EURASIP Journal on Applied Signal Processing 2003:12, 1238–1249

Hammerstein Model for Speech Coding

Jari Turunen

*Department of Information Technology, Tampere University of Technology, Pori, Pohjoisranta 11,
P.O. Box 300, FIN-28101 Pori, Finland
Email: jari.j.turunen@tut.fi*

Juha T. Tanttu

*Department of Information Technology, Tampere University of Technology, Pori, Pohjoisranta 11,
P.O. Box 300, FIN-28101 Pori, Finland
Email: juha.tanttu@tut.fi*

Pekka Loula

*Department of Information Technology, Tampere University of Technology, Pori, Pohjoisranta 11,
P.O. Box 300, FIN-28101 Pori, Finland
Email: pekka.loula@tut.fi*

Received 7 January 2003 and in revised form 19 June 2003

A nonlinear Hammerstein model is proposed for coding speech signals. Using Tsay's nonlinearity test, we first show that the great majority of speech frames contain nonlinearities (over 80% in our test data) when using 20-millisecond speech frames. Frame length correlates with the level of nonlinearity: the longer the frames the higher the percentage of nonlinear frames. Motivated by this result, we present a nonlinear structure using a frame-by-frame adaptive identification of the Hammerstein model parameters for speech coding. Finally, the proposed structure is compared with the LPC coding scheme for three phonemes /a/, /s/, and /k/ by calculating the Akaike information criterion of the corresponding residual signals. The tests show clearly that the residual of the nonlinear model presented in this paper contains significantly less information compared to that of the LPC scheme. The presented method is a potential tool to shape the residual signal in an encode-efficient form in speech coding.

Keywords and phrases: nonlinear, speech coding, Hammerstein model.

1. INTRODUCTION

Due to the solid theory underlying linear systems, the most widely used methods for speech coding up to the present day have been the linear ones. Numerous modifications of those methods have been proposed. At the same time, however, the application of nonlinear methods to speech coding has gained more and more popularity. An early example of nonlinear speech coding is the a-law/μ-law compression scheme in pulse code modulation (PCM) quantization. With a-law (8 bits per sample) or μ-law (7 bits per sample) compression, the total saving of 4–5 bits per sample can be achieved compared to linear quantization (12 bits per sample). However, these nonlinearities do not involve modeling and are purely based on the fact that the human hearing system has logarithmic characteristics.

Probably, the most well-known linear model-based speech coding scheme is the linear predictive coding (LPC), where model parameters together with the information about the residual signal need to be transmitted. For example, in the ITU-T G.723.1 speech encoder, the linear predictive filter coefficients can be represented using only 24 bits while the excitation signal requires either 165 bits (6.3 kbps mode) or 134 bits (5.3 kbps mode). In analysis-by-synthesis coders, such as G.723.1, the excitation signal is used for speech synthesis to excite the linear filter to produce synthesized speech sound similar to the original speech sound. The G.723.1 codec itself is robust and has successfully served multimedia communications for years. However, only 13–15% of the encoded speech frame contains information about the filter while 85–87% is spent on the excitation signal. In other words, over 80% of the transmitted data is information that the linear filter cannot model.

The residual signal in speech coding is a modeling error that is left out after filtering. The excitation signal has similar characteristics to the residual signal and it is used to excite the inverse linear filtering process in the decoder.

A lot of research has been done recently to study the nonlinear properties and to find an efficient model for the speech signal. For example, Kubin shows in [1] that there are several nonlinearities in the human vocal tract. Also, several studies suggest that linear models do not sufficiently model the human vocal tract [2, 3]. In [4], Fackrell uses a bispectral analysis in his experiments. He found that generally there is no evidence of quadratic nonlinearities in speech, although, based on the Gaussian hypothesis, voiced sounds have a higher bicoherence level than expected. In some papers, efforts have been made to model speech using fluid dynamics, as in [5]. In [6, 7, 8] chaotic behavior has been found mainly in vowels and some nasals like /n/ and /m/. In [9], speech signal is modeled as a chaotic process. However, these types of models have not proved to be able to characterize speech in general, including consonants, and therefore they have not become widely used.

In other studies, hybrid methods, combining linear and nonlinear structures, have been applied to speech processing. For example, in [10] nonlinear artificial excitation is modulated with a linear filter in an analysis-synthesis system while in [11, 12] Teager energy operator has been found to give good results in different speech processing contexts.

Another approach to dealing with nonlinearities in speech is to use systems that can be trained according to some training data. These systems must have the capability of learning the nonlinear characteristics of speech. In [13, 14, 15, 16, 17, 18], radial basis function and multilayer perceptron neural networks were tested as short- and long-term predictors in speech coding. The results in these studies are encouraging. However, the use of neural networks always entails a risk that the results may be totally different if the copy of the originally reported system is built from scratch using the same number of neural nodes and so forth even when the same training data is used. The platform may be different; the way how the training is performed and the possibility of over- and undertraining will affect the training result. Also, a mathematical analysis of the model structure which the neural network has learned is usually not feasible.

All these studies suggest that nonlinear methods enhance speech processing when compared to the traditional linear speech processing systems. However, the form of the fundamental nonlinearity in speech is still unknown. From a practical point of view, the speech model should be easy to implement, and computationally efficient, and the number of transmitted parameters should be as low as possible, or at least have some benefit when compared to traditional linear coding methods. It may be possible that speech contains different types of linear/nonlinear characteristics, for example, vowels have either chaotic features or types of higher-order nonlinear features, while consonants may be modeled by random processes.

Based on the ideas presented above, a parametric model consisting of a weighted combination of linear and nonlinear features and capable of identifying the model parameters from the speech data could be useful in speech coding. One such model is the Hammerstein model that has been used in different types of contexts, for example, in biomedical signal processing and noise reduction in radio transmission, but not for speech modeling in the context of coding. Recently, the parameter identification of the Hammerstein model has turned from an iterative to a fast and accurate process in the approach presented in [19, 20, 21]. The proposed method is derived from system identification and control science. It has been used, for example, in biological signal processing [22] and acoustic echo cancellation [23], but it can also be used in speech processing. In this paper, we present the use of a noniterative Hammerstein model parameter identification applied to speech modeling in coding purposes.

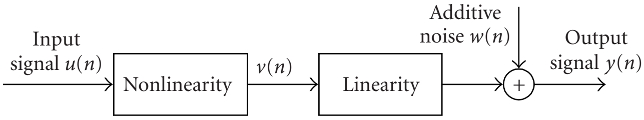

Figure 1: Hammerstein model.

