SPEECH ENHANCEMENT USING VOICE SOURCE MODELS

*Anisa Yasmin**, *Paul Fieguth*** *and Li Deng**

*Department of Electrical and Computer Engineering,
**Department of Systems Design Engineering,
University of Waterloo,
Waterloo, ON N2L-3G1 Canada.

ABSTRACT

Autoregressive (AR) models have been shown to be effective models of the human vocal tract during voicing. However the most common model of speech for enhancement purposes, AR process excited by white noise, fails to capture the periodic nature of voiced speech. Speech synthesis researchers have long recognized this problem and have developed a variety of sophisticated excitation models, however these models have yet to make an impact in speech enhancement. We have chosen one of the most common excitation models, the four-parameter LF model of Fant, Liljencrants and Lin, and applied it to the enhancement of individual voiced phonemes. Comparing the performance of the conventional white-noise-driven AR, an impulsive-driven AR, and AR based on the LF model shows that the LF model yields a substantial improvement, on the order of 1.3 dB.

1. INTRODUCTION

Broadly speaking, the field of speech enhancement is interested in addressing three (not necessarily compatible) objectives: the improvement of the perceptual quality of noisy speech, the immunization of speech encoders against input noise, and the improvement of the performance of speech recognition systems in the presence of noise[4]. This paper investigates the first of these: in our context, the speech enhancement problem concerns the estimation of "clean" (de-noised) speech $\hat{x}(t)$ from noisy speech $z(t)$.

To be sure, the de-noising problem has been well-studied and estimation techniques are quite mature, however our research objective in this paper is very specific. Rather than developing turnkey enhancement systems, applicable to large speech corpa, whose performance have been improving over time but where the *limits* to performance are unclear, we seek to establish performance benchmarks or limits by studying speech-enhancement in *detail* for individual phonemes under arbitrarily well-characterized circumstances. Although such circumstances might appear artificial, they are essential in understanding the intrinsic factors which limit enhancement performance — an understanding which may improve enhancement algorithms in much broader, less constrained conditions.

Our approach involves *model-based* speech enhancement, in which prior stochastic models of the clean speech and of the corrupting noise are used for estimation. Clearly, accurate estimation requires that these models be robust and faithful representations of reality. By far the most common choice of model is a white-noise excited autoregressive process; we will discuss the limitations of white-noise excitation and will propose two more appropriate ones, based on the concept of the source-filter theory of speech production [5].

Although models based on speech production are regularly used in speech synthesis, their application to speech enhancement is new. Furthermore, there is a subtle, but extremely important, difference in how models are used for synthesis versus enhancement: for synthesis a model generates a speech signal *ex nihilo*, whereas for enhancement the model must be made consistent or compatible with all of the non-ideal vagaries of "real" speech (e.g., rapidly time-varying pitch periods). For example, a shift in the time-origin is irrelevant in speech synthesis, however a shift in the relative origins between a model and a speech signal can lead to catastrophic mismatch in enhancement.

The following section will review autoregressive models, followed by two proposed alternatives and a presentation of results.

2. BACKGROUND

Autoregressive (AR), or all-pole, models driven by white noise have been one of the most popular models for representing speech waveforms [2]. An N-th order AR model represents speech $x(t)$ as a linear combination of past speech samples added to white noise:

$$x(t) = \sum_{i=1}^{N} a_i x(t-i) + w(t) \qquad (1)$$

where $w(t)$ is a zero mean, white Gaussian process with variance σ^2 and $\{a_i\}$ is the set of AR coefficients. The popularity of the AR model stems from its simplicity, and because the human vocal tract during voicing can be modeled by an all-pole system[2]. Furthermore, although unvoiced speech and nasals introduce zeros into the system, since the zeros of the transfer function of the vocal tract lie inside the unit circle, they can be approximated by an all-pole system with sufficiently many poles[2]. Finally, because (1) can be rewritten in state-space form, the Kalman filter[10] can be used to compute the optimal estimates $\hat{x}(t)$.

