IMPROVING THE MODELING OF THE NOISE PART IN THE HARMONIC PLUS NOISE MODEL OF SPEECH

Yannis Pantazis , *Yannis Stylianou*

Institute of computer Science, FORTH, Crete, Greece, and
Multimedia Informatics Lab, Computer Science Department, University of Crete, Greece
email: {pantazis, yannis}@csd.uoc.gr

ABSTRACT

Harmonic + Noise model (HNM) is a hybrid model of speech with a harmonic component and a noise component. While harmonic part describes efficiently the periodicities in speech signals (voiced parts), modeling of noise part introduces artifacts primarily because of the specific time-domain characteristics of noise in voiced speech. In this paper, we are concentrated on the modeling of noise in voiced frames. To model the temporal characteristics of noise, we study three time envelopes in the context of HNM; Triangular envelope, Hilbert envelope and Energy envelope. Listening tests showed a clear preference on the Energy envelope and Hilbert envelope for male voices and in a less extent the same conclusions can be drawn for female voices.

***Index Terms*—** Speech Synthesis, Noise modeling, Time envelope, Energy Distribution

1. INTRODUCTION

A usual approach for applications in speech analysis, synthesis, and coding is to split speech into components [1] [2]. These models are usually referred to as hybrid models because they suggest the modeling of speech using components with different statistical properties (both in time and frequency). In many parametric representations of speech, there is a quasi-periodic component which is usually modeled as a sum of harmonically related sinusoids. Then, there is another component, sometimes referred to as stochastic component, to model the non-periodic characteristics of speech. This component is usually modeled by white noise after appropriate processing (modulation in frequency and time). Note that voiced speech usually contains both parts.

One well known hybrid model for speech is HNM which was developed by Stylianou et al. [3] and it is used for high quality time/pitch scale modification of speech and voice transformation. HNM has also been suggested for speech synthesis [1]. HNM decomposes speech into two bands; the lower band where the signal is modeled as a sum of harmonically related sinusoids and the upper band where the signal is modeled by colored noise. While HNM produces very good quality of speech, a background noise is sometimes perceived. This is mostly noticeable for the case of male voices. We believe that a major source of this background noise comes from the modeling of the noise part in HNM, and especially from the modulation in time of this component. In this paper, we suggest two alternative ways to the standard approach suggested in [4] and [3] for the time-domain modulation of the noise part. Furthermore, we have conducted a listening test to compare the three approaches.

As far as the frequency characteristics of the noise part are concerned, the use of techniques based on linear prediction provide satisfactory results [4]. However, if the time characteristics of noise are not taken into account, then the noise part is not fused into the harmonic part, and then a second source of background noise is perceived [4] [5]. An example of the time characteristics of the noise part is depicted in Fig. 1. The upper panel the original speech signal which is sampled at 16kHz, while the lower panel in this figure shows the result after highpass filtering of the original signal with a cutoff frequency of 4kHz. It is important to note the time-domain structure of the noise part which is synchronized with the pitch period of speech. For voiced frames, this time-domain characteristic of noise is partially derived from the turbulence and the friction noise that is produced at the time instants of opening and closing of vocal folds. This effect is more prominent in the male voices and for voiced fricatives like /z/ and /v/. The example in Fig. 1 has been extracted from the /z/ sound uttered by a male speaker.

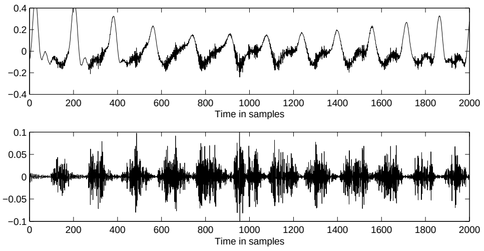

Fig. 1. Upper plot shows 12 pitch periods of voiced fricative phoneme $/z/$. Lower plot shows the same speech signal filtered by a highpass filter at 4kHz (noise component). Obviously, the energy of the noise part is not distributed uniformly.