2. MATHEMATICAL BACKGROUND

2.1. Hammerstein model

The Hammerstein model consists of a static nonlinearity followed by a linear time-invariant system as defined in [24] and presented in Figure 1. The Hammerstein model can be viewed as an extension of the conventional linear predictive structure in speech processing. The motivation to implement this model in speech processing can be traced to the exact mathematical background of the combined nonlinear and linear subsystem parameter identification. It is possible to augment static nonlinearity in front of the LPC system with fixed coefficients, but the Hammerstein model offers, in the presented form, frame-by-frame adaptive coefficient optimization for both nonlinear and linear subsystems. Traditionally, the Hammerstein model is viewed as a black-box model, but in speech coding, the inverse of the Hammerstein model must also be found in order to decode the compressed signal in the destination. The coding-based aspects are discussed later in this paper.

In Figure 1, the nonlinear subsystem includes a preselected set of nonlinear functions. The monotonicity of the nonlinear functions, required in the decoder, is the only limitation that restricts the selection and the number of the nonlinear functions. The linear subsystem consists of base functions whose order is not limited.

The general form of the model is as follows:

$$y(n) = \sum_{k=0}^{p-1} b_k B_k(q) \sum_{i=1}^{r} a_i g_i(u(n)) + w(n), \tag{1}$$

where $a = [a_1, \ldots, a_r]^T \in \mathbb{R}^r$ are the unknown nonlinear coefficients, g_i represents the set of nonlinear functions, r is the number of nonlinear functions and coefficients, B_k are finite impulse response (FIR), Laguerre, Kautz, or other base functions, and $b = [b_0, \ldots, b_{p-1}]^T \in \mathbb{R}^p$ are the linear base function coefficients. The integer p is the linear model order. The signal $w(n)$ represents the modeling error or additive noise in this case. In our coding scheme, the original speech signal is used as the model input $u(n)$ while $y(n)$ can be viewed as a residual, that is, a part of the input signal which the model is not able to represent. We assume that the mean of the original speech signal has been removed and the amplitude range has been normalized between $[-1, 1]$.

As it can be seen from (1), the parameter coefficient sets (b_k, a_i) and $(\alpha b_k, \alpha^{-1} a_i)$ are equivalent. In order to obtain unique identification, either b_k or a_i is assumed to be normalized.

Based on the model given by (1), the following two vectors can be formed: the parameter vector θ, containing the multiplied nonlinear and linear coefficient combinations, and the data vector ϕ, containing the input signal passed through the individual components of the set of nonlinear functions g_i.

The parameter vector θ, parameter matrix Θ_{ab}, and data vector ϕ can be defined as

$$\theta = \left[b_0 a_1, \ldots, b_0 a_r, \ldots, b_{p-1} a_1, \ldots, b_{p-1} a_r\right]^T, \tag{2a}$$

$$\Theta_{ab} = \begin{bmatrix} a_1 b_0 & a_1 b_1 & \cdots & a_1 b_{p-1} \\ a_2 b_0 & a_2 b_1 & \cdots & a_2 b_{p-1} \\ \vdots & \vdots & & \vdots \\ a_r b_0 & a_r b_1 & \cdots & a_r b_{p-1} \end{bmatrix} = ab^T, \tag{2b}$$

$$\begin{aligned} \phi = \big[&B_0(q) g_1(u(n)), \ldots, B_0(q) g_r(u(n)), \ldots, \\ &B_{p-1}(q) g_1(u(n)), \ldots, B_{p-1} g_r(u(n))\big]^T. \end{aligned} \tag{3}$$

Using vectors θ and ϕ, (1) can be written as

$$y(n) = \theta^T \phi + w(n). \tag{4}$$

The set of values $\{y(n),\ n = 1, \ldots, N\}$ can be considered as a frame and expressed as a vector Y_N. For the whole frame, (4) can be written in a matrix form:

$$Y_N = \Phi_N^T \theta + W_N, \tag{5}$$

where Y_N, Φ_N, and W_N can be expressed as

$$\begin{aligned} Y_N &\hat{=} \left[y(1), y(2), \ldots, y(N)\right]^T, \\ \Phi_N &\hat{=} \left[\phi(1), \phi(2), \ldots, \phi(N)\right]^T, \\ W_N &\hat{=} \left[w(1), w(2), \ldots, w(N)\right]^T. \end{aligned} \tag{6}$$

Estimating θ by minimizing the quadratic error $\| W_N \|_2^2$ between the real signal and the calculated model output in (5) (least squares estimate) can be expressed as [25]

$$\hat{\theta} = \left(\Phi_N \Phi_N^T\right)^{-1} \Phi_N Y_N. \tag{7}$$

The $\hat{\theta}$ vector obtained using (7) contains products of the elements of the coefficient vectors a and b in (2a). To separate the individual coefficients vectors a and b, the elements of θ can be organized into a block column matrix, corresponding to the matrix defined in (2b), as

$$\hat{\Theta}_{ab} = \begin{bmatrix} \hat{\theta}_1 & \cdots & \hat{\theta}_p \\ \hat{\theta}_{p+1} & \cdots & \hat{\theta}_{2p} \\ \vdots & \ddots & \vdots \\ \hat{\theta}_{(r-1)p+1} & \cdots & \hat{\theta}_{rp} \end{bmatrix}. \tag{8}$$

From this matrix, the model parameter estimates $\hat{a} = [\hat{a}_1, \ldots, \hat{a}_r]^T$ and $\hat{b} = [\hat{b}_0, \ldots, \hat{b}_{p-1}]^T$ can be solved using economy-size singular value decomposition (SVD) [25], which yields factorization

$$\hat{\Theta}_{ab} = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix} \begin{bmatrix} V_1^T \\ V_2^T \end{bmatrix} \tag{9}$$

which is partitioned so that $\dim(U_1) = \dim(a)$ and $\dim(V_1) = \dim(b)$. The block Σ_1 is in fact the first singular value σ_1^2 of $\hat{\Theta}_{ab}$. It is proved in [21] that the optimal parameter vector estimates are obtained as follows:

$$(\hat{a}, \hat{b}) = \arg\min_{a,b} \left\{ \left|\left|\hat{\Theta}_{ab} - ab^T\right|\right|_2^2 \right\} = (U_1, V_1 \Sigma_1), \tag{10}$$

$$\hat{a} = U_1, \tag{11}$$

$$\hat{b} = V_1 \Sigma_1. \tag{12}$$

In addition, it is proved in [21] that (11) and (12) are the best possible parameter estimates for parameter vectors a and b. It is also proved in [21] that under rather mild conditions on the additive noise $w(n)$ and input signal $u(n)$ in (1), $\hat{a}(N) \to a$ and $\hat{b}(N) \to b$, with probability 1 as $N \to \infty$. Notice however that in (11) and (12) it is assumed that $\|a\|_2 = 1$, that is, the a-parameter vector is normalized. More details can be found in [19, 20, 21].

