Uniform Concatenative Excitation Model for Synthesising Speech without Voiced/Unvoiced Classification

João P. Cabral

School of Computer Science and Informatics, University College Dublin, Ireland
joao.cabral@ucd.ie

Abstract

In general, speech synthesis using the source-filter model of speech production requires the classification of speech into two classes (voiced and unvoiced) which is prone to errors. For voiced speech, the input of the synthesis filter is an approximately periodic excitation, whereas it is a noise signal for unvoiced. This paper proposes an excitation model which can be used to synthesise both voiced and unvoiced speech, thus overcoming the problem of degradation in speech quality caused by those classification errors. Basically this model consists of representing two contiguous segments of the residual signal pitch-synchronously. The first segment is represented by the original residual in a fraction of the period around the pitch-mark (obtained using an epoch detector), in order to capture the most important aspects of the residual during voiced speech. Instead, the remaining part of the period is modelled by a set of parameters of the amplitude envelope of the residual waveform and its energy. The technique for synthesising the excitation combines these shaping parameters with a novel method for regeneration of the residual waveform and a method to mix a periodic signal with noise based on the Harmonic plus Noise model. Besides producing high-quality speech, this technique is computationally fast.

Index Terms: speech synthesis, excitation model, uniform concatenative model

1. Introduction

Speech can be produced by shaping the vocal tract filter on the excitation signal representing the source, according to the source-filter model. In this model, the speech signal is generally classified into two different classes, *voiced* and *unvoiced*, depending on whether it is excited by a quasi-periodic source signal or noise respectively. There are several methods to estimate the model components, such as the popular Linear Prediction (LP) analysis. In this method, the synthesis filter is defined by the LP parameters and the source can be represented by the LP residual, which is calculated by inverse filtering the speech.

The residual waveform is often used in voice transformation applications. For example, it is possible to modify the pitch or voice quality of *voiced* speech by time-scaling segments of the residual waveform, such as in [1]. The great advantage of using the residual waveform is to preserve details of the source which are important for speech quality. For example, in [2] the authors keep the segment around the instant of maximum excitation (epoch) unchanged and modify the other segments pitch-synchronously to transform the pitch. However, a parametric model of the source is often employed for transformation of voice aspects which requires a higher degree of parametric flexibility. This model of the source is also very important for several speech processing applications which use speech modelling methods, such as in speech coding and text-to-speech synthesis.

The simplest model of the source consists of using a periodic pulse train to synthesise voiced speech and white Gaussian noise for unvoiced, such as in the traditional implementation of the Linear Prediction Coding (LPC) vocoder [3]. One of the problems of this model is that the impulse train does not represent other important aspects of the source than the periodicity characteristic and the resulting speech quality is poor. High-quality speech vocoders use more advanced excitation models for synthesising voiced speech which try to better approximate the excitation to the residual signal. For example, a popular speech coding technique is to use a combination of pulse positions and amplitudes, such as the multipulse excited linear prediction coding (MELPC) [4]. Several models have also been proposed to improve the representation of the residual in voice transformation and parametric text-to-speech applications. For example, the Deterministic plus Stochastic Model (DSM) [5] which combines a pitch-synchronous residual frame with a noise component, was proposed to improve the quality of HMM-based speech synthesis and pitch-scale transformations. However, speech synthesis methods based on the source-filter model rely on a robust voiced/unvoiced (V/UV) classification of speech. For example, if a speech frame is incorrectly classified as unvoiced its excitation is modelled by Gaussian noise only which causes deterioration of speech quality. This is an important problem in speech processing applications because the automatic classification methods are not completely reliable and their performance depends on several factors, such as the background noise of the speech recordings and the speaker's voice characteristics. It is also affected by the type of speech sound, because it is usually more difficult to classify speech sounds which are a mix of voiced and unvoiced, such as fricatives.

This paper proposes a source-filter model, named Uniform Concatenative Excitation (UCE), which does not require a V/UV decision. This model divides the signal into two contiguous parts pitch-synchronously. Another advantage, compared with a fully parametric model, is that it permits to represent part of the excitation by the residual waveform (to capture important details which are difficult to model) and the other by parameters.

2. Uniform Concatenative Excitation Model

2.1. General Description of the Model

The UCE model divides a residual signal into two parts using pitch-marks and represents these parts separately. The pitch-marks are obtained by detecting the epochs, which correspond approximately to maximum amplitude peaks in the residual waveform. The first part represents a segment of the residual, $x_a(t)$, which is located around the pitch-mark and has duration

equal to a fraction of the period T_0, while the second part represents the remaining segment till the end of the period, $x_b(t)$.

