Optimised source signal modelling for Linear Predictive speech synthesis

A Mamombe and Bea Lacquet

Department of Electrical and Information Engineering
University of the Witwatersrand, Johannesburg, South Africa.
a.mamombe@ee.wits.ac.za

Abstract

Linear Predictive (LP) speech synthesisers still play an important role in linguistic analysis and speech processing. However, the quality of speech produced from such synthesisers still falls short of many people's expectations. This paper discusses ways of improving the quality of speech-produced by LP synthesisers through unique source signal models. Popular models of the source signal include the Rosenburg-Klatt (R-K), the triangular pulse, codebooks and the unit impulse [1]. Tests have proved that the R-K model is the most favourable [2], though it has limitations related to the processing difficulties and accounting for fricative noise. Two fairly new source signal modelling techniques that solve this problem are discussed in this paper namely 1) A linear modification of the R-K signal and 2) A modification of the Harmonic plus Noise Model (HNM) speech processing technique to model the source signal [2],[6]. Favourable results were obtained when using the HNM technique for vowel sounds.

Keywords: Linear Prediction, Source Signal Modelling, Harmonic plus noise.

1. Introduction

Linear Predictive synthesis, is a technique based on the auto-regressive model as shown in equation 1,2 [3]. The two main parameters of LP synthesis are the predictive coefficients a_k (*vocal tract filter characteristics*) and the source signal $e(n)$ (*the glottal pulse source signal*).

$$\widetilde{x}[n] = \sum_{k=1}^{p} a_k x[n-k] \tag{1}$$

$$e[n] = x[n] - \widetilde{x}[n] \tag{2}$$

$x[n]$ is the actual speech signal. $\widetilde{x}[n]$ is the predicted sample at instant n and a_1,a_2,....,a_k are predictor coefficients.

There are various methods of obtaining the filter parameters a_k and the residual signal $e(n)$ as discussed in [4]. Once the filter parameters and the residual signal *(source signal)* is known, speech can be synthesised by passing the residual signal $e(n)$ through an all pole filter with transfer characteristics shown in equation 3[4]. The filter parameters are stored in a codebook and residual signal (source signal) is either stored or modeled using the unit impulse, triangular or R-K methods [1]. Modelling the residual signal greatly reduces the need for a bigger memory but compromises quality. This paper presents a brief critical overview of the existing source signal modelling techniques. Proposed techniques for improving the quality of the source signal models are presented and discussed.

$$H(z) = \frac{1}{\sum_{k=1}^{p} a_k z^{-k}} \tag{3}$$

2. Source signal modelling

The following sections will discuss various ways of modelling the source signal accurately whilst maintaining highly natural and intelligible speech. In order to archive this an algorithm was developed in MATLAB to obtain the residual signal and LPC parameters, for the vowel /a/ shown in Fig 1.0 sampled at 8 KHz. The residual signal obtained from the MATLAB algorithm using 20 LP parameters is shown in Fig 2.0.