The flaws in this AR model (1) become apparent when

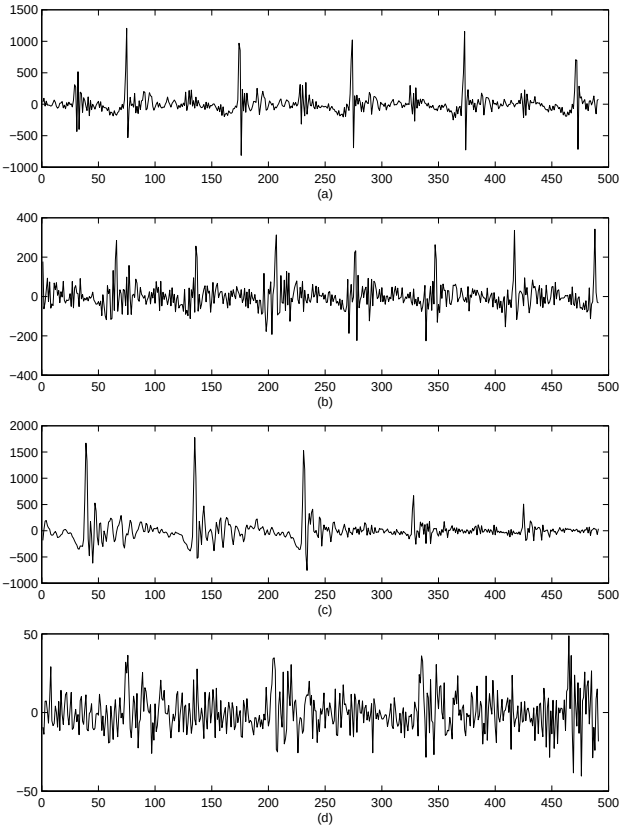

Figure 1: Plots of AR residuals for four voiced speech phones: (a) front /ae/, (b) diphthong /ay/, (c) semivowel /r/, (d) nasal /n/. The model (1) predicts that each of these signals be white (random) — clearly incorrect.

the model residuals,

$$x(t) - \sum_{i=1}^{N} a_i x(t-i) \tag{2}$$

are examined, as shown in Figure 1. The model (1) asserts that these residuals should be white (random), an assertion which is flatly contradicted by the figure, since obvious quasi-periodic (deterministic) components are present in each of the four phones shown. The remainder of this paper investigates more consistent alternatives to $w(t)$ in (1).

3. NEW VOICE SOURCE MODELS AND PARAMETER ESTIMATION TECHNIQUES

3.1. Impulsive Models

The obvious flaw with the conventional autoregressive model (1) is that the vocal tract is modeled as being driven by white noise, whereas vowels, diphthongs, semivowels and nasals all have quasi-periodic glottal pulse excitation of the vocal tract. Quasi-periodic pulses are produced when air is forced through the glottis, causing the vocal cords to vibrate and periodically interrupt the subglottal airflow.

We can begin to account for a quasi-periodic vocal-tract excitation by modifying the AR forcing function:

$$x(t) = \sum_{i=1}^{N} a_i x(t-i) + \quad w(t) + a_{_{N+1}} u_{_I}(t) \tag{3}$$

where $a_{_{N+1}}$ is the amplitude of the driving term, and where $u_{_I}(t)$ is a train of impulses:

$$u_{_I}(t) = \sum_j \delta(t - t_j) \tag{4}$$

where the times t_j mark the times of the glottal pulses. The times are approximated manually from the residual signal (2) in which the pulses are conspicuous, followed by an automated local peak-finder to guarantee accurate positioning.

The inclusion of the weighted excitation term in (3) implies that the conventional covariance LP analysis[2], which applies to (1), needs to be modified. The principle of covariance LP analysis is just parameter estimation to minimize a least-squares criterion

$$C_K = \sum_{t=0}^{K-1} \epsilon(t)^2 \tag{5}$$

where K is length of the speech segment (frame) being processed, and where the error is given by the model residual

$$\epsilon(t) = x(t) - \sum_{i=1}^{N} \hat{a}_k x(t-i) + \hat{a}_{_{N+1}} u_{_I}(t) \tag{6}$$