2.2. Nonlinearity test for speech

In order to find out nonlinearities in speech, it must be tested somehow. There are some methods available that will measure the signal nonlinearity against a hypothesis and will give a statistical number as a result. Several objective tests have been developed to estimate the proportion of nonlinearities in time series. In the following, the nonlinearity of a conversational speech signal is analyzed using Tsay's test [26], which is a modification of Keenan nonlinearity test [27] having several benefits over Keenan test yet maintaining the same simplicity. The Keenan test is originally based on Tukey's nonadditivity test [28].

Tsay's test was selected for our experiments due to its simplicity and usability for time series. It uses linear autoregressive (AR) parameter estimation, which has proven to work with speech data in several other contexts. The idea of this test is to remove the linear information and delayed regression information from the data and see how much information remains in these two residuals. These two residuals are then regressed against each other and the regression error is obtained. The output of the test is the information of the two residual signals normalized by the energy of the error.

A stationary time series $y(n)$ can be expressed in the form

$$\begin{aligned} y(n) = \mu &+ \sum_{i=-\infty}^{\infty} b_i e(n-i) + \sum_{i,j=-\infty}^{\infty} b_{ij} e(n-i) e(n-j) \\ &+ \sum_{i,j,k=-\infty}^{\infty} b_{ijk} e(n-i) e(n-j) e(n-k) + \cdots, \end{aligned} \tag{13}$$

where μ is the mean level of $y(n)$, b_i, b_{ij}, and b_{ijk} are the first-, second-, and third-order regression coefficients of $y(n)$, and $e(n-i)$, $e(n-j)$, and $e(n-k)$ are independent and identically distributed random variables. If one of the higher-order coefficients (b_{ij}), $(b_{ijk}), \ldots$ is nonzero, then $y(n)$ is nonlinear. If, for example, b_{ij} is nonzero, then it will be reflected in the diagnostics of the fitted linear model if the residuals of the linear model are correlated with $y(n-i)y(n-j)$, a quadratic nonlinear term. Tsay's test for nonlinearities is motivated by this observation and performed by the following way using only the first- and second-order regression terms.

(1) Regress $y(n)$ on vector $[1, y(n-1), \ldots, y(n-M)]$ and obtain the residual estimate $\hat{e}(n)$. The regression model is then

$$y(n) = K_n\Phi + e(n), \tag{14}$$

where $K_n = [1, y(n-1), \ldots, y(n-M)]$ is the vector consisting of the past values of y, and $\Phi = \{\Phi(0), \Phi(1), \ldots, \Phi(M)\}^T$ is the first-order autoregressive parameter vector, where M presents the order of the model and $n = [M+1, \ldots, \text{sample\_size}]$.

(2) Regress the vector Z_n on K_n and obtain the residual estimate vector $\hat{X}_n$. The regression model is

$$Z_n = K_n H + X_n, \tag{15}$$

where Z_n is a vector of length $(1/2)M(M+1)$. The transpose of Z_n and Z_n^T are obtained from the matrix

$$[y(n-1), \ldots, y(n-M)]^T [y(n-1), \ldots, y(n-M)] \tag{16}$$

by stacking the column elements on and below the main diagonal. The second-order regression parameter matrix is denoted by H, and $n = [M+1, \ldots, \text{sample\_size}]$.

(3) Regress $\hat{e}(n)$ on $\hat{X}(n)$ and obtain the error $\hat{\varepsilon}(n)$:

$$\hat{e}(n) = \hat{X}(n)\beta + \varepsilon(n), \quad n = [M+1, \ldots, \text{sample\_size}], \tag{17}$$

where β is the regression parameter matrix of two residuals obtained from (1) and (2).

(4) Let $\hat{F}$ be the F ratio of the mean square of regression to the mean square of error:

$$\begin{aligned}\hat{F} = {} & \frac{\left(\sum \hat{X}(n)\hat{e}(n)\right)\left(\sum \hat{X}(n)^T\hat{X}(n)\right)^{-1}}{(1/2)M(M+1)\sum \hat{\varepsilon}(n)^2} \\ & \times \left(\sum \hat{X}(n)^T\hat{e}(n)\right)\left(n - M - \frac{1}{2}M(M+1) - 1\right),\end{aligned} \tag{18}$$

which is used to represent the value of rejection of the null hypothesis of linearity. It follows approximately the F-distribution with degrees of freedom $n_1 = (1/2)M(M+1)$ and $n_2 = \text{sample\_size} - (1/2)M(M+3) - 1$. A more detailed analysis of the nonlinearity test can be found in [26].

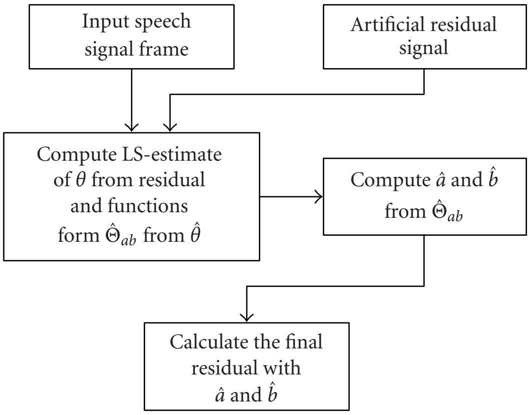
Input speech signal frame
Artificial residual signal
Compute LS-estimate of θ from residual and functions form $\hat{\Theta}_{ab}$ from $\hat{\theta}$
Compute $\hat{a}$ and $\hat{b}$ from $\hat{\Theta}_{ab}$
Calculate the final residual with $\hat{a}$ and $\hat{b}$

Figure 2: Structure of the identification system.

3. THE PROPOSED MODEL FOR SPEECH CODING

In case of the Hammerstein model, the process that alters the input signal can be viewed as a black-box model. This model has an input signal and an output signal which is the black-box process modification of the input signal. In order to identify this kind of model parameters, we need both signals, model input $u(n)$ and output $y(n)$. The original speech signal can be used as $u(n)$, but $y(n)$ is unknown.