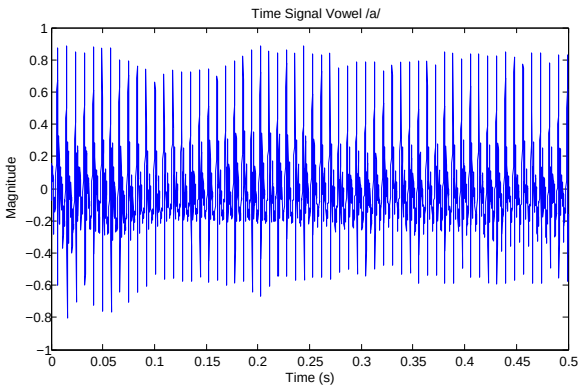

Fig 1.0 Time domain signal for vowel /a/

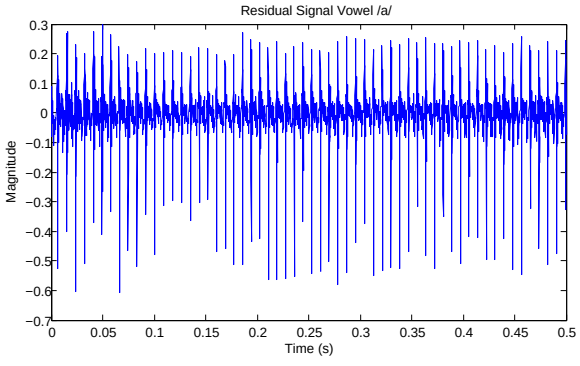

Fig 2.0 Time domain residual signal for the vowel /a/

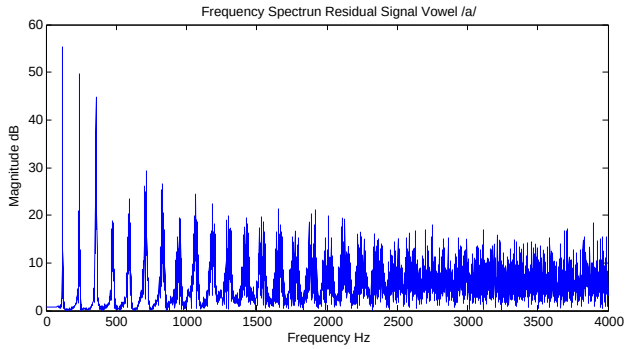

Fig 3.0 Frequency domain residual signal for the vowel /a/

Most linear Predictive (LP) synthesisers tend to simplify matters once the residual signal is obtained, by using an impulse train, R-K or a triangular pulse signal in the modelling *(source signal)* [1].

Two fairly new techniques that employ the *modified R-K and the HNM synthesis* to model the source signal are also discussed [6]. The criterion used to quantify the quality of the source signal models discussed is that ideally the model should exhibit characteristics similar to those of the actual residual signal in Fig 2.0 , 3.0 and produce intelligible speech.

2.1. Current source signal modelling techniques

In this section, we give descriptions of the current residual/source signal modelling techniques namely the triangular pulse, the unit impulse and Rosenburg-Klatt (R-K). By applying them in synthesis to the vowel /a/.

2.1.1. Impulse train

The impulse train Fig 4.0 was used to model the source signal for a vowel /a/ shown in Fig 2.0 The method produced reasonable speech quality for the vowel /a/; however, comparing the frequency and magnitude components of the signal in Fig 2.0 it is evident that the impulse train Fig 4.0 is far from the ideal residual signal.

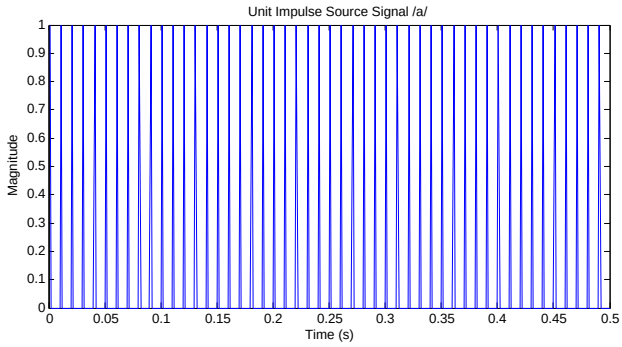

Fig 4.0 Impulse train source signal model

The pitch period To of the impulse train is derived from the pitch frequency Fo, that is the frequency of the largest harmonic in the source signal [5]. Such that $To = 1/Fo$.

2.1.2. Triangular pulse approximation

Most LPC based speech synthesisers use the triangular pulse Fig 5.0 as the source signal [1]. The triangular pulse is a good estimate of the source signal *(actual glottal pulse)* and is easier to generate unlike the R-K signal. The triangular signal in Fig 5.0 was applied as the source signal to synthesise the vowel /a/ using Linear prediction. The resulting synthetic speech produced was fairly intelligible and is further discussed in the results section.

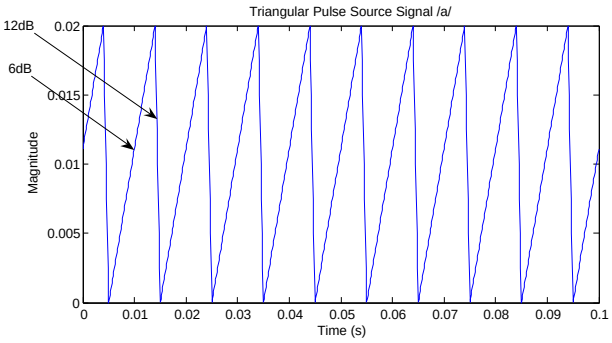

Fig 5.0 Triangular pulse source signal model

2.1.3. The R-K source signal model

Literature suggests that a better way of modelling the source-signal is the use of the R-K model [1]. Rosenburg reported that the source signal produced a more natural speech when modelled similar to the glottal excitation signal Fig 6.0. He derived a polynomial that closely modelled the glottal pulse shown in equation 4 [1]. Modern research has simplified this polynomial as a unit impulse driven through a filter or simply modelled the signal as in equation 5 [1]. The R-K signal was modified for the experiment in order to reduce the computational requirements as shown in the next section.

