

Analysis of Voiced Speech using Harmonic Models

Yannis Stylianou

AT&T Labs Research, 180 Park Ave., Florham Park, NJ 07932, USA

Summary: It is very common to consider voiced speech as a periodic or quasi-periodic signal. Therefore, harmonic models are usually proposed for the modeling of this kind of speech. This paper focuses on the harmonic analysis of speech. Three models are presented and compared. The first model is the simplest among the three models supposing that the speech signal is a stationary signal. The second model does not consider speech as a stationary signal and even though it uses the fundamental frequency as the basic frequency of the signal, it is not a harmonic model. The third model does not make the hypothesis of the stationarity of the speech signal, however, it is a harmonic model.

INTRODUCTION

It is very common to separate speech into a non-periodic part and into a periodic or quasi-periodic part. This kind of decomposition of the speech signal is a subject of considerable importance with applications in speech synthesis, speech coding, psychoacoustic research and pathological voice detection. Because the signal is decomposed into two parts, different modification methods can be applied to each part, yielding, for example, more natural sounding pitch and time-scale modifications. The naturalness of the prosodically modified speech signal is very important for high-quality text-to-speech synthesis based on acoustical unit concatenation. For speech coding different coding schemes can be applied to each part [1].

One of well-known differentiators in the analysis of speech is a simple voiced/unvoiced classification of speech segments over time. However, observations of short-time spectra of speech indicate that there are regions in spectrum dominated by harmonics of a fundamental frequency and other regions dominated by noise-like energy. In this paper, we will label these regions that are dominated by harmonics in the time-frequency space of speech as “voiced speech” while we will refer to all the other regions as “unvoiced”. This paper focuses on the harmonic analysis of voiced speech. Three models are presented and compared. The development and description of the three models is given first. The second part of the paper is devoted to estimating parameters for the three models, followed by a comparison of the three models. The comparison will be done on the variance of the residual signal obtained from the three models, the spectral content of the residual signal and the error modeling. Finally, the paper concludes with a discussion of the usefulness of the proposed models for speech coding and for prosodic modifications of speech.

PRESENTATION OF THE MODELS

The basic approaches to speech analysis proposed in this paper are to approximate a discrete-time sequence $s[n]$ using one of three harmonic models. The models assume the speech signal

to be composed of a periodic component $h[n]$ and a non-periodic component $r[n]$. The periodic component is designated as sums of harmonically related sinusoids

$$h[n] = \sum_{k=-L(n_i)}^{L(n_i)} A_k[n] e^{j2\pi k f_0(n_i)(n-n_i)} \quad (1)$$

where $L(n_i)$ denotes the number of harmonics included in the harmonic part at $n = n_i$, $f_0(n_i)$ denotes the fundamental frequency at $n = n_i$, while $A_k[n]$ can take on one of the following forms:

$$A_k[n] = a_k(n_i) \quad (2)$$

$$A_k[n] = a_k(n_i) + (n - n_i) b_k(n_i) \quad (3)$$

$$A_k[n] = a_k(n_i) + (n - n_i) c_k(n_i) + (n - n_i)^2 d_k(n_i) \quad (4)$$

where $a_k(n_i)$, $b_k(n_i)$, $c_k(n_i)$, and $d_k(n_i)$ are assumed to be complex numbers with $\arg\{a_k(n_i)\} = \arg\{c_k(n_i)\} = \arg\{d_k(n_i)\}$ (assumption of constant phase). These complex numbers denote the amplitude of the k th harmonic, its first derivative (slope) and its second derivative. These parameters are measured at time $n = n_i$ referred to as analysis time instants. The number of harmonics, $L(n_i)$, depends on the fundamental frequency $f_0(n_i)$ as well as the frequency bands labeled as “voiced”. For $|n - n_i|$ small, the main model assumes that $f_0(n) = f_0(n_i)$ and $L(n) = L(n_i)$, which means that the fundamental frequency and the number of pitch-harmonics are held constant within each analysis frame and equal to the values at the center of the analysis window, n_i . Note that the second model is not a harmonic model, though it uses a fundamental frequency as a main frequency, because the derivative of the phase function is not a multiple of this fundamental frequency.

ESTIMATION OF MODEL PARAMETERS

Considering that the fundamental frequency as well as the number of harmonics is already estimated [2], the next step consists of estimating the harmonic amplitudes, slopes and phases. This is done using a weighted least-squares method aimed at minimizing the following criterion with respect to $A_k[n]$:

$$\epsilon = \sum_{n=n_a^i-N}^{n_a^i+N} w^2[n] (s[n] - \hat{h}[n])^2 \quad (5)$$

where $w[n]$ is a weighting window and N is the integer closest to the local pitch period $T(n_i)$. The above criterion has a quadratic form for the parameters of the first and the second model, and is solved by inverting an over-determined system of linear equations. For the third model, a non-linear system of equations has to be solved. Once the periodic part of the speech is estimated using the above three harmonic models, the non-periodic part (or the so called *residual signal*) can be obtained by subtracting the harmonic part from the original signal. Although the paper focuses on the analysis of the periodic part, the non-periodic part will be helpful for our analysis as it is a function of the periodic models.