In the speech coding environment, the output signal $y(n)$ is viewed as a residual. It is desirable that $y(n)$ be represented with as few parameters as possible. For estimating model parameters in our experiments, we used three different artificial residual signals: white noise, unit impulse, and codebook-based signals. The selection and properties of these signals will be discussed later in this paper.

If the model structure is adequate, applying the model with the estimated parameters gives a true residual which resembles the artificial residual signal used for the estimation. Therefore, we can assume that the information contained in the true residual can also be represented using few parameters, a codebook or coarse quantization. The structure of the system proposed for the parameter estimation is presented in Figure 2.

The identification algorithm is forced to find the coefficients for the nonlinear and linear parts of the current model so that the final residual is very close to the artificial residual signal. The least squares estimate of the parameter vector θ is calculated from the artificial output vector and the input which is fed through the nonlinear and linear parts of the model in question. The block column matrix $\hat{\Theta}_{ab}$ is formed, and nonlinear and linear coefficient estimates $\lfloor \hat{a}, \hat{b} \rfloor$ are obtained. The proposed system attached to the speech coding framework is presented in Figure 3.

In Figure 3, the whole coding-decoding system using the Hammerstein model is presented. The residual of the Hammerstein process can be compressed using coarse quantization, codebook-based, or any other suitable compressing scheme. This information, together with the model coefficients, is packed for transmission.

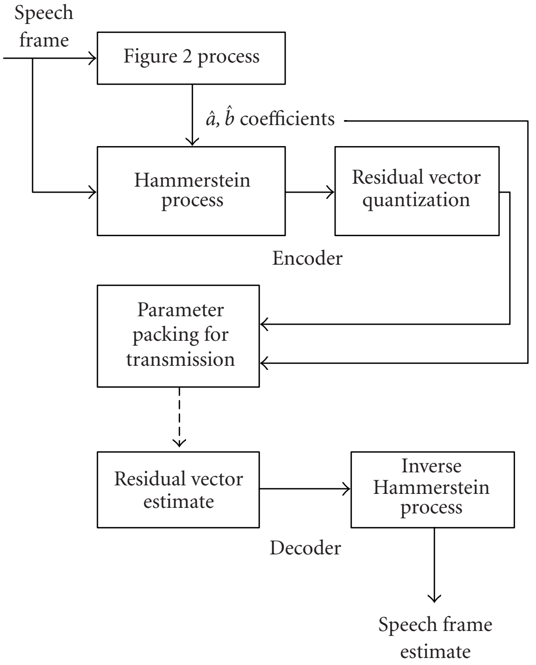
Speech frame
Figure 2 process
$\hat{a}, \hat{b}$ coefficients
Hammerstein process
Residual vector quantization
Encoder
Parameter packing for transmission
Residual vector estimate
Inverse Hammerstein process
Decoder
Speech frame estimate

Figure 3: The Hammerstein mode-based speech coder.

The aim of this paper, however, is to evaluate the capability of the Hammerstein model for speech modeling by estimating the amount of information contained in the residual signal.

As expressed by (1) and Figure 1, the Hammerstein model consists of two submodels, a linear and a nonlinear one. In our experiments, FIR base functions

$$B_k(q) = q^{-k} \tag{19}$$

were used in the linear substructure. These base functions are easy to implement. In the decoder, the inverse model has to be implemented. This is usually not a problem for the linear part of the model.

The nonlinear substructure of the Hammerstein model can be viewed as a preprocessor, turning the nonlinear task of speech modeling into a linearly solvable one. In the decoder, finding the inverse of the nonlinear subsystem might constitute a problem. For the inverse to be unique, the functions must be monotonic in the amplitude range $[-1, 1]$. The inverse can be implemented, for example, using numerical methods or lookup tables, depending on the type of functions used. The nonlinear subsystem is a memoryless unit and stability can be ensured by checking whether the nonlinear coefficients are below the predetermined threshold values. The linear subsystem must have its poles inside the unit circle. The parameter quantization also affects the encoded/decoded speech quality. However, depending on the system, the proposed Hammerstein model can be built on an analysis-by-synthesis system where the quantized parameters are part of the encoding process and thus try to maximize the quality of the encoded speech.

In the Hammerstein model, nonlinearity is a kind of preprocessing to the speech sound before linear processing. In this case, the nonlinear part is assumed to reduce or modify the features of the speech signal that the linear part cannot model.

4. RESULTS

4.1. Nonlinearities in speech

We tested about 89 minutes of conversational speech sampled at 8000 Hz. The speech samples consisted of professional speakers' talks, interviews, and telephone conversations in low-noise conditions. Three frame lengths were used: 160, 240, and 320 samples. All the speech samples were normalized so that the amplitude range was between $[-1, 1]$. Frames were nonoverlapping and for each frame length two tests were performed—one with rectangular-windowed frames and the other with Hamming windowing. Hamming windowing was selected due to its popularity in some speech-related applications and to see if the windowing itself would affect the results. In our analysis, the model order M was $M = 10$ and the number of samples was equal to the frame length. The frame energy was calculated as the sum of absolute values, and if this sum was less than the predetermined threshold 15, the frame was regarded as a silent frame and was left out. In some cases also frames containing very low-amplitude /s/ phonemes might have been left out. Of all the test data, about 45 minutes were judged as silent frames and 44 minutes had an amplitude high enough to perform the test. The test results are presented in Table 1. In the table, "$p = 99\%$" means that the null hypothesis confidence limit was 99 percent and the numbers listed in the corresponding column indicate the number of frames for which the F-distribution confidence limit was exceeded.

This test clearly demonstrates the existence of nonlinearities in speech in over 80% of the frames. This correlation may be caused by the fact that the frame length was fixed so that a single frame might have contained parts of different types of phonemes. Table 1 also shows that the percentage of nonlinear frames increases significantly due to windowing. When the Hamming-windowed frames are compared with the frames with rectangular windowing, it seems that Hamming windowing enhances the nonlinear properties of the speech signal. This is due to the nonoverlapped Hamming windowing, where the edges of the frames may affect the result.

In Table 2, the results of hand-labeled phonemes from TIDIGITS database /a/, /s/, and /k/ are presented. The frame length was fixed, and in /s/ and /a/ the frame is taken from the middle of the phoneme. In the case of /k/, the plosive is within the frame in a way that the rest is silence or near background noise level.