The optimal parameters are found by finding the roots of the squared error (5),

$$\frac{\partial C_K}{\partial \hat{a}_j} = 0, \quad 1 \le j \le N \quad \& \quad \frac{\partial C_K}{\partial \hat{a}_{_{N+1}}} = 0 \tag{7}$$

leading to a set of linear equations:

$$\begin{bmatrix} \Phi(i,j) & \Psi(i,0) \\ \Psi^T(i,0) & R_u \end{bmatrix} \begin{bmatrix} \hat{\mathbf{a}} \\ \hat{a}_{_{N+1}} \end{bmatrix} = \begin{bmatrix} \Phi(i,0) \\ \Psi(0,0) \end{bmatrix} \tag{8}$$

which is easily solved, using the Cholesky decomposition, for the unknowns $\hat{\mathbf{a}} = [\hat{a}_1, \ldots, \hat{a}_N]^T$ and $\hat{a}_{_{N+1}}$. The terms in the square matrix are the correlation terms: Φ the cross-correlation matrix of clean speech,

$$\Psi(i,j) = \sum_{t=0}^{K-1} x(t-i)u(t-k) \tag{9}$$

the cross-correlation between clean speech and the excitation, and

$$R_u = \sum_{t=0}^{K-1} u_{_I}(t)^2 \tag{10}$$

the energy (zero-lag autocorrelation) of the excitation $u_{_I}$.

Figure 2 shows the impulse-AR residuals (6) for the same four phonemes of Figure 1. In general the residual pulses in Figure 2 are thinner or narrower than before, but still conspicuously present. Clearly a more sophisticated voice-source model is required.

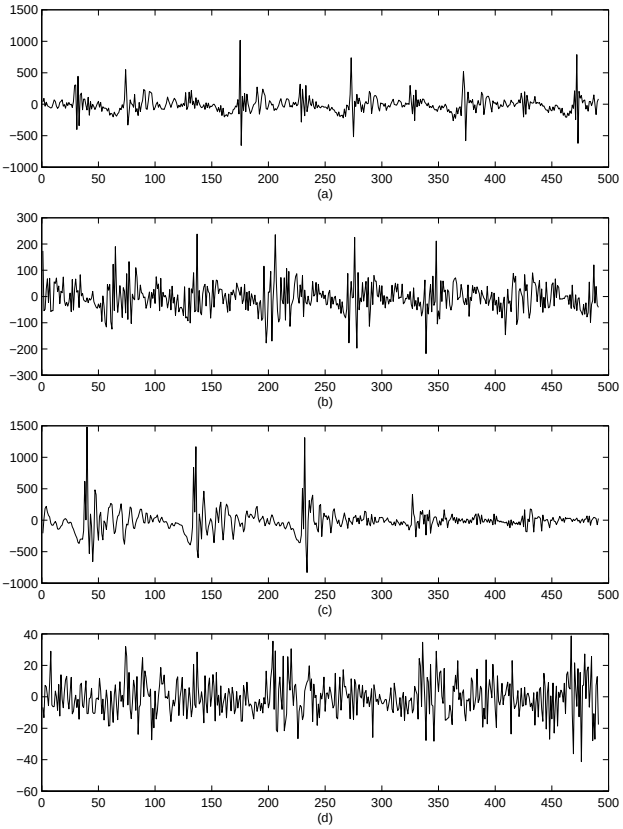

Figure 2: AR residuals for the impulsive model (3) for the voiced phonemes of Figure 1.

3.2. LF Model

An impulsive model is a highly simplified approximation of the human voice. Indeed, impulsive-driven systems were found to make poor speech synthesizers, so the synthesis field has proposed a number of more complex glottal pulse models[1, 7]. Of these, the four parameter LF model[7] proposed by Fant, Liljencrants and Lin has widely been used practically in speech synthesis and theoretically in speech analysis[3].

The LF excitation model, sketched in Figure 3, is the derivative of the LF glottal pulse function, and is parameterized in terms of

- t_c – the fundamental period,
- t_p – the instant of maximum flow,
- t_e – the instant of maximum glottal closing,
- t_a – exponential recovery time constant.

