Harmonic Weighting for All-Pole Modeling of the Voiced Speech

Davor Petrinovic
davor.petrinovic@fer.hr

Faculty of Electrical Engineering and Computing
University of Zagreb, Croatia

Abstract

A new distance measure for all-pole modeling of voiced speech is introduced in this paper. It can easily be integrated within the concept of discrete Weighted Mean Square Error (WMSE) all-pole modeling, by a suitable choice of the modeling weights. The proposed weighting will address the problems such as: harmonic estimation reliability, perceptual significance of the harmonic and the model mismatch errors. The robust estimator is proposed, to reduce the effect of outliers caused by spectral nulls or additive non-speech contributions (e.g. background noise or music). It is demonstrated that the proposed all-pole estimation can significantly improve the performance of speech coders based on sinusoidal model, since the harmonic magnitudes are modeled better by the WMSE all-pole model.

1. Introduction

Voiced speech can be modeled very well by a set of sine-wave oscillators with harmonically related frequencies $\omega_i = i\cdot\Omega$, $i=0,1,....,N$, where Ω is the pitch frequency and N is the index of the highest harmonic with $\omega_N < \pi$. This is known as a sinusoidal model [1]. Basic idea of the sinusoidal model representation is to estimate sine-wave complex amplitudes (magnitudes and phases) directly from the speech signal segment. Thus, the signal segment is represented by a set of model parameters (fundamental frequency and complex amplitudes of harmonics) instead of its time samples. Sine-wave complex amplitudes can be encoded directly, or by a two-step procedure depicted in the upper part of Fig. 1. The input speech is processed by a conventional LPC analysis followed by sinusoidal modeling performed on the gain normalized residual signal of the LPC prediction. The LPC model roughly describes sine-wave amplitudes and consequently the residual signal exhibits much lower dynamics of the harmonic magnitudes then the input speech signal (cca. +/- 5dB). The log-magnitude of the i^{th} sine-wave oscillator y_i can be represented as a sum of the estimated residual harmonic log-magnitude, e_i, and the log-magnitude sample $P(\omega_i)$ of the LPC model frequency response $H(e^{j\omega})$ sampled at the harmonic frequency $\omega_i = i\cdot\Omega$, $i=0,1,....,N$:

$$y_i = e_i + P(\omega_i) \tag{1}$$

$$P(\omega_i) = 20\cdot\log_{10}\left(\left|H(e^{j\omega_i})\right|\right) \tag{2}$$

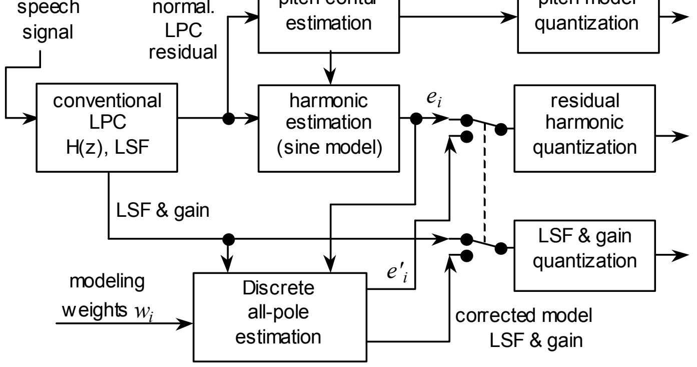

Figure 1. Application of the proposed WMSE all-pole model

The first part of the model can be quantized very efficiently by any of the known and well-studied spectrum quantization techniques, typically based on the Line Spectrum Frequencies, LSF. The second part is usually quantized by some of the techniques for residual harmonic magnitude/phase quantization. This is not a trivial problem, since the number of model parameters is variable (pitch dependent) and thus requires appropriate variable dimension quantization techniques. Although the LPC model captures a great deal of the speech signal information, it has been shown that a significant percentage of the overall codec bit-rate (typ. >40%) must be dedicated to the representation of the quantized residual harmonic magnitudes e_i.