The test also shows that there are nonlinearities in phonemes /a/, /s/, and /k/ as seen in Table 2. The vowel /a/ seems to be highly nonlinear while the amount of nonlinearities in /s/ is very low. In the case of /s/ phonemes, their frequency content is near the white noise frequency content,

Table 1: Tsay nonlinearity test results of conversational speech.

Frame size	Window	No. of all frames	No. of nonlinear frames $p = 99\%$	No. of nonlinear frames $p = 99.9\%$	No. of nonlinear frames $p = 99.99\%$
160	Rectangular	74401	69117 (92.9%)	64761 (87.0%)	59660 (80.2%)
160	Hamming	74401	73932 (99.4%)	73159 (98.3%)	71828 (96.5%)
240	Rectangular	71795	68879 (95.9%)	66956 (93.3%)	64645 (90.0%)
240	Hamming	71795	71524 (99.6%)	71066 (99.0%)	70331 (98.0%)
320	Rectangular	65613	63036 (96.1%)	61903 (94.3%)	60678 (92.5%)
320	Hamming	65613	65302 (99.5%)	64811 (98.8%)	64087 (97.7%)

Table 2: Tsay nonlinearity test results for hand-labeled phonemes.

Frame size	phoneme	No. of all frames	No. of nonlinear frames $p = 99\%$	No. of nonlinear frames $p = 99.9\%$	No. of nonlinear frames $p = 99.99\%$
256	/a/	670	670 (100%)	669 (99.8%)	669 (99.8%)
256	/s/	669	175 (26.2%)	100 (15.0%)	59 (8.8%)
256	/k/	224	194 (86.6%)	181 (80.8%)	163 (72.8%)

and thus the linear model will be appropriate to present the phoneme accurately. The phoneme /k/ is a plosive burst that has fast changes, and thus it seems to include nonlinearities.

4.2. Modeling nonlinearities of speech with Hammerstein model

In order to estimate the model parameters, artificial residuals must be chosen. Artificial residual, in this context, means a signal with properties that are also required for the true residual after the Hammerstein model process. Although ideally the residual would be zero, estimating the model parameters according to the zero residual will end up with the trivial result of zero-valued coefficients. The artificial residuals chosen for our experiments are shown in Figure 4.

The white noise residual was uniformly distributed with amplitude range $[-0.1, 0.1]$. The second residual was obtained by collecting a 1024-vector codebook from true residuals of a tenth-order LPC filter from which the periodical spikes were removed. The codebook vectors were 32-sample long and the artificial residual for our experiment was formed by combining 8 randomly selected vectors from the codebook. As the third residual, a unit impulse was used. There are lots of good candidate signals available, but the ones were chosen for the following reasons: first, the random signal is very difficult to model with linear methods; second, the codebook-based signal was chosen because of the fact that it is widely used in modeling and vector quantization; and third, unit impulse was chosen due to its simple form.

The nonlinearity chosen for the experiments is

$$\begin{aligned} g(u(n)) &= a_1 g_1(u(n)) + a_2 g_2(u(n)), \\ g_1(u(n)) &= u(n), \\ g_2(u(n)) &= \operatorname{sign}(u(n)) \, |u(n)|^{3/2}. \end{aligned} \tag{20}$$

The exponent 3/2 can be changed to almost any finite number, but it was selected for demonstrative purposes, in this case, based on our knowledge. The purpose was to show the behavior of the Hammerstein model using a very simple model structure.

The linear substructure constitutes a first-order FIR filter:

$$L(v(n)) = \sum_{k=0}^{1} b_k B_k(q) = b_0 v(n) + b_1 v(n-1). \tag{21}$$

The selection of the linear substructure is analyzed more in the discussion. The modeling experiment was done 670 times for hand-labeled phonemes /a/. The Hammerstein model with the three artificial residuals is shown in Figure 4. The used sampling frequency of the signals was 8000 Hz. For comparison, the coefficients of the third-order LPC model are also presented. The distribution of the estimated coefficients is shown in Figure 5. The first linear parameters are normalized to one, and thus left out from Figure 5.

Figure 5 shows that in this test with variable phoneme /a/ data, the Hammerstein model coefficient values are finite and stable. Interestingly, the deviation of the nonlinear parameters is limited to a very narrow area. Also the distribution of the linear component in the unit-impulse signal case is more concentrated near -0.5 when compared to the other linear parameter deviations. The coefficient parameters with phonemes /k/ and /s/ are distributed in the same manner, however the peaks are in different places (the coefficients of /k/ are deviating more than the coefficients of /a/ or /s/). This concentration property is useful especially in speech coding and possibly in speech recognition purposes.

In Figure 6, the results of two phoneme modeling experiments are shown. Two sections of female speech, one voiced (/a/) and another unvoiced (/s/), were modeled using structures of the Hammerstein and LPC models similar to those in

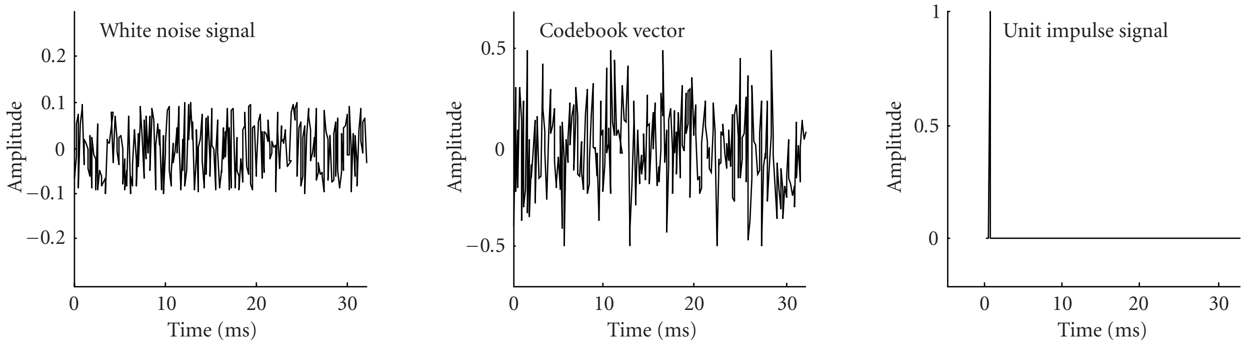

Figure 4: Three artificial residual signals: the leftmost is white noise, the middle signal is codebook vector, and the rightmost is unit impulse with zero padding.

the first experiment. The estimated coefficients of the Hammerstein model for all the experimental cases are presented in Table 3 for speech sections /a/ and /s/, respectively.