PROPERTIES OF THE RESIDUAL SIGNALS

Variance of the residual signals

It can be shown [2] that if the input signal, $s[n]$, is a white noise with unit variance then the covariance matrix is given by:

$$E(\mathbf{r}\mathbf{r}^h) = \mathbf{I} - \mathbf{W}\mathbf{P}(\mathbf{P}^h\mathbf{W}^h\mathbf{W}\mathbf{P})^{-1}\mathbf{P}^h\mathbf{W}^h \quad (6)$$

where \mathbf{I} is the identity matrix, \mathbf{W} is a diagonal matrix with values of the analysis window as entries and matrix \mathbf{P} has as elements the exponential functions of Eq.1. The variance of the residual signal, $r[n]$, is the diagonal of the above matrix. Using the same weighting window $w(t)$, typically a Hamming window, the variance of the residual signal from each one of the harmonic models has been computed. The three variances are depicted in the Fig.1. It is clear by Fig.1 that the

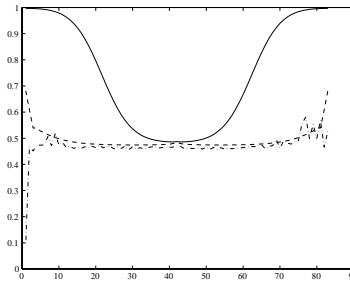


Figure 1: Variance of the least-squares residual signals from: the first model (solid line), the second model (dashed line) and the third model (dash dotted line). The weighting window was a typical hamming window.

variance of the residual signal from the first model is far away from the ideal least-squares variance while both the second and the third model approximate this pretty well.

Time domain comparison of the residual signals

The amplitudes and the phases estimated by the first model correspond to the values at the center of each analysis frame and there is, therefore, no information about the parameters variation within the analysis frame. This causes low frequencies to appear in the residual signal of the first model increasing the modeling error. On the other hand, the other two models provide information about the evolution of amplitude and phases within an analysis frame. We define the modeling error as:

$$E = 10 \log_{10} \frac{\sigma_{r(t)}^2}{\sigma_{s(t)}^2} \quad (7)$$

where $\sigma_{r(t)}^2$ denotes the variance of the residual signal $r(t)$, and $\sigma_{s(t)}^2$ denotes the variance of the original speech signal $s(t)$. Fig.2(a) shows a segment of a speech signal ('wazi waza') and in (b) the modeling error in dB produced from the three models is plotted (first model: solid line, second model: dashed line, and third model: dash dotted line). It is clear from this figure that the

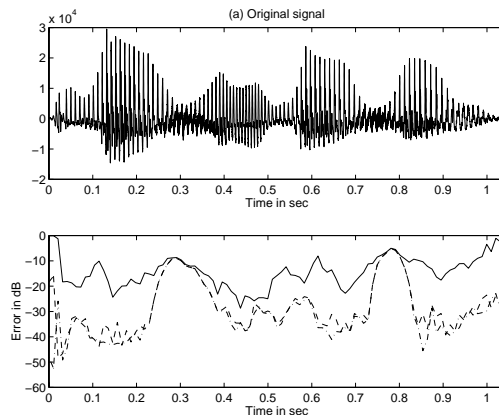


Figure 2: Modeling error in dB using the three harmonic models.

modeling error using the first model is greater than those produced by the two other models. As the full frequency band (from $0Hz$ up to half of the sampling frequency) of the residual signal is used in the definition of the modeling error, the error is large on voiced fricative regions where not many frequencies belong to the periodic part (for example, between the 0.25 and 0.35 sec.). While one might expect that the two more sophisticated models perform in a similar way, the following experiment has showed that this is not true. Adding wideband noise to the original signal, the second model tries to model the noise (as being part of the periodic part) while the third model leaves the additive noise “untouched”. This is mainly due to the fact that the phase of the second model is a non-linear function of frequency in contrast to the phase function of the third (and first) model.

DISCUSSION AND CONCLUSION

This paper discussed the modeling of voiced speech based on three harmonic models. The first model is the simplest one but has the largest modeling error among the proposed models. This was expected, because the first model does not include any information about the variation of the harmonic amplitudes. The second model has a low modeling error and fast transitions are well modeled. However, the phase function of this model is not linear having as result to model additive noise as part of the voiced (periodic) signal. The third model has also a small modeling error and as the second model, fast transitions are efficiently modeled. However, because the phase of the model is linear, additive noise is not “folded” into the harmonic signal. This makes the third model suitable for robust modeling of the speech signal as well as for prosodic modifications of speech.

REFERENCES

- [1] D. Griffin and J. Lim, “Multiband-excitation vocoder,” *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-36, pp. 236–243, Feb 1988.
- [2] Y. Stylianou, J. Laroche, and E. Moulines, “High-Quality Speech Modification based on a Harmonic + Noise Model.,” *Proc. EURO-SPEECH*, pp. 451–454, 1995.