The first segment $x_a(t)$ is not parameterised, in order to capture well the important information of the residual in the region around the epoch, during voiced speech. For example, the perception of pitch is directly related to the duration between contiguous epochs. There are also other details of the source in this region which are important for the perceptual characteristics of voiced speech, such as the decaying and rising slopes just before and after the epoch, respectively. Figure 1 shows examples of $x_a(t)$ and $x_b(t)$, for voiced and unvoiced speech.

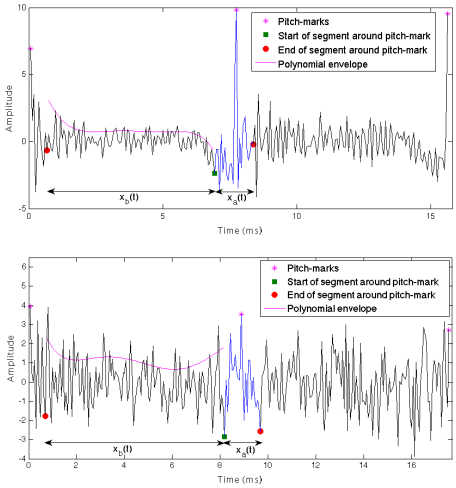

Figure 1: Example of the two residual segments represented by the UCE model in a voiced region (top) and unvoiced (bottom).

In contrast to the "original" segment, $x_b(t)$ is modelled by parameters representing the amplitude envelope of the residual waveform and its energy. The assumption is that these parameters can model well the shape of the residual waveform for both voiced and unvoiced speech. For example, the waveform of the unvoiced residual is approximately flat and unpredictable, and its power is approximately constant. Conversely, in the voiced residual the amplitude envelope is characterised by a non-flat pattern and the power of $x_a(t)$ is usually higher than that of $x_b(t)$. These differences can be observed in Figure 1.

Figure 2 shows the general block diagram of the speech synthesis method. The first part of the excitation represented by the original waveform is concatenated with a segment synthesised using the model parameters and scaled in energy by using a measure of the power ratio between the two segments.

2.2. Speech Analysis

The speech analysis is performed pitch-synchronously using the epochs estimated using the SEDREAMS algorithm [6]. Epochs are not defined in segments of unvoiced speech, but the peaks detected by SEDREAMS in these segments are also used in the UCE model. By using all the peaks, the problem of discarding an epoch, which was correctly detected, due to a voicing classification error is avoided. Meanwhile, LP analysis of order 18 is performed by using a Hamming window which is 25 ms long and centered at a fixed distance from the epoch (around 1ms). The LP coefficients are then used to calculate the residual signal by inverse filtering the speech.

For obtaining the segment $x_a^p(t)$ of the UCE model, its

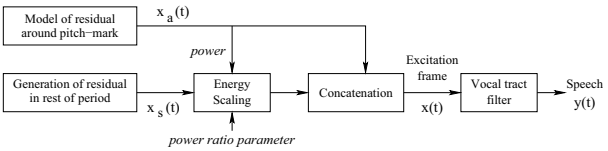

Figure 2: *Speech production using the UCE model.*

starting instant t_i^p is calculated as the first negative peak in a fixed time interval centered at the epoch p. This interval is set equal to 1.3 ms (derived experimentally), because it is sufficiently long to capture the peak with minimal amplitude in the neighborhood of the epoch and short enough for modelling the decay of the amplitude envelope. Meanwhile, the end of this segment, t_f^p, is calculated by detecting the last zero-crossing in the same time interval. This permits to avoid an amplitude discontinuity in the concatenation of this segment with the following synthesised part of the excitation, which also starts at a zero crossing. Figure 1 shows an example of the detected instants.

The next analysis step is to parameterise the segment $x_b^p(t)$, which is defined between t_f^{p-1} and t_i^p. The amplitude envelope is estimated by detecting all the local maxima in this segment and calculating a linear interpolation of these values. Then, the resulting envelope is parameterised by using a non-linear polynomial fitting algorithm. The polynomial order of six was chosen experimentally by visual inspection of the resulting polynomial functions for several segments. Finally, the power ratio between the two segments is also calculated.

2.3. Speech Synthesis

2.3.1. The Method

Speech is synthesised by passing the excitation of the UCE model through the vocal tract filter, defined by the LP coefficients, as shown in Figure 2. This excitation signal is obtained by synthesising a segment $x_s(t)$, scaling its energy and then concatenating it with the original residual segment $x_a(t)$. The block diagram of Figure 3 shows how $x_s(t)$ is generated. This method can be divided into two parts, which are described in the following sections. The first part consists of the generation of a signal $x_r(t)$ with an approximately flat amplitude envelope, which is then shaped by the polynomial curve of the model. The other step is the mixture of this signal with a noise component based on the Harmonic plus Noise model (HNM).