Ideally, the second step wouldn't be necessary at all if the original LPC model could be replaced by some other all-pole model, of the same order p, that passes exactly through the spectrum peaks corresponding to the harmonic magnitudes y_i. This idea is illustrated at the bottom of Fig 1. where a new all-pole model is determined from the transfer function of the initial LPC model $H(z)$ and from the estimated residual harmonic magnitudes e_i. Several approaches had been proposed in order to find this better fitting all-pole model [2],[3],[4]. If the underlying physical process is indeed an all-pole process, then the exact match is possible resulting with an ideal excitation signal with equal amplitude of all excitation harmonics (e'_i=0dB). On the other hand, for voiced sounds with spectral nulls, for sounds with mixed mode excitation or for corrupted speech with significant additive tonal or noise contributions, the estimated spectrum peaks do not follow the all-pole envelope.

Nevertheless, even for such signals, the all-pole model can be forced to approximate the harmonic peaks in a sense of minimizing certain distance measure. Ideally, the distance measure should be chosen in a way that the resulting all-pole model captures magnitudes of the largest possible number of harmonics. The modeling error of the remaining few that cannot be fitted by the all-pole envelope can be captured by the residual signal model e'_i.

A new distance measure for estimation of the discrete all-pole model for voiced speech will be proposed in this paper, that will try to address the main aspects related to the real speech. It will be shown that the variances of the residual harmonic magnitudes of the improved all-pole modeling are much lower compared to the conventional LPC model. This represents the grounds for reduction of the required bit-rate for residual harmonic magnitude quantization and improved performance of the speech coder based on such model.

2. Discrete WMSE all-pole estimation

A method for estimation of the all-pole model with the best possible fit to a discrete set of spectrum peaks was proposed in [4] and was validated on a synthetic vowel example. It is based on minimizing the weighted mean square error (WMSE) between the target harmonic log-magnitudes $\mathbf{Y}=[y_0,y_1,...,y_N]^{\mathrm{T}}$ given in decibels and the log-magnitude samples $P(\omega_i)$ of the all-pole model frequency response. Although the algorithm allows arbitrary frequencies of the peaks, in the context of voiced speech modeling the spectrum peaks are equidistant and equal to the harmonic magnitudes y_i estimated from the voiced speech segment. The Weighted Root Mean Square (WRMS) modeling error D of the current speech frame is determined as follows:

$$D^2 = D^2(H,\mathbf{Y}) = \left(\sum_{i=0}^{N} w_i \cdot e_i^2 \right) \Bigg/ \sum_{i=0}^{N} w_i \qquad (3)$$

where e_i is the modeling error of the i^{th} harmonic and w_i is its modeling weight.

This paper will follow the theoretic background proposed in [4] and it will demonstrate how the proposed estimation can be applied to the real voiced speech. Discrete all-pole estimation is based on an iterative algorithm that performs refinement of the all-pole model by directly modifying its LSF vector and gain term. Optimal modification is computed in each iteration of the algorithm from the following three elements:

- spectral sensitivity of the LSF vector components of the current all-pole model evaluated at ω_i,
- modeling error of the current all-pole model $e_i=y_i-P(\omega_i)$,
- modeling weights w_i for each of the harmonics.

If the target harmonic magnitudes y_i are indeed samples of some p^{th} order all-pole envelope, then the choice of the harmonic fitting weights w_i is not critical and the algorithm converges to the ideal solution. This was demonstrated in [4] on a synthetic vowel example with unit weights. However, for the real speech, solution is always a compromise and the proper choice of weights is of crucial importance for good performance of the estimator as will be explained next.

3. Modeling weights for the real speech

As with any optimization problem, the final solution is as good as is the distance measure minimized in the optimization procedure. The following factors must be taken into account in the definition of the modeling weight:

- reliability of the harmonic magnitude estimate
- perceptual significance of the harmonic
- estimation bias due to the all-pole outliers

A modeling weight is given to each of the above listed factors and the final weight is computed as a product of the three:

$$w_i = w_{rel}(i) \cdot w_{per}^2(i) \cdot w_{out}(i) \qquad (4)$$

The details related to each of the weighting factors will be explained in the following sections.