To reconstruct the time-domain characteristics of the noise part, a time-domain envelope is usually used. In [4], a deterministic triangular-like envelope has been used, a solution which sometimes fails to provide satisfactory results, especially for male voices as it has been mentioned previously. There are two main drawbacks by using a deterministic envelope. First, it may not follow the perceptually important time-characteristics of the high frequency signal. Second, while the position of the envelope is fixed during the duration of a pitch period, the harmonic part is moving inside the same interval. Another solution has been suggested by McCree et al. [5] for the expansion of narrowband speech to the wideband speech. In [5] the short-time energy of the signal corresponding to the frequency band above 3 kHz (3-4 kHz) is used as a time modulator of the high band noise. Using about the same frequency band, we construct an envelope as the instantaneous amplitude of the analytic signal (through Hilbert transform) that corresponds to the signal contained in this band.

In this paper, we suggest a third envelope that is obtained by developing in Fourier series the short-time energy of the noise part. To compare the three envelopes, we have conducted a listening test using high quality recordings for male and females voices.

This paper is organized as follows; a brief review of HNM is presented in Section 2. Then, the three different time-domain envelopes are described in Section 3. Section 4 describes the listening test that was used for the evaluation of the effect of different envelopes, while, in Section 5 results and conclusions are thoroughly discussed.

2. HARMONIC + NOISE MODEL

HNM decomposes speech into two components: a harmonic component and a noise component. HNM analysis (and synthesis) is performed in a frame-by-frame basis in a pitch synchronous way. Pitch synchronous analysis, in the HNM context, it only means that the distance between two consecutive analysis time instants is equal to one local pitch period and the length of the analysis window is an integer multiple of the local pitch period. Analysis time instants are not, however, related to any reference time instant of speech like glottal closing instant (GCI). Depending on the voicing decision a frame is labelled either as voiced or unvoiced. If a frame is unvoiced, the signal is only modeled by the noise part as an AR process. When the frame is voiced, then speech is modeled as the sum of two components:

$$s[n] = h[n] + u[n] \tag{1}$$

where $h[n]$ is the harmonic part and $u[n]$ is the noise part. Harmonic part is described by a sum of harmonically related sinusoids:

$$h[n] = \sum_{k=-L}^{L} a_k e^{j2\pi k(f_0/f_s)n} \tag{2}$$

where L denotes the number of harmonics, f_0 denotes the fundamental frequency, a_k are the complex amplitude of the kth harmonic and f_s is the sampling frequency.

The direct estimation of the unknown parameters (f_0, L, a_k) is a nonlinear problem, therefore it is broken into two subproblems. First, the local fundamental frequency, f_0, and maximum voiced frequency, f_{max} are estimated based on an analysis-by-synthesis scheme [4]. The number of harmonics, L, is given by $L = \lfloor \frac{f_{max}}{f_0} \rfloor$ where $\lfloor \cdot \rfloor$ denotes the floor operator. The estimation of the unknown complex amplitudes, a_k, is obtained by minimizing a weighted time-domain least-squares criterion with respect to a_k,

$$\epsilon_\alpha = \sum_{n=-T}^{n=T} w^2[n](s[n] - h[n])^2 \tag{3}$$

where $s[n]$ denotes the original speech signal, $h[n]$ denotes the harmonic signal, $w[n]$ denotes the weighted window (which usually is a Hamming window) and T denotes the local fundamental period in samples, $T = (f_s/f_0)$.
Noise part, $u[n]$, is modeled as:

$$\hat{u}[n] = e[n](u_G[n] * q[n]) \tag{4}$$

where $u_G[n]$ denotes a white Gaussian noise process filtered by an AR filter with impulse response $q[n]$ and $e[n]$ is the time-domain envelope for the time modulation of the colored noise.

3. ANALYSIS OF NOISE ENVELOPES

There are various methods to obtain the envelope of a signal, most common being through the analytic signal. The analytic signal is obtained using the Hilbert transform. Another possibility which is sufficient for our purpose and much faster than Hilbert transform is to obtain the envelope by filtering the absolute value of $u[n]$ with a moving average filter of order $2N + 1$,

$$e[n] = \frac{1}{2N+1} \sum_{k=-N}^{N} |u[n-k]| \tag{5}$$

In Fig. 2, the same noise part shown in Figure 1 (lower panel) is depicted along with the corresponding envelopes estimated by (5) using $N = 7$.