$$g(t) = \begin{cases} 0 & for\ 0 \le t \le t_1 \\ A(\frac{(t-t_1)}{(t_2-t_1)})^2(3-2\frac{(t-t_1)}{(t_2-t_1)}) & for\ t_1 \le t \le t_2, \\ A(1-\frac{(t-t_2)}{(b-t_2)}) & for\ t_2 \le t \le b \end{cases} \tag{4}$$

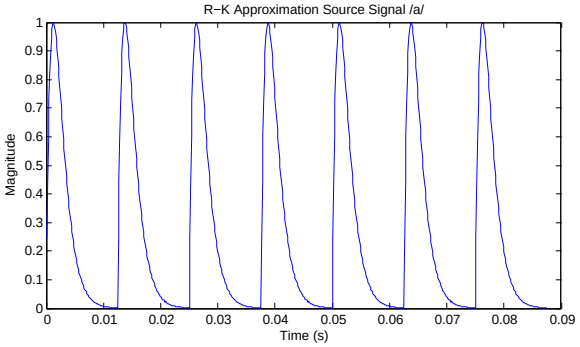

Fig 6.0 R-K Approximate source signal model

The approximate R-K equation

$$g(t) = A\frac{t}{T_0}\exp(1 - \frac{t}{T_0}) \tag{5}$$

Were T_0 is the period of the pitch frequency and $g(t)$ the source signal.

2.2. Optimised source signal modelling

The following sections of the paper describe two fairly new source signal models that the authors used for LPC speech synthesis. The first is the linear modification of the R-K signal and the second is the use of the HNM synthesis to model the source signal.

2.2.1. Modification of the R-K source signal

A new technique discussed in this paper is a linear modification of the R-K source signal. A set of linear ratios were used to simplify the computation of the signal by relating the values t_1, t_2, b *from equation 4* to the pitch period To. The ratios used in relating the variables t_1, t_2 , b and T_0 are presented in equation 6. By specifying the variable ratios, the R-K polynomial was reduced to Equation 7. The derived model from this modification is shown in Fig 7.0. The resulting source signal was used to synthesise the vowel /a/ and produced equally intelligible speech as the R-K polynomial.

$$b = T_0 \quad t_1 = 0.111b = aT_0 \quad t_2 = 0.667T_0 = cT_0 \tag{6}$$

$$g(t) = \begin{cases} 0 & 0 \le t \le aT_0 \\ A(\frac{(t-aT_0)}{(cT_0-aT_0)})^2(3 - 2\frac{(t-aT_0)}{(cT_0-aT_0)}) & aT_0 \le t \le cT_0 \\ A(1 - \frac{(t-cT_0)}{(T_0-cT_0)}) & cT_0 \le t \le T_0 \end{cases} \tag{7}$$

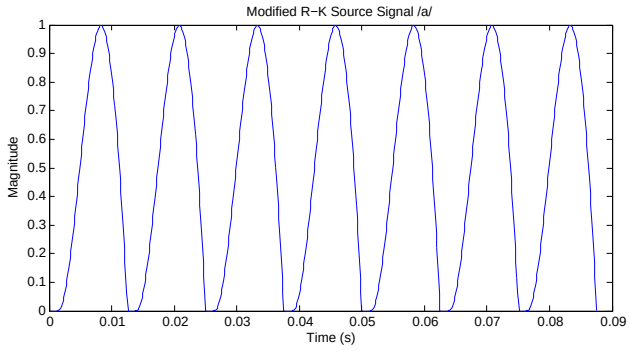

Fig 7.0 Modified R-K source signal model

2.3. HNM synthesis

A fairly new technique discussed in this paper is modelling the source signal using HNM [6]. HNM is a speech synthesis and modelling technique on its own [2]. Research has generally shied away from this technique because of the complication in finding the HNM model parameters [7].