3.1. Harmonic reliability weight

Due to high-pass or band-bass pre-filtering of the speech signal prior to analysis, certain spectrum regions are significantly attenuated and the estimated harmonic components within those regions should not be used for all-pole estimation. Based on the pitch frequency Ω, two sets of harmonic indices are formed: in-band set I_{in} and out-band set I_{out}, defined as:

$$I_{in} = \{ i,\ \text{for } \omega_{min} \le \Omega \cdot i \le \omega_{max} \} \qquad (5)$$

$$I_{out} = \left\{ i,\ \begin{matrix} \text{for } 0 \le \Omega \cdot i < \omega_{min} \\ \text{or } \omega_{max} < \Omega \cdot i \le \pi \end{matrix} \right\} \qquad (6)$$

where ω_{min} and ω_{max} are the lower and the upper cutoff frequency of the pre-filtered speech. Due to non-reliable estimation of the out-band harmonics the target log magnitudes y_i for $i \in I_{out}$ are replaced with the initial LPC model samples $P(\omega_i)$, thus setting e_i for $i \in I_{out}$ to zero for the initial iteration. However, in the iterative procedure, the all-pole model is modified relative to the initial LPC model, and out-band errors become nonzero again. Since the out-band modeling error is not as important as is the inbound error, a constant weighting factor $w_{ob}<1$ is introduced for out-band harmonics:

$$w_{rel}(i) = \begin{cases} 1, & i \in I_{in} \\ w_{ob}, & i \in I_{out} \end{cases} \qquad (7)$$

Assigning a zero weight (w_{ob}=0) to out-band harmonics is not recommended, since it may cause unpredictable behavior of the all-pole model in those regions. The value used in the experiments was w_{ob}=0.25. Weight $w_{rel}(i)$ can also be combined with multi-band voicing factor proposed in [5], thus giving higher weight to genuine voiced harmonics, and lower weight to the noise like bands of the spectrum.

3.2. Perceptual weight

A perceptually based model for improved LSF quantization was proposed in [6]. It mimics the quantization error integration along the nonlinear Bark frequency scale, by assigning certain weights to each of the LSFs. The weights are determined by evaluating the derivative $w_B(f)$ of the Bark scale mapping at frequencies of the LSF components. Weighting $w_B(f)$ is given in (8) and is shown in Fig. 2.

$$w_B(f) = 1 \Big/ \left(25 + 75 \cdot \left(1 + 1.4 (f/1000)^2 \right)^{0.69} \right) \qquad (8)$$

A similar approach is followed in this paper, by assigning a perceptual weight $w_{per}(i)$ to each harmonic:

$$w_{per}(i) = \frac{1 - 0.5 e^{-f_i/250}}{25 + 75\sqrt{1 + 1.4 (f_i/1000)^2}} \qquad (9)$$

where f_i is the frequency of the i^{th} harmonic in [Hz], i.e. $f_i=\omega_i/(2\pi)\cdot f_s$ and f_s is the sampling frequency. The proposed

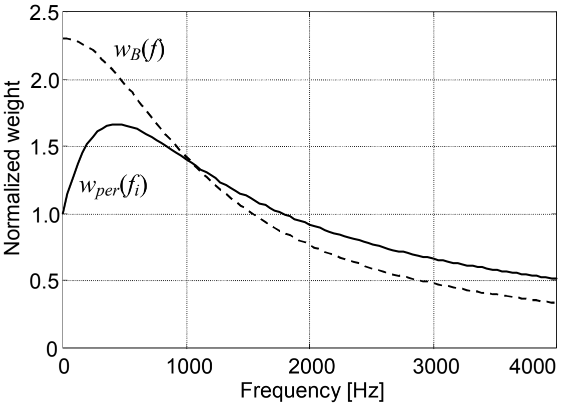

Figure 2. Perceptual weighting

weighting $w_{per}(f_i)$ shown with solid line in Fig. 2. is closer to the experimentally determined perceptual weighting in [6]. It is obvious that compared to $w_B(f)$, $w_{per}(f_i)$ gives lower weights to very low frequency region and slightly higher weights to high frequency band.

