15%	25%	60%
2	20%	25%	55%
3	35%	45%	20%
4	5%	15%	80%
5	35%	45%	20%
6	35%	20%	45%
7	35%	40%	25%
8	30%	45%	25%
9	70%	5%	25%
10	55%	15%	30%
All	33.5%	28%	38.5%

5. TTS SYSTEM

The synthesis model introduced in this paper has been integrated in a full concatenation-based text-to-speech system, where the text processing and prosody generation modules are the same as described in [6][7]. The unit inventory has 455 diphone-like units, which are encoded using the algorithm described previously. Spectral smoothing in the unit boundaries is implemented as LAR coefficients smoothing over two signal periods, one on each side of the concatenation point. Subjective listening evaluation of this system remains to be done, as well as a study on quantization of the system parameters.

6.CONCLUSIONS

In this paper we presented several improvements to our earlier source model that make it suitable for both high quality analysis-synthesis [1] and text-to-speech applications.

The bigger gains in quality came from the introduction of an explicit gaussian noise generator in the model plus a generalised concept of stability.

The model was successfully applied to our concatenative synthesiser, generating high quality speech.

The presented model has shown its suitability for extracting the speaker characteristics in copy-synthesis as well as for text-to-speech tasks. These two facts make it appropriate for introducing source information in a voice conversion system.

REFERENCES

[1] J.M. Gutiérrez Arriola, F.M. Giménez de los Galanes, M.H. Savoji. *"Improvement of the quality of speech synthesis by analysis using segmentation and modelling of the excitation signal"*. EUROSPEECH'95. pp1097-1100 Madrid.

[2] P. Hedelin. *"High quality glottal LPC-vocoding"* ICASSP 86. Vol 1 pp 465-470. Tokio

[3] F.M. Giménez de los Galanes, M.H. Savoji, J.M. Pardo. *"Marcador automático de excitación glotal"*. Proc. URSI 93: 189-193. Valencia.

[4] P.H. Milenkovic. *"Voice source model for continuous control of pitch period"*. JASA 93: 1087-1096. 1993.

[5] D.G. Childers, H.T. Hu. *"Speech synthesis by glottal excited linear prediction"*. JASA 96(4): 2026-2036. October 1994.

[6] J.M. Pardo et all. *"Spanish text-to-speech: from prosody to acoustics"*. International Conference on Acoustics. 1995.

[7] P.J. Moreno, M. Martinez, J.M. Pardo, J. A. Vallejo. *"Improving naturalness in a text to speech system with a new fundamental frequency algorithm"*. EUROSPEECH'89 vol. 1 pp 360-363. Paris 1989.

[8] F. M. Giménez de los Galanes. *"Síntesis de voz de alta calidad en castellano"* PhD thesis. ETSIT Universidad Politécnica de Madrid.1995