The harmonic plus noise model (HNM) is based on the fact that speech can be viewed as two components, namely the harmonic part $h(t)$ a quasi periodic signal and the non periodic component noise $n(t)$. These two components are distinctly separated by a time varying quantity $Fmax$ (maximum voiced frequency). The lower component is solely composed of harmonics and the upper band noise as shown in Fig 8.0 and Equation 8 [2] on the frequency spectrum of the residual signal for vowel /a/.

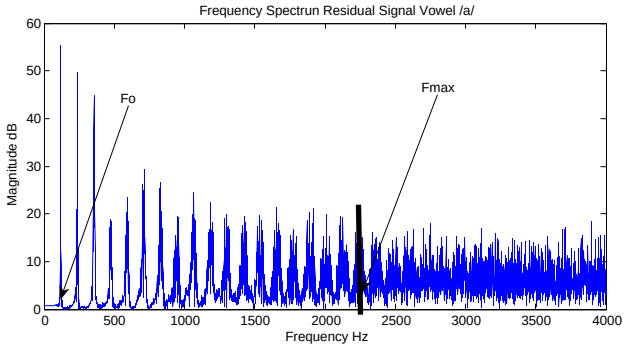

Fig 8.0 Characteristics HNM signal model

$$h(t) = \sum_{k=1}^{K} A_k(t)cos(k\theta(t) + \theta_k(t) \tag{8}$$

$n(t)$ noise component is derived from filtered white Gaussian noise.

The number of harmonics K is given by $Fmax/Fo$ were Fo is the pitch frequency [7].

2.3.1. HNM source signal modelling

This section describes how the HNM technique was used to model the source signal, as well as deriving the parameters for HNM equation 8. From Fig 8.0 it is clear the residual signal exhibits characteristics equivalent to those of the actual speech signal. Therefore, the source signal can be described as a sum of the harmonic and noise of the residual. The major complication as stated earlier is the derivation of HNM parameters Fo, $Fmax$, θ and A_k. The techniques we applied in solving the HNM parameters are explained below.

2.3.2. F_0 and $Fmax$ estimation

F_0 is defined as the pitch frequency and is given as the frequency of the first harmonic [5]. Once the F_0 was obtained then $Fmax$ and the number of harmonics K were calculated based on the relationship in equation 9 [7].

$$maxA_i - A_n \ge 13db \tag{9}$$

A_n is the average magnitude of the noise spectrum
Were A_k is the peak amplitude in the range specified in equation 10

$$[F_k - \frac{F_0}{2}, F_k + \frac{F_0}{2}] \tag{10}$$

F_k is a multiple of F_0 the fundamental Frequency such that $F_k = kF_0$

*The first instant that equation 10 is not satisfied defines the number of harmonics in the signal spectrum as K and the maximum voiced frequency as Fmax.

2.3.3. Phase modelling

One complexity of HNM is computing the phase from the frequency domain waveform [8]. A method of linearity was used to model the phase relationships between HNM harmonics [2]. Tests were performed by observing the quality produced for vowel sounds /a/, /e/, /i/, /o/, /u/ when the phase of all the harmonics was varied linearly over a 360, 180, 270, 90 degree intervals equation 11. Positive results were obtained for all vowels when the phase was varied on the 360 degree interval.

$$\theta_k = (\frac{2\pi}{K})(k-1) \tag{11}$$

2.3.4. Modelling the harmonic and noise interaction

The source signal models discussed thus far fail to model effectively the noise interaction between the harmonics *(voiced source)* and the noise *(unvoiced source)* [2]. This is because the R-K, triangular and impulse signal models assume the source signal to be purely harmonic or noise [1]. As a solution to this problem the HNM synthesis model developed, allows the modelling of the noise interaction by multiplying the developed source signal with a noise window at the interaction of the two components. The noise window is equivalent to passing a white Gaussian noise through a band pass filter bounded by $0.75Fmax \quad and \quad 0.85Fmax$. The resulting residual is shown in Fig 9; clearly this is a better approximate of the residual signal. The vowel /a/ was synthesised using this source signal model and satisfactory results were obtained when comparing the intelligibility with the other source signal models.