3.3. All-pole outlier weighting

Modeling outliers are usually caused by the spectral nulls or additive tonal (narrowband) errors. Simple squared Euclidean distance gives very high emphasis to any such outlier harmonic that is either far above or far bellow the all-pole envelope, thus affecting the fit to all neighboring harmonics. As a result, the envelope is biased toward these outliers. Therefore, the new distance function $d(e)$ is proposed that is defined as:

$$d(e)=\begin{cases}-2c_{ns}e-c_{ns}^2 & e<-c_{ns}\\ e^2 & -c_{ns}\le e\le c_{ps}\\ 2c_{ps}e-c_{ps}^2 & e>c_{ps}\end{cases}\qquad(10)$$

Distance $d(e)$ is equal to the squared Euclidean distance only for e within the two thresholds $-c_{ns}$ and c_{ps}, while for the larger errors, a linear model is used. The latter is proportional to the absolute value of the error with slopes $2c_{ns}$ and $2c_{ps}$ for negative and positive errors respectively. Similar type of distance measure is used in Robustness Theory for maximum-likelihood estimation of sources with heavy-tailed non-Gaussian distributions [7]. An example of $d(e)$ with thresholds c_{ns}=0.5 and c_{ps}=1 is shown in Fig. 3.

Since the values of outliers are speech dependent, it is very hard to impose some fixed thresholds. Therefore the thresholds c_{ns} and c_{ps} are computed by scaling a pair of fixed thresholds c_n and c_p with scale estimates of the negative and the positive errors. If the all-pole modeling error is divided in two groups $I_p=\{i,\text{ for } e_i\geq 0\}$ and $I_n=\{i,\text{ for } e_i<0\}$, then c_{ns} and c_{ps} are computed as:

$$c_{ns}=c_n\sqrt{\frac{\sum\limits_{i\in I_n}e_i^2}{\mathrm{C}(I_n)-1}}\,,\quad c_{ps}=c_p\sqrt{\frac{\sum\limits_{i\in I_p}e_i^2}{\mathrm{C}(I_p)-1}}\qquad(11)$$

where C( ) denotes the cardinal number of the sets.

The proposed distance measure can be easily implemented with conventional WMSE distance, by simply applying the following outlier weights $w_{out}(i)$ to each of the harmonics:

$$w_{out}(i)=d(e_i)\big/e_i^2\qquad(12)$$

where $d(\ )$ is the distance given in (10) and e_i is the modeling error of the i^{th} harmonic. The proposed distance measure is very flexible, since the emphasis can be given to either positive or negative modeling errors. Generally, the all-pole model should represent a spectral envelope passing as close as possible to the spectrum peaks. Therefore, a convenient choice for the thresholds (also used in the experiment) is c_n=0.3 and c_p=1.5, thus significantly de-emphasizing the effect of spectral nulls. The positive threshold is still low enough to ensure good estimation in the case of significant tonal outliers.

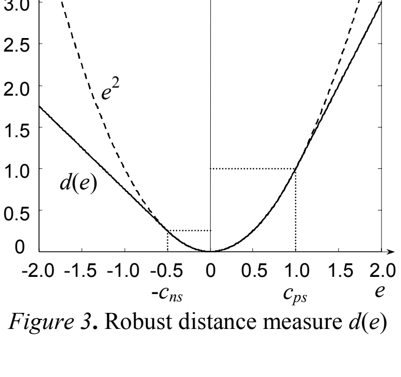

Figure 3. Robust distance measure $d(e)$

4. Convergence and stability of the solution

The LSF log-spectral sensitivity formulation used in the all-pole modification algorithm is based on the second order Taylor expansion of the log-spectral distance between the initial and the modified all-pole model [4]. If the required modification of the LSF vector is small enough, then the higher order terms can indeed be neglected. However, if the initial model is far from the target harmonic magnitudes, then the required modification of the LSF vector is also large, thus violating the initial small variations assumption. The solution to this problem, also applied in this experiment, was to scale the computed optimal modification of the LSF vector by a step size α<1, similar to the approach proposed in [3].

Additional restriction that is imposed on the solution is the stability of the all-pole model. Furthermore, minimum formant bandwidth should also be limited to avoid sharp peeks. Since the all-pole estimation is performed by modifying the LSFs, stability (ordering property) and minimum separation of the neighboring LSF components can be easily checked and enforced. Similar problem exists in certain LSF quantization schemes and can be solved by stabilization techniques that spread the LSF pairs that come too close to each other [8]. However, it must be noted that this kind of brute force modification within the optimization loop may sometimes result in divergence of the procedure.