Figure 6 shows that the Hammerstein model gives a significantly reduced residual compared to the LPC model. This indicates the adaptation capability of the model in amplitude. For our experiments we selected a simple nonlinear function of (20). By optimizing the form of the nonlinearity, the performance of the Hammerstein model could be further improved. The coefficients shown in Table 3 indicate the different emphasis with different artificial residual even with this small model. The results presented in Table 4 in the case of phoneme /a/ are a typical case of the results presented in Figure 5 with dotted vertical line.

Figure 7 shows male vowel results. The coefficients are more oriented to the edges of the statistical data presented in Figure 5 (dash-dotted vertical lines) when compared to the female speech. However, both the processed female and male speech frames suggest that signal residuals processed by the Hammerstein model have smaller amplitude levels when compared to the linear prediction-based residual. Although the Hammerstein model is formed from simple linear and nonlinear substructures, the coefficient determination algorithm gives different weights to the linear and nonlinear coefficients, computed with different artificial residuals. The true residual output from the Hammerstein model is not the optimal one, due to the selected nonlinearity, but it indicates the adaptation possibilities that will be acquired by carefully selecting the nonlinear functions.

The performance of the model can be evaluated by measuring the amount of information in the true residual signal using, for example, Akaike's information criterion (AIC). However, AIC is not directly targeted in speech processing because the purpose of AIC is to measure the amount of information stored in the signal in the sense of information theory.

The AIC can be defined as

$$\mathrm{AIC}(i) = N \ln \hat{\sigma}_i^2 + 2i, \tag{22}$$

where N is the number of data samples, $\hat{\sigma}$ is the maximum likelihood estimate of the white noise variance for an assumed autoregressive process, and i is the assumed autoregressive model order. AIC estimates the information criterion for the signal by using estimation error from model and the model order number.

We calculated the AIC value for 670 /a/, 669 /s/, and 224 /k/ phoneme residuals for the codebook-based artificial residual (residual 2). The AIC model order $i = 6$ was chosen to be greater than the linear model order (LPC order $= 4$) used in the tests. The codebook artificial residual was chosen for the modeling for the reason that it is the worst signal in the sense that it may contain LPC-based information, and this information may be transferred to the true residual signal. For comparison, the consequent residuals for LPC were calculated. The averaged results are shown in Table 5.

The table shows clearly that the true residual of the Hammerstein model contains significantly less information compared to the LPC residual. This again indicates the ability of the Hammerstein model to capture the features of the speech signal.

5. DISCUSSION

The potential of nonlinear methods in speech processing is tremendous. The assumption that speech contains nonlinearities can be indicated with different types of tests, including Tsay's test for nonlinearity. This test shows clearly that speech contains nonlinear features. As shown in this paper, the Hammerstein model is applicable to speech coding. Figures 6 and 7 indicate that the shape of the artificial residual used in estimating the model parameters is significant as the true residuals differ from each other. This suggests that speech signal contains variable information that cannot be modeled using a single artificial residual but the residual shaping is possible to a certain extent. However, Figure 5 shows that the nonlinear parameter deviation is small in all the Hammerstein model experiment cases, and this property might be useful in speech recognition purposes. The AIC results also indicate that the information is clearly reduced

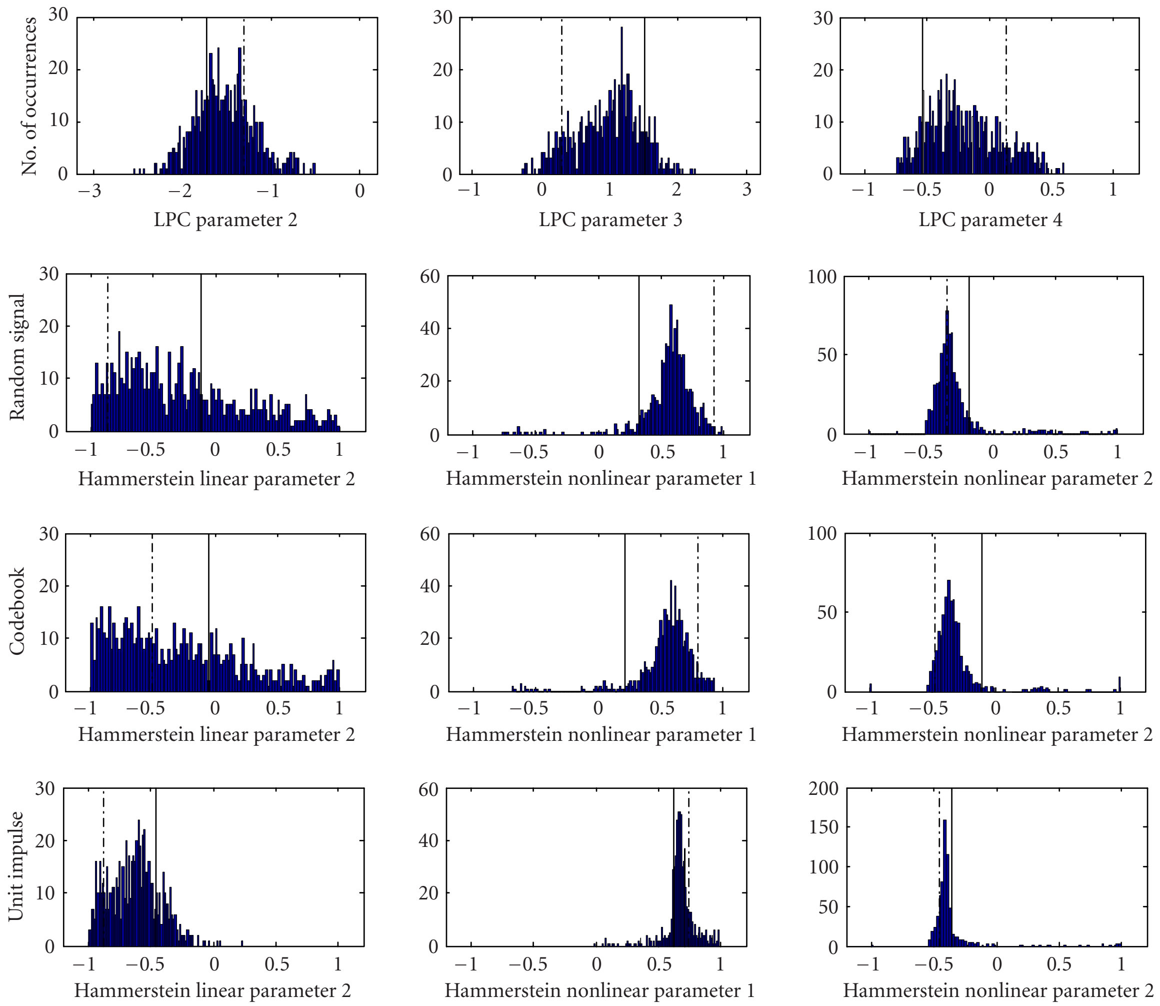

Figure 5: The distribution of LPC and Hammerstein model parameters for phoneme /a/. The first linear parameters are normalized to 1, and thus left out from the figure. The dotted vertical line indicates the phoneme /a/ parameter values of Table 3 and the dash-dotted line indicates the respective parameter values of Table 4.