The LF model is then given by

$$u_{LF}(t) = \begin{cases} e^{\alpha t}\sin\omega_g t & t \le t_e \\ \frac{-1}{\beta t_a}[e^{-\beta(t-t_e)} - e^{-\beta(t_c-t_e)}] & t_e \le t \le t_c \end{cases} \tag{11}$$

where α, β satisfy the transcendental equations

$$1 - e^{-\beta(t_c-t_e)} = \beta t_a$$
$$e^{\alpha t_e}\sin(\pi t_e/t_p) = -1,$$

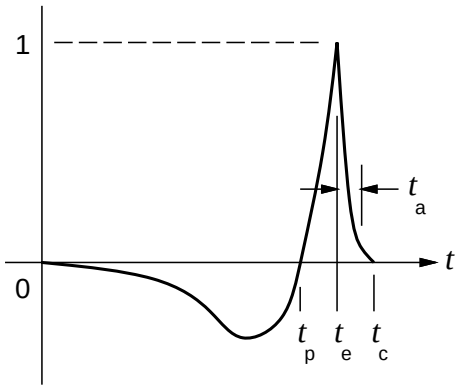

Figure 3: The LF deterministic excitation model.

leading to the revised AR model

$$x(t) = \sum_{i=1}^{N} a_i x(t-i) + \quad w(t) + a_{N+1}u_{LF}(t). \tag{12}$$

The main challenge with using model $u_{LF}(t)$ in (12) is the need to estimate the seven parameters $t_c, t_e, t_p, t_a, \alpha, \beta, a_{N+1}$. Only a_{N+1} enters the problem linearly, so it is solved using least-squares as in (8). The point of maximum glottal closing t_e is set to coincide with the impulsive points t_j determined in the previous section, leaving five remaining parameters to be found by nonlinearly optimizing the mean-squared error C_K (equivalently the output SNR) via coordinate gradient descent.

4. DISCUSSION & RESULTS

Speech data, used for testing our speech models, were extracted from twenty sentences spoken by ten female and ten male speakers from the TIMIT data base. In order to assess enhancement limits we learn the model parameters separately for each phoneme. The phoneme boundaries given in the TIMIT data base were initially used to accomplish this separation, followed by the inspection of spectrograms and temporal plots to verify the exact phoneme boundaries. Each speech signal, representing a single phoneme, is segmented into frames of $K = 256$ data points. The Kalman filter was used as the estimation algorithm, using one of three different models (1),(3),(12). The speech signals were corrupted with additive white noise to an SNR of 5dB; for each signal the identical noise process was added, so that output SNR results are meaningfully comparable.

Figure 4 shows the AR-LF residuals, paralleling the earlier results of Figures 1 and 2. In moving from the purely impulsive to the LF model, the top two panels, in particular, show a reduction and thinning of residual spikes and exhibit less deterministic structure. A close examination of the figures reveals a substantial limitation in u_I which begins to be addressed in u_{LF}: an impulse $\delta(t)$ is exactly one sample wide, whereas the width of the residual spikes in Figure 1 and of the peak in u_{LF} are clearly sampling-rate dependent, and are frequently, although not always, more than one sample in width. A similar issue can be raised in terms

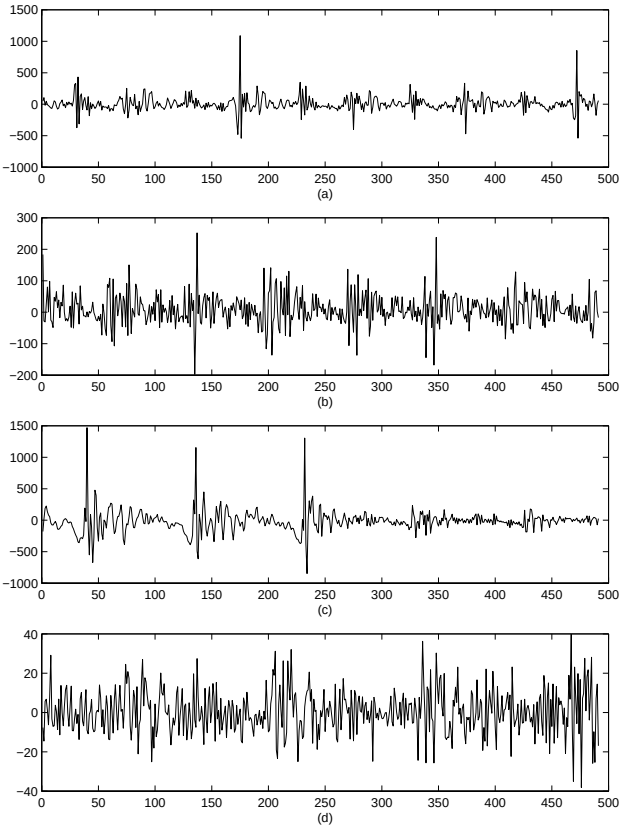

Figure 4: AR residuals for the LT model (12) for the voiced phonemes of Figures 1, 2.

of sampling origin: a single glottal burst may, depending on the sampling origin, be captured as a single impulse or as two smaller impulses. An impulse-train u_I cannot properly address this issue, whereas u_{LF} is a continuous signal and lends itself naturally to resampling.