2.3.2. Algorithm for Regeneration of the Residual Waveform

Initial experiments conducted in this work indicated that synthesising the segment $x_s(t)$ by shaping the polynomial envelope on white noise produced speech which sounded excessively noisy in the voiced regions, although this model was good in the unvoiced regions. This result reflected the problem in the UCE model of generating a waveform with flat amplitude envelope that would be less noisy for voiced speech but that could model well the excitation of unvoiced speech. For solving this problem, an iterative algorithm was developed to generate this waveform from the segment of the original residual $x_a(t)$.

Figure 4 shows the block diagram of this algorithm. In the first iteration, $x_a(t)$ is padded with zeros and passed through the vocal tract filter to generate a synthetic speech segment. Next, this segment is inverse filtered to produce a signal which has non-zero energy in the extended part. This signal is expected to be a better approximation of the original residual than zero sam-

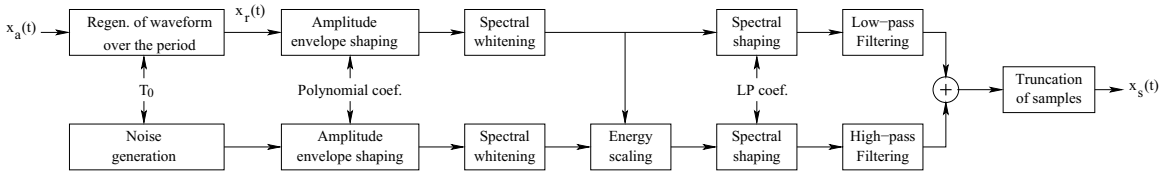

Figure 3: *Block diagram of the method to generate the synthetic part of the excitation.*

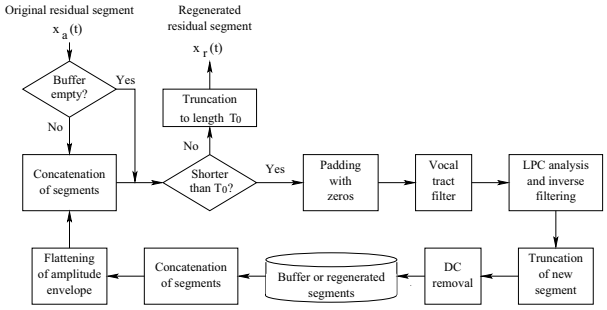

Figure 4: *Block diagram of the algorithm to regenerate the residual waveform in the UCE model.*

ples, because it uses the signal $x_a(t)$ of the original residual to predict the missing samples using the LP coefficients. That is, a speech sample generated by the LP synthesis filter of order n is a linear weighted combination of the past n speech samples plus a prediction error, which is the residual. The assumption is that the amplitude of the original residual is sufficiently small in the zero region to obtain a good approximation of the speech signal in this region. For voiced speech, the residual segment $x_b(t)$ has relatively small amplitude compared with the segment $x_a(t)$, which strengthens this assumption. On the other hand, since the residual of unvoiced speech can be represented by a random sequence, the prediction of the missing part from the original residual is expected to be less important. However, the error in the regeneration of the residual by inverse filtering is expected to propagate and increase along the zero region. Also, the energy of the resulting signal decays along the time. These problems are avoided by truncating the regenerated residual signal to obtain a segment that starts at t_f^p with shorter duration than the period. This duration was set equal to 1.5 ms by assuming that the length of this segment (24 samples at $F_s = 16\ kHz$) should not be much higher than the LPC order and based on visual comparison of the regenerated and original residual segments. Then, this segment is subtracted by its mean value and it is stored in a buffer to be concatenated with the other segments regenerated in the subsequent iterations. The amplitude of the resulting segment is also flattened by multiplication with a curve, which is the inverse of the envelope contour calculated from the local maxima of the waveform. Finally, the original segment $x_a(t)$ is concatenated to the new segment. This cycle of operations is repeated until the segment obtained from the concatenation can be truncated to have the duration of the pitch period. An example of the regenerated residual, $x_r(t)$ in the voiced region of speech is shown in the top of Figure 5.

An example of the excitation generated after envelope shaping and energy scaling of the regenerated waveform is shown in Figure 5 (bottom). The quality of speech produced using the regenerated residual signal is better compared with white noise, in voiced regions. However, the quality of the synthetic speech in unvoiced regions becomes poor using this technique. The explanation is that the regenerated signal has more periodicity than white noise in these regions. In order to overcome this problem, the HNM is used to mix a noise component with the component of the regenerated residual.