5. Example of the optimal all-pole model

To illustrate the exceptional performance of the proposed method in resolving the all-pole model from the harmonic samples, one example is shown in Figs. 4 to 6. Segment of the voiced female speech from the beginning of the phoneme "*ð*" in pronunciation of the word "*the*" with pitch frequency of 230 Hz was sampled with f_s=8kHz and analyzed. Initial LPC estimation was performed as described in the G729 standard [8], while the harmonic magnitude estimation was performed in the domain of Generalized Fourier Transform, GFT, using the highly accurate algorithm described in [9]. Due to polynomial phases of the complex exponentials used as transformation basis functions, the variations of the pitch frequency within the analysis window are canceled by the same type of variations of the basis function frequencies.

In Fig. 4 the GFT spectrum of the signal is shown with a dotted line, while the GFT spectrum of the synthetic all-harmonic signal model constructed from the estimated harmonic magnitudes y_i (depicted with circles 'o') is shown with a solid line. Samples of the LPC model $P(\omega_i)$ on harmonic positions are shown with a '+' symbol. As can be observed, the initial autocorrelation LPC model (shown with the thin dash-dot line) does not fit harmonic samples y_i very well.

Modeling error $e_i= y_i-P(\omega_i)$ of the LPC model is shown in Fig. 5 with a star symbol, together with the GFT of the LPC residual signal (dotted line) and GFT of the synthetic all-harmonic excitation signal to the LPC model (solid line). It is obvious that the error e_i of the initial LPC model is between –7 and +5 dB at harmonic positions. Scaled outlier thresholds c_{ns} and c_{ps} defining the outlier weights $w_{out}(i)$ are also shown.

Based on the weights w_i proposed in this paper, the WMSE all-pole model was determined as in [4] with only two iterations of the algorithm. The resulting WMSE all-pole model is shown with a thick dashed line in Fig. 4 together with corrected harmonic samples $P(\omega_i)$ depicted with '*'. The improved harmonic modeling can be observed even better in Fig. 6 showing the modeling error of the WMSE model. The GFTs of the residual signal and synthetic excitation corresponding to the modified all-pole model are also shown. All magnitude errors are within +/-1dB except for the 4th and 5th harmonic that have errors e_4=1.42 and e_5=−4.3 dB respectively. WRMS modeling error was reduced from D=3.85dB for the initial LPC model to D=1.03 dB for the proposed WMSE all-pole model. For the WMSE model, the 5th harmonic is treated as an outlier and its fit is sacrificed to improve the modeling of the neighboring harmonics.