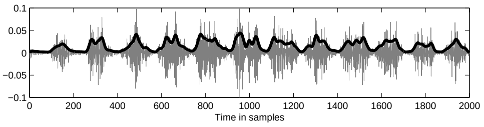

Fig. 2. The envelope of the noise component.

3.1. Triangular Envelope

A triangular-like envelope for the time-domain modeling of the noise part has been proposed in [4]. This envelope is depicted in Fig. 3 and it is controlled by four parameters. However, assuming that the envelope is symmetric and that we are only interested in the relative amplitude i.e. only in A_0/A_1, T_1 is computed from T_0 and we set $A_1 = 1$. As we will see later, similar normalization is performed to the other types of envelope. So, we have to estimate two variables, T_0 and A_0. In [4], the design parameters have been specified ad hoc: $T_0 = 0.15\,T$ and $A_0 = 0.5$, where T is the local pitch period.

Such a deterministic approach may degrade the synthesized signal by introducing background noise, as it was already mentioned previously. In different terms, the main reason for this degradation is that different phonemes have different (T_0, A_0) pairs, as well as, different speakers are also observed to have different (T_0, A_0) pairs. Actually, if the same person utters the same phoneme twice we may observe differences in the pair (T_0, A_0). Moreover, the location where the envelope is placed inside a pitch period is fixed (i.e., at its center) independently of the position of the harmonic part. Therefore, the fusion of the noise part into the harmonic part is not always achieved.

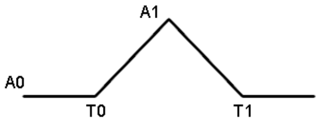

Fig. 3. The deterministic time-domain envelope for the noise part.

3.2. Hilbert Envelope

In [5] McCree used the spectral information of 3-4kHz for the expansion of the speech spectrum to the upper frequency bands. We use the information of approximately the same band for the computation of the time envelope of noise. To obtain the contents of the band, the M highest harmonics (in the harmonic part) are used:

$$\tilde{e}_H[n] = \sum_{k=L-M+1}^{L} a_k e^{2\pi k(f_0/f_s)n} \qquad (6)$$

We recall that the parameters $\{L, a_k, f_0\}$ are the same parameters described in (2). M is the number of harmonics used for the reconstruction. Signal $\tilde{e}_H[n]$ is the analytic signal of the signal contained in this frequency band and, therefore, the absolute of it, $|\tilde{e}_H[n]|$ provides its instantaneous amplitude which is then used for the modulation of the noise part in time. We will refer to this envelope as Hilbert envelope. The envelope is normalized so the largest value of it to be 1. Hilbert envelope depends on the bandwidth on which it is estimated as well as the center of this frequency band since fundamental frequency and maximum voiced frequency are time-varying parameters. This may increase the variance of the computed envelopes from one frame to the next. To cope with this, we fix the maximum voiced frequency to 4kHz.

3.3. Energy Envelope

In the Energy envelope approach, signal $u[n]$ is used to estimate the time envelope of the noise part. As we have already seen in the example depicted in Fig. 2, noise envelope, as this is computed by (5), is a smooth function over time. Therefore, the Energy envelope can be modeled by Fourier series with a few number of harmonics. Due to the fact that time modulation of noise is pitch related, the fundamental frequency of the noise envelope is selected to be the same as that of the harmonic part.

The energy envelope, $e[n]$, estimated by (5) is therefore approximated by:

$$\hat{e}[n] = \sum_{k=-L_e}^{L_e} A_k e^{j2\pi k(f_0/f_s)n} \qquad (7)$$

where L_e is the number of harmonics which is a small integer (about 3 or 4), and f_0 is the fundamental frequency of the current frame. The amplitudes are estimated using Least Squares in a similar manner to that of the harmonic part.

The major advantage of this representation of the Energy envelope is that it can approximate well the distribution of energy of the noise part while it can be easily manipulated for pitch and time-scale modifications. Similar to the other methods, normalization is performed by setting the maximum value of the envelope to 1.

To summarize, the Triangular envelope is independent of the signal and it is easily manipulated for speech modifications, the Hilbert envelope depends on the higher frequencies of the harmonic part but it is not easy to manipulate for speech modification since it is not parameterized and finally, the Energy envelope depends on the noise part if the speech signal (higher frequencies) and it is easily manipulated for speech modifications since it is parameterized.