when the residuals of the Hammerstein and LPC models were compared although the tests were performed with a third-order LPC filter against the Hammerstein model with a first-order linear subsystem, one nonlinearity, and linear scaling.

Usually, in speech processing, either the source or the output of the model in question is unknown. However, in the proposed model, both input and output signals are needed. In all speech coding, the purpose is to send as small a number of parameters as possible to the destination while keeping the quality of the decoded speech as good as possible. This means that the model, intended to characterize the vocal tract, works so well that either there is no residual signal after the filtering process or the residual can be presented with very few parameters. On the other hand, the expectation of the zero residual can be dangerous when using input-output system parameter identification processes. There is a risk that the identification process will give zero-coefficients to all nonlinear and linear filter components and there is no true filtering at all. This is why some type of residual must exist in the identification process.

Codec using the Hammerstein model requires the inversion of the nonlinear function in the decoder. This means that the nonlinear function must be monotonic in the selected amplitude range in order to reconstruct the estimate of the original speech signal. The Hammerstein model allows the usage of a very wide range of nonlinear functions, for example, polynomials, exponential series $\{e^{0.1x}, e^{0.2x}, e^{0.3x}, \dots\}$, and so forth, including their mixed combinations. In speech coding, however, the amount of information to be transmitted must be as low as possible. Therefore, finding the suitable combination of nonlinear components, characteristic to speech signal, is very important. This issue requires a lot of research in the future.

Another important issue is the balance between the linear and nonlinear substructures. For example, in our

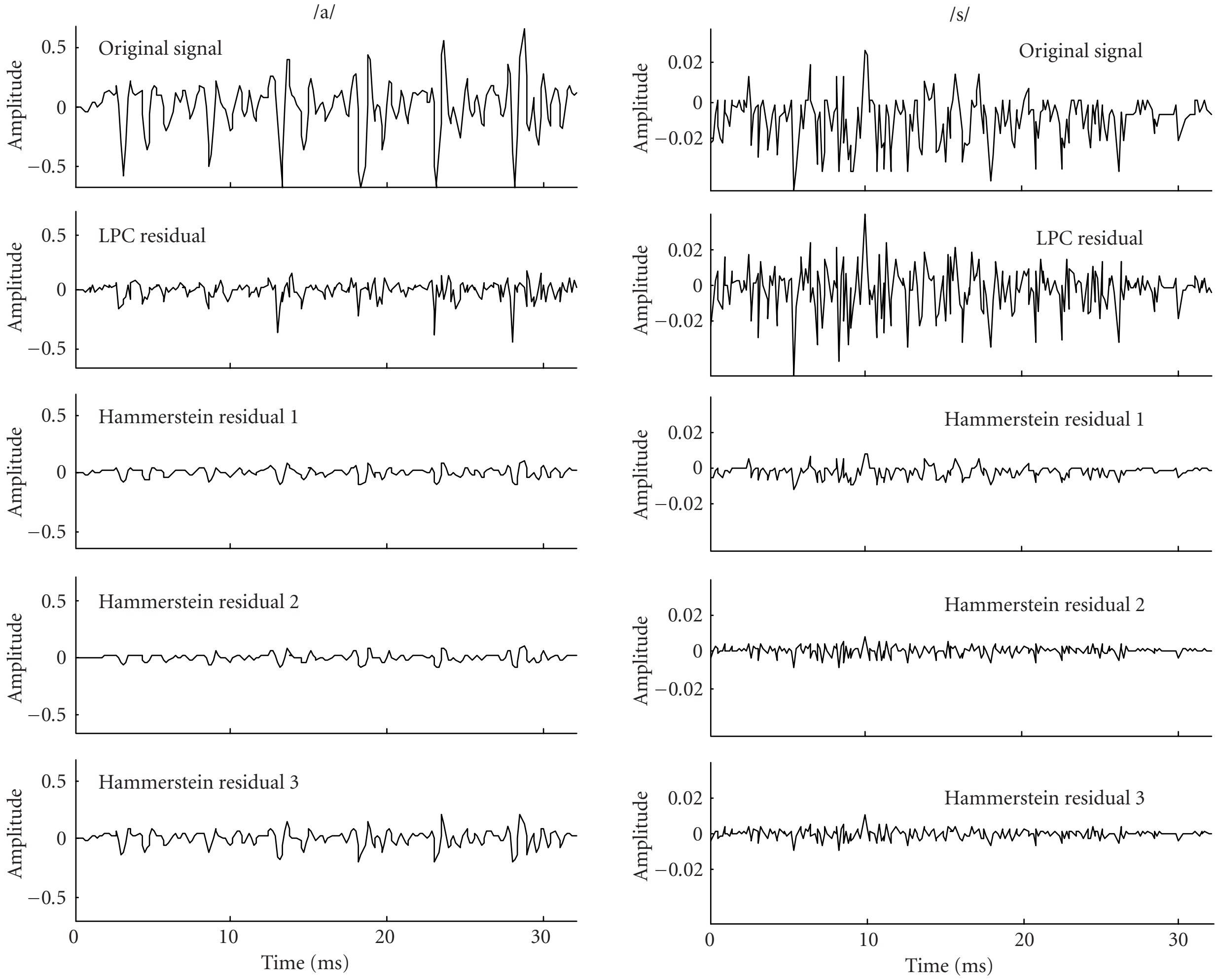

Figure 6: Comparison between the original signal, LPC-filtered residual signal, and Hammerstein residuals in the case of a random artificial residual (Hammerstein residual 1), codebook-based artificial residual (Hammerstein residual 2), and unit-impulse residual (Hammerstein residual 3). The artificial residuals are the input signals for the model, and residuals presented in the figure are the true output of the model.