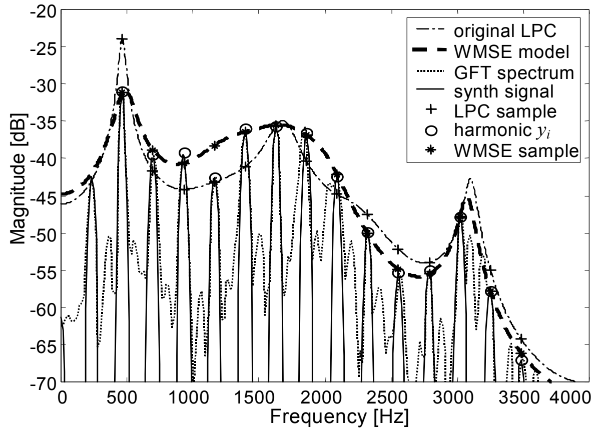

Figure 4. Conventional LPC and proposed WMSE modeling

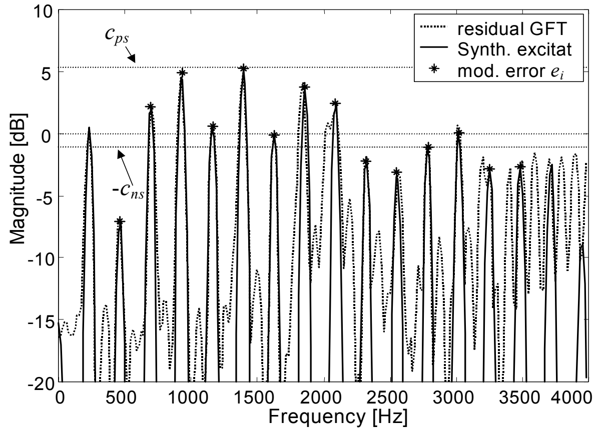

Figure 5. Modeling error of the conventional LPC

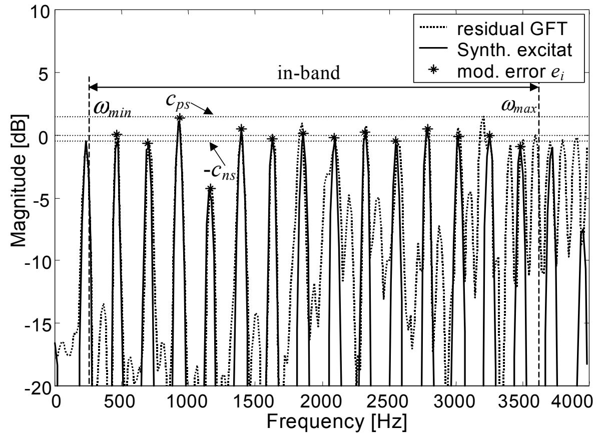

Figure 6. Modeling error of the WMSE all-pole model

6. Estimation performance

To get a better insight in the performance of the proposed harmonic all-pole estimation, 15 sentences spoken by 8 male and 7 female speakers were processed as already described. WRMS modeling error D was computed for all voiced frames and averaged separately for male and female speakers. Error D was computed for the initial LPC model and for 1 to 10 iterations of the proposed all-pole estimation. Three step sizes were used: α=1, α=1/2 and α=1/3 as shown in the Table 1. The best results are obtained with α=1/2, thus reducing the modeling error D in only two iterations by 31% for female and 24% for male speakers.

Table 1. WRMS modeling error vs. step size α and number of iterations of the estimation algorithm

# of iterat	Female speakers			Male speakers		
	α=1	α=1/2	α=1/3	α=1	α=1/2	α=1/3
LPC	2.62			3.01		
1	2.12	2.03	2.12	2.55	2.50	2.56
2	1.92	1.81	1.87	2.35	2.30	2.35
3	1.87	1.75	1.79	2.30	2.26	2.29
10	1.84	1.71	1.70	2.28	2.23	2.23

7. Conclusion

Harmonic weighting for all-pole modeling of the voiced speech has been proposed in the paper. It has been demonstrated that by simple modification of the conventional LPC model, the harmonic magnitude modeling error can be reduced by up to 30%. Due to the reduced residual magnitude variance, a significant bit-rate savings can be achieved in actual speech coder implementations. Proposed robust distance measure with adaptive thresholds ensures good estimation performance even for corrupted speech.

8. References

[1] McAulay, R.J. and Quatieri, T.F., “Speech analysis/synthesis based on sinusoidal representation”, *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 34, no. 4, pp. 744-754, August 1986

[2] McAulay, R.J., Quatieri, T.F., "Sinusoidal coding" in *Speech Coding and Synthesis*, ed. Kleijn, W.B., Paliwal, K.K., Elsevier, 1995, pp. 121-173

[3] El-Jaroudi, A., Makhoul, J., "Discrete all-pole modeling", *IEEE Transaction on Signal Processing*, vol. 39, no. 2, pp. 411-423, Feb. 1991

[4] Petrinovic, D., "Discrete weighted mean square all-pole modeling", *Proceedings of IEEE Int. Conf. Acoust., Speech, Signal Processing,* 2003, SPEECH-P10.12

[5] Griffin, D.W. and Lim, J.S., “Multiband excitation vocoder”, *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, no. 8, pp. 1223-1235, August 1988.

[6] Cohn, R.P.; Collura, J.S., "Incorporating perception into LSF quantization some experiments", *Proceedings of IEEE Int. Conf. Acoust., Speech, Signal Processing,* 1997, Vol. 2 , pp 1347 -1350

[7] Huber, P.J. *Robust Statistics*, New York, Wiley, 1981

[8] ITU-T G.729, *Coding of speech at 8kbit/s using conjugate-structure algebraic-code-excited linear-prediction (CS-ACLEP)*, ITU, 1996.

[9] Petrinović, D. Cuperman V., "Analysis of the voiced speech using the generalized Fourier transform with quadratic phase", *Proceedings of the 7th European Conference on Speech Communication and Technology, EUROSPEECH*, 2001. vol. 4, pp. 2479-2482