4. LISTENING TEST

A listening test has been conducted for the evaluation of the three envelopes using two sets of speech data; the first one contains high quality recordings in the French language provided by France Telecom R&D, while the second one contains examples of speech extracted by the TIMIT database. The test is a variation of the ABX [6], chap.13, listening test. In our test, A and B denotes the two synthetic signals while X refers to the original speech signal. The major difference between the test we performed and that of ABX is the use of one more option for the listener; the option of A=B has been added, meaning that the synthesized examples are equivalent. This option has been inserted because sometimes it was difficult to perceive any difference between A and B. Moreover, the differences, if any, are in the high frequencies making the test quite difficult. Note that the harmonic part was the same for all the choices of the envelope for the noise part.

In order to perform the synthesis, we have to specify the parameters of HNM. An important parameter is that of the maximum voiced frequency. To avoid any problem with the envelope using the Hilbert approach, it was decided to set the maximum voiced frequency at 4kHz, with sampling frequency at 16kHz. Lowering more the maximum voiced frequency will introduce artifacts in the synthesized signals (quasi-harmonic frequencies will be modeled as modulated noise) posing serious problems in the evaluation of the different options for the time modulation of noise. On the other hand, increasing more the maximum voiced frequency (above 4kHz) will decrease the importance of the noise part reducing therefore the effect of the envelopes under evaluation.

The width and height of the triangular envelope was set at the same values as proposed in [4]. The 8 highest harmonics ($M = 8$) were used for the reconstruction of the envelope in case of Hilbert envelope. This means that for the male voice the bandwidth of bandpass signal was about 800 to 1000Hz whereas for female voice the bandwidth was doubled. Finally, 4 harmonics ($L_e = 4$) were used for the approximation of the energy envelope.

The acoustic inventory used for the listening test consisted of, high quality recordings in the French language (8 sentences from one female voice and 8 sentences from one male voice), and a set of randomly selected 8 sentences with 5 male voices and 8 sentences from 6 female voices from the TIMIT database. Therefore, for each language a stimuli of 48 sentences have been presented to the listeners. The entire test battery was divided into two series of subsets; one for each language. The listening test was conducted in a quiet private walled office, using a closed type headphones and high quality equipment for digital to analogue conversion. Listening tests were interactive, using an interface developed in Matlab for the easy access to the sound examples and decisions recording. Listeners could listen as many times as they wanted to the speech signals before submit their preference. Two examples for each case, dissimilar and similar to the original, emphasizing the effect of the background noise was initially provided to them. The original signal has been presented without any additive noise modulation. Twelve listeners without any known hearing problem have participated in the test. The majority of the listeners were not used in speech listening tests while most of them they represented a different language background from the stimuli; however, native language is not considered relevant for this auditory task. Most of listeners reported that the test was difficult and the differences between the different versions of envelope, were not always easily noticeable. All the synthetic signals have been considered to be very good quality reconstructions of the corresponding originals.

5. RESULTS AND CONCLUSIONS

Fig. 4 shows the noise part of phoneme /z/. The best re-synthesis in terms of signal similarity (similar time-domain distribution of energy) is the one obtained by the Energy envelope (lower panel). This is expected, however, since Energy envelope models the envelope of

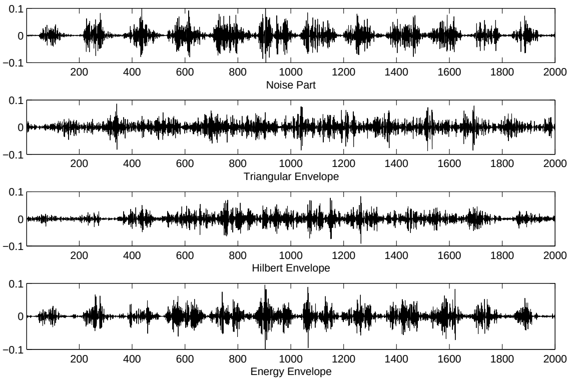

Fig. 4. A few periods of the noise part for phoneme $/z/$. First plot is the original noise part while the other three plots are the synthesized noise part. From 2nd to 4th panel, Triangular, Hilbert and Energy envelopes are presented, respectively. The closest to the distribution of energy of the original noise part is the noise produced using the Energy envelope.