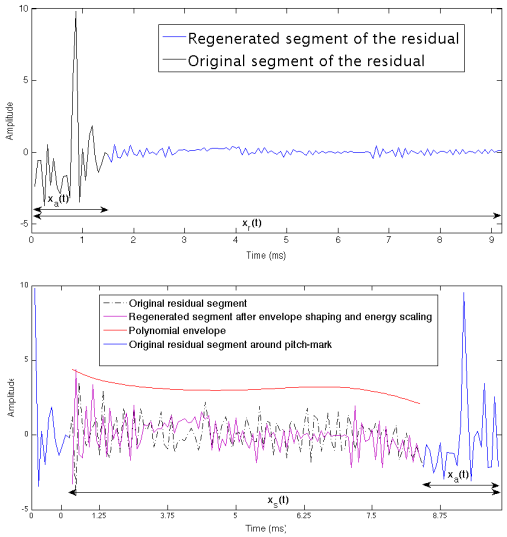

Figure 5: Example of the regenerated residual waveform (top) and a segment of the excitation synthesised by shaping the envelope of the regenerated signal and scaling its energy (bottom).

2.3.3. Harmonic plus Noise Model

In the HNM [7] the spectrum of the speech signal is divided into two frequency bands by the voiced frequency parameter F_m. The low- and high-frequency parts are represented by a harmonic and a modulated noise signal, respectively. Typically, the noise is modulated in the time-domain pitch-synchronously using a function such as the triangular window. In this work, the HNM was implemented using a constant $F_m = 4$ kHz (speech sampled at 16 kHz), for simplicity.

The HNM was incorporated into the UCE model as shown in Figure 3. The periodic component of the HNM is obtained from the regenerated signal $x_r(t)$, whereas the noise component is obtained from white noise. Both signals, which have duration T_0, are first shaped in amplitude by the polynomial envelope in the region that starts at the end of the "original" part of $x_r(t)$ till the end of the period. A spectral whitening operation is also performed on the resulting signals, which permits to approximate their spectral envelope to that of the original residual and to scale the energy of the noise to be the same as the periodic component, based on the assumption that the residual

has an approximately flat spectrum. In order to better model the two signals in the frequency-domain, their spectral envelope is shaped by the LP spectrum of the residual afterwards. Then, the periodic and noise components are low and high-pass filtered, respectively, and added together. The resulting signal is truncated, in order to be scaled in energy and concatenated with the original residual segment, as shown in Figure 2. However, since the two segments have different spectra, the spectral whitening and shaping is also performed on the concatenated signal to better model the spectrum of the excitation after the concatenation.

3. Perceptual Evaluation

3.1. Baseline LPC vocoder with HNM

The speech synthesis method using the UCE model was compared with a baseline vocoder which used a similar implementation of HNM, with the exception that the periodic component was represented by the impulse train and the aperiodic component was white noise. These components were also shaped by the spectral envelope of the residual, before band-pass filtering.

The V/UV classification of the LPC vocoder was performed using the pitch-tracking algorithm (called RAPT) of the ESPS/Waves+ tools [8]. This detector was chosen because it performed well for the speech data used in this experiment and it enabled the control over the amount of V/UV detection errors by changing the value of a voicing probability parameter. Speech was synthesised twice using this vocoder. In the first case, the voicing parameter (range from 0 up to 1) was manually tuned to obtain the best V/UV classification, by visualization of the resulting F0 contours and the speech waveform. In the other case, the voicing probability parameter was decreased by 0.4 to intentionally increase the V/UV classification errors and simulate the situation in which the V/UV errors have an important effect on degradation of speech quality.

3.2. Experiment

The recorded speech consisted of six sentences of the CMU ARCTIC speech database (US English BDL male voice) [9]. The sentences were synthesised using the method described in Section 2.3, which uses the UCE model, and the LPC vocoder with the two types of V/UV classification respectively.

The evaluation was conducted via the web and eight listeners participated in the experiment, of which three were native speakers of English. It was divided into two parts: the pairwise comparison and Mean Opinion Scores (MOS). In the MOS part, listeners heard one utterance and chose a score which represented how natural or unnatural the sentence sounded on a scale of 1 ("Completely Unnatural") to 5 ("Completely Natural").

3.3. Results

The results of the pairwise comparisons and MOS are shown in Figure 6. Speech synthesised using the LPC vocoder with reduced V/UV detection errors (by manual tuning) was preferred over speech synthesised with the UCE model, on average. However, when the amount of V/UV classification errors of the LPC vocoder was higher the UCE model obtained better results. These results show the advantage of using the UCE model, when the V/UV decision errors of the baseline produce significant speech distortion. The MOS results also show that the LPC vocoder can produce speech which sounds better than using the UCE model, but the speech naturalness of the two methods was similar when the V/UV decision errors increased.