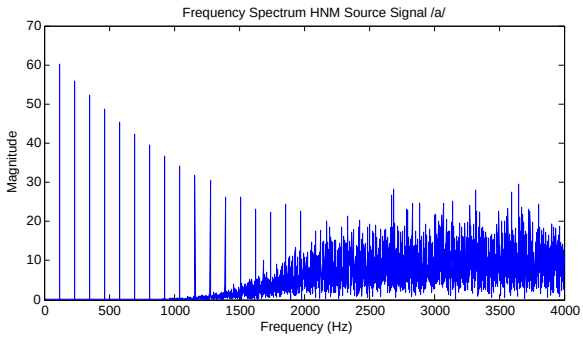

Fig 9.0 HNM Source signal model

3. Discussion

Listening tests were conducted on vowels sounds /a/, /e/, /i/, /o/, /u/ produced from each of the previously described models. A sample of 20 students (10 male) and (10 female) from the School of Electrical and Information Engineering at the University of the Witwatersrand South Africa conducted tests on wave files generated from the discussed models in MATLAB. For each vowel the original speech segment used in analysis was played as well as the synthetic sound from each model, each student was then asked to give her opinion and perception using a scale of (1-5) on the quality, intelligibility and audibility of the sound produced from each of the five source signal models when applied to LP synthesis. On the (1-5) scale (1) represented poor inaudible quality speech and (5) the best audible quality. The listener was asked to give his or her score on the (1-5) range after listening to each synthetic vowel on all discussed models at least three times. The average score from all the listeners for a each specific model and sound was tabulated and recorded. The results of the listening tests for the source signal **(SS)** models are shown in Table 1.0.

SS Model	**/a/**	**/e/**	**/i/**	**/o/**	**/u/**
Impulse Train	2.5	3	3	3	2.5
Triangular Pulse	3	3.2	3.2	3	3
R-K Signal	3.5	3.8	3.8	3.5	3.5
Modified R-K	3.4	3.8	3.8	3.4	3.5
HNM Source Signal	3.6	4.0	4.0	3.3	3.3

Table 1.0 Performance of the source signal models (SS models) for vowel LP synthesis

From the results it is evident that the HNM model produced better synthetic speech. It is also evident that the modified R-K and the original R-K source signal models were comparable.

4. Conclusions

The paper has described two fairly new approaches to source signal modelling for LPC synthesis based on HNM and a linearised model of the R-K model. Other well documented source signal modelling methods for LPC synthesis were briefly described in this paper. The two modified models produced better quality synthetic speech when compared to previously renowned simplified models such as Impulse train for the vowels /a/, /e/, /i/, /o/, /u/. Further testing still has to be done for fricative and nasal sounds using these described models.

5. Acknowledgments

The authors would like to thank the Electronic Engineering research group at the University of Witwatersrand Johannesburg and the department of trade and industry in South Africa for providing funding through the THRIP. Finally yet importantly, the authors would also like to thank Gedion Klompje previously of the language-processing group at the University of Stellenbosch in South Africa for sharing ideas in the field of speech synthesis.

6. References

[1] I.H. Witten, *Principles of Computer Speech*, Academic Press, 1982.

[2] Y Stylianou "On the implementation of the harmonic plus noise model for concatenative speech synthesis," *In Proceedings. of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP Volume 2, Issue 2000*. pp II957 - II960 , Istanbul Turkey, 9 June 2000.

[3] J. Makhoul. "Linear prediction: A tutorial review," *In Proceedings of the IEEE, vol 63*. pp 561-580, April 1975.

[4] F.J. Owens, *Signal Processing of Speech*, The Macmillan Press Ltd, 1993.

[5] S Roa, M Bennewitz, S Behnke "Fundamental frequency estimation based on pitch-scaled harmonic filtering," *In Proceedings. of the IEEE International*

Conference on Acoustics, Speech, and Signal Processing, ICASSP Volume: 4. pp IV-397-IV-400, Honolulu Hawaii, 15-20 April 2007.

[6] G Klompje, T.R Niesler, "A parametric monophone speech synthesis system", *In proceedings of the seventeenth annual symposium of the Pattern Recognition Association of South Africa (PRASA)*, Parys South Africa, November 2006.

[7] Y Stylianou "Applying the harmonic plus noise model in concatenative speech synthesis," *IEEE Transactions on speech and audio processing, Volume 9, Issue 1*. pp 21 - 29 , January 2001.

[8] Y Stylianou "Concatenative speech synthesis using the harmonic plus noise model," *Third ESCA Speech Synthesis Workshop*. pp 261 - 266 , November 1998.