	Triangular	No pref.	Hilbert
Male	8 (8.3%)	43 (44.8%)	45 (46.9%)
Female	40 (41.7%)	47 (48.9%)	9 (9.4%)

	Hilbert	No pref.	Energy
Male	22 (22.9%)	47 (49.0%)	27 (28.1%)
Female	22 (22.9%)	54 (56.3%)	20 (20.8%)

	Energy	No pref.	Triangular
Male	43 (44.8%)	50 (52.0%)	3 (3.2%)
Female	16 (16.7%)	67 (69.8%)	13 (13.5%)

Table 1. Results from the listening test for the English sentences.

	Triangular	No pref.	Hilbert
Male	10 (10.4%)	47 (49.0%)	39 (40.6%)
Female	8 (8.3%)	71 (74.0%)	17 (17.7%)

	Hilbert	No pref.	Energy
Male	11 (11.5%)	58 (60.4%)	27 (28.1%)
Female	13 (13.5%)	58 (60.4%)	25 (26.1%)

	Energy	No pref.	Triangular
Male	42 (43.7%)	48 (50.0%)	6 (6.3%)
Female	16 (16.7%)	68 (70.8%)	12 (12.5%)

Table 2. Results from the listening test for the English sentences.

the noise part as a Fourier series. Hilbert envelope partially achieves to follow the energy distribution of the noise part, while the result using the Triangular envelope presents the highest dissimilarity with the energy distribution of the noise part.

In both tables, Table 1 and Table 2, the total number of preferences as well as the average rate of them (in parenthesis) made by the listeners is provided. Note that in each row there are in total 96 preferences since 12 listeners evaluated 8 sentences in each case.

Table 1 shows the results for French. The Energy envelope has similar preference score as the Hilbert envelope, while both envelopes outperformed the Triangular envelope. However, in the case of female voice only, the Triangular envelope was clearly preferred over the Hilbert envelope while this is not the case when the comparison was performed between the Energy and the Triangular envelopes. A possible reason for this it may be the number of harmonics used in the Hilbert envelope reconstruction. For female voices the reconstructed signal covers more spectrum than in male voices which possibly results in an over-estimation of the energy of the noise part. Noticeable is also that the score of no preference which means that there is no clear preference among the different versions of the envelope is high.

In Table 2, the results for English is presented. Almost the same conclusions made for Table 1 can be drawn. However the difference in the comparison of the Hilbert and the Triangular envelope for female voices is not present here. The Hilbert envelope seems to be preferred over the Triangular window.

Overall, in all the experiments, the Energy envelope was clearly preferred in comparison to the other envelopes, while the Hilbert envelope seems to be the next choice. Especially for male voices where Triangular envelope performed poor the Energy envelope is a sufficient method for modeling the time characteristics of the noise part.

6. ACKNOWLEDGEMENT

We would like to thank the listeners who have participated in the test as well as Olivier Rosec from France Telecom R&D for his valuable comments and continuous discussion on this topic and for providing us with the high quality recordings in French.

7. REFERENCES

[1] Yannis Stylianou. Applying the harmonic plus noise model in concatenative speech synthesis. *IEEE Trans. on Speech and Audio Proc.*, 9:21–29, 2001.

[2] A.V. McCree and T.B. Barnwell. A mixed excitation LPC vocoder for low bit rate speech coding. *IEEE Trans. on Speech and Audio Processing*, 3:242–250, 1995.

[3] Y. Stylianou J. Laroche and E. Moulines. High quality speech modification based on a harmonic + noise model. *EUROSPEECH*, 1995.

[4] Yannis Stylianou. *Harmonic plus Noise Models for Speech, combined with Statistical Methods, for Speech and Speaker Modification.* PhD thesis, Ecole Nationale Supèrieure des Télécommunications, 1996.

[5] Alan McCree. A 14kb/s wideband speech coder with a parametric highband model. *ICASSP*, 2000.

[6] W.B. Kleijn K.K. Paliwal (editors). *Speech Coding and Synthesis.* Elsevier, 1995.