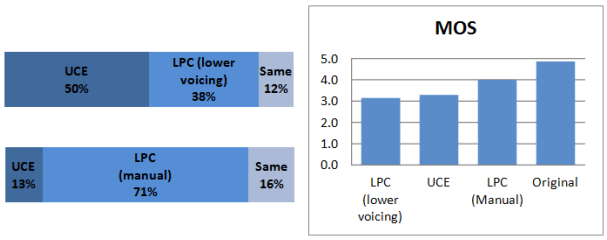

Figure 6: *Preference rates and MOS obtained for the the speech synthesis methods based on the UCE model and LPC vocoder.*

From informal perceptual evaluation of speech quality by the author, the UCE can produce segments of speech which sound very natural and close to the quality of the recorded speech. However, sometimes it produces voiced segments which sound clearly noisier than the original. Currently, the HNM component of this method is being improved, by estimating its F_m parameter from the speech signal (instead of using a constant value), in order to reduce this noise distortion.

4. Conclusion

The V/UV classification has always been a limitation for synthesising speech using a source-filter model. The reason is that it is not possible to ensure that automatic V/UV detection is always reliable and its performance affects the speech quality. The UCE model of the source is proposed in this paper for synthesising speech without the need for a V/UV decision. Basically, this model consists of concatenating two consecutive segments which represent the residual pitch-synchronously. The first segment represents the original residual waveform in a short region around the pitch-mark (obtained for both voiced and unvoiced speech). Meanwhile, the second is modelled by parameters of the amplitude envelope of the waveform and energy. This part can potentially provide parametric flexibility for voice transformation and enables a more compact representation of the residual for speech coding and synthesis applications.

The UCE model was compared with a baseline LPC vocoder using a mixed excitation model, in two different conditions. In one case, the V/UV classification of the LPC vocoder was manually tuned to avoid the effect of voicing errors on the speech quality, whereas in the second those errors were artificially increased. The UCE model did not obtain as good speech quality as the baseline in the first condition. However, on average, it was preferred over the baseline when the V/UV errors became a significant problem for this system. These results show that the UCE model is a solution for overcoming the problem of V/UV classification in speech synthesis application, but there is scope for further improvements of this synthesis method. More extensive experiments will also be conducted in the future, such as an evaluation of the UCE against the baseline using different V/UV classifiers and its robustness to epoch detection errors.

5. Acknowledgements

This research is supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (`www.cngl.ie`) at University College Dublin. The opinions, findings and conclusions, recommendations expressed in this material are those of the authors and do not necessarily reflect the views of Science Foundation Ireland.

6. References

[1] Cabral, J. P. and Oliveira, L. C., “Pitch-synchronous time-scaling for prosodic and voice quality transformations”, Proc. of INTERSPEECH, 1137–1140, Lisbon, 2005.

[2] Rao, K.S. and Yegnanarayana, B., “Prosody modification using instants of significant excitation”, IEEE Transactions on Audio, Speech, and Language Processing, 14(3), 972–980, 2006.

[3] Deller, J. R., Proakis, J. G. and Hansen, J. H., “Discrete Time Processing of Speech Signals”, Macmillan, New York, USA, 1993.

[4] B.S. Atal, B.S. and Remde, J.R., “A New Model of LPC Excitation for Producing Natural Sounding Speech at Low Bit Rates”, Proc. ICASSP, Paris, 614–617, 1982.

[5] Drugman, T. and Dutoit, T., “The Deterministic plus Stochastic Model of the Residual Signal and its Applications”, IEEE Transactions on Audio, Speech and Language Processing, 20(3), pp. 968–981, 2012.

[6] Drugman, T. and Dutoit, T., “Glottal Closure and Opening Instant Detection from Speech Signals”, Proc. of INTERSPEECH, Brighton, U.K., 2009.

[7] Stylianou, Y., “Harmonic plus Noise Models for Speech, combined with Statistical Methods, for Speech and Speaker Modification”, PhD thesis, Ecole Nationale Sup‘erieure des Telecommunications, 1996.

[8] Talkin, D., “A robust algorithm for pitch tracking (RAPT)”, in Speech Coding and Synthesis, Elsevier Science, 495–518, 1995.

[9] Kominek, J. and Black, A., “The CMU Arctic speech databases”, in Proc. of 5th ISCA Speech Synthesis Workshop (SSW5), Pittsburgh, USA, 2004.