To assess the models more objectively a global measure of SNR was used,

$$\mathrm{SNR} = 10 \log \frac{\sum_{t=1}^{J} x^2(t)}{\sum_{t=1}^{J} [x(t) - \hat{x}(t)]^2} \qquad (13)$$

where J is the total length of the speech signal.

Table 1 summarizes the SNR improvement for each of the three proposed models, tested on eight different voiced phonemes. Most importantly, consistent and nontrivial improvements in SNR are realized, first by the impulsive model, then additionally by the LF model, for *all* voiced phonemes tested.

5. CONCLUSIONS

The main objective of this work was to find appropriate models for voiced speech enhancement, and to investigate limits to performance. The obtained results are promising, however there is excellent potential for improvement. At the very least, the LF model needs to have subsampling issues addressed, and the nonlinear parameter optimization via coordinate descent may be sensitive to local minima and should be robustified.

Phone	Output SNR (dB) White Noise AR-Model	Output SNR (dB) Impulsive AR-Model	Output SNR (dB) LF-base AR-Model
Front vowel /ae/	8.061	9.31	10.053
Front vowel /iy/	9.216	10.078	10.702
Mid vowel /ah/	8.292	9.171	9.500
Back vowel /ao/	9.592	10.264	10.836
Diphthong /ay/	8.781	9.639	10.587
Diphthong /iu/	9.106	9.694	10.035
Semivowel /r/	8.302	9.376	9.535
Nasal /n/	9.395	9.633	10.111

Table 1: Enhancement results for voiced speech; input SNR of 5dB, AR order $N = 10$.

The other substantial step is the automation of procedures undertaken manually in this work — the identification of phoneme boundaries and global pulse locations; algorithms already exist which can accomplish these tasks to varying degrees of accuracy. The exciting challenge then is the development of algorithms, applicable to sentences or whole conversations, which are capable of achieving enhancement performance on the aggregate scale at the same level which has been demonstrated, in principle, for individual phonemes.

6. REFERENCES

[1] T.V.Ananthapadmanabha "Acoustic Analysis of Voice Source Dynamics," *STL-QPSR 2-3*, pp. 1-24, 1984

[2] B.S.Atal and S.L.Hanauer. "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave," *J. Acoust. Soc. Am.*, V50 #2, pp. 637-655, 1971

[3] D.G.Childers and C.K.Lee. "Vocal quality Factors: Analysis, Synthesis and Perception," *J. Acoust. Soc. Am.*, V90 #5, pp. 2394-2410, 1991

[4] Y.Ephraim. "Statistical-Model-Based Speech Enhancement Systems," *Proc. IEEE*, V80 #10, pp. 1526-1555, 1992

[5] G.Fant. "Acoustic Theory of Speech Production," Mouton's Co., Hague, 1960

[6] G.Fant. "The Source Filter Concept in Voice Production," *STL-QPSR 1*, pp. 21-37, 1981

[7] G.Fant, J. Liljencrants and Q. Lin. "A Four Parameter Model of Glottal Flow," *STL-QPSR 4*, pp. 1-12, 1985

[8] J.D.Gibson, B.Koo and S.D.Gray. "Filtering of Colored Noise for Speech Enhancement and Coding," *IEEE Trans. Audio Signal Processing*, V39 #8, pp. 1732-1741, 1991

[9] J.S.Lim and A.V.Oppenheim. "All-pole Modeling of Degraded Speech," *IEEE Trans. ASSP*, V26 #3, pp. 197-210, 1978

[10] K.K.Paliwal and A.Basu. " A Speech Enhancement Method Based on Kalman Filtering," *Proc. IEEE ICASSP*, pp. 177-180, 1987