preliminary tests, the selected nonlinear series function

$$\begin{aligned} g_1(u(n)) &= a_0u(n), \\ g_2(u(n)) &= a_1 \tan(0.5u(n)), \\ g_3(u(n)) &= a_2 \tan(0.75u(n)), \\ g_4(u(n)) &= a_3 \tan(0.875u(n)), \\ g_5(u(n)) &= a_4 \tan(0.9688u(n)), \\ g_6(u(n)) &= a_5 \tan(u(n)), \end{aligned} \tag{23}$$

was used as nonlinearity in the Hammerstein model together with a tenth-order linear filter. The nonlinearity reduced the information too much so that after quantization in the coding process the decoder oscillated and produced unwanted frequencies in the decoded speech signal. However, with carefully balanced combined nonlinear and linear structure, it is possible to quantize the final residual with very coarse quantization scheme and obtain a stable speech estimate as in [29, 30]. In these studies, the stability of the inverse system was obtained by checking the linear system stability and, if necessary, correcting it by using the minimum phase correction.

The form of the linear subsystem is also important. Either autoregressive moving average (ARMA), AR, or MA model can be used. Another choice to be made concerns the basis functions. Orthonormal bases with fixed poles, Kautz bases, and so forth provide a good foundation for different ARMA structures, but finding the poles and/or zeros from the current speech frame before calculating the coefficients of the model will increase the overall computational load. Another problem with the ARMA model is that the parameter estimation method may lead to poles within the z-plane unit circle and zeros outside the unit circle. The latter nonminimum phase property will lead to unstability of the inverse system. The zeros of the numerator and denominator must lie within the unit circle as the inverse system is needed in the decoder. It is possible to place the zeros and poles inside the unit circle by performing minimum phase correction, that is,

Table 3: The coefficient values for phonemes /a/ and /s/ in Figure 6.

Linear coefficient values for /a/				Linear coefficient values for /s/			
LPC	Hamm. 1	Hamm. 2	Hamm. 3	LPC	Hamm. 1	Hamm. 2	Hamm. 3
1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
−1.73	−0.12	−0.05	−0.46	−0.50	−0.05	−0.81	−0.60
Nonlinear coefficient values				Nonlinear coefficient values			
1.52	0.33	0.21	0.62	0.06	0.28	0.20	0.24
−0.53	−0.19	−0.11	−0.36	−0.29	−0.17	−0.11	−0.13

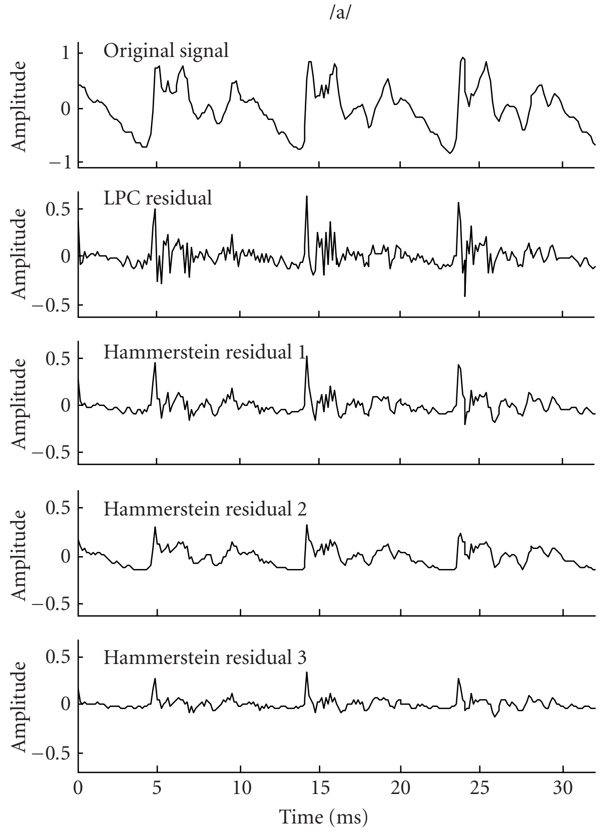

Figure 7: The original speech frame /a/ taken from male speech.

Table 4: The coefficient values for phoneme /a/ in Figure 7.

Linear coefficient values for /a/			
LPC	Hamm. 1	Hamm. 2	Hamm. 3
1.00	1.00	1.00	1.00
−1.31	−0.86	−0.50	−0.87
Nonlinear coefficient values			
0.30	0.92	0.80	0.74
0.14	−0.37	−0.48	−0.46

Table 5: The AIC results.

Signal	AIC	RMS
/a/ LPC residual	−5.31	0.11
/a/ Hammerstein residual	−7.00	0.09
/s/ LPC residual	−9.73	0.01
/s/ Hammerstein residual	−14.03	< 0.01
/k/ LPC residual	−9.09	0.01
/k/ Hammerstein residual	−12.52	< 0.01

moving the zeros and poles outside the unit circle to their reciprocal locations. The base functions utilizing pole location information need also extra calculations for defining the pole locations.

By using the rational orthonormal bases with fixed poles (OBFP) in the linear subsystem, the estimation accuracy can be improved compared to the Kautz, Laguerre, and FIR bases where the knowledge of only one pole can be incorporated [20]. The OBFP can utilize the knowledge of multiple poles in the orthonormal system and they are defined as

$$B_k(q) = \left(\frac{\sqrt{1 - |\xi_k|^2}}{q - \xi_k}\right) \prod_{m=0}^{k-1} \left(\frac{1 - \overline{\xi}_m q}{q - \xi_m}\right), \tag{24}$$

where q is the unit delay, ξ_k is the kth pole, and $\overline{\xi}_k$ is its conjugate. This structure is valid if the poles of the basis functions are real. If the poles are complex conjugate pairs, which is the case in speech analysis, the base function conversion to real pole bases maintaining orthonormality is described in [31]. Using ARMA filter with the Hammerstein model would be a fascinating idea but the calculation of the ARMA filter by adding up the base functions with their weighted coefficients will increase the number of total calculations. Also, in speech processing, there is no a priori knowledge of the locations of zeros and/or poles of the linear subsystem. This knowledge must be obtained using LPC or other methods before the actual model parameter identification. Naturally, this will increase the number of calculations in the speech frame analysis.

Computational complexity is always a big concern. The Hammerstein model identification process needs more computation compared to LPC model. However, the overhead of calculations and memory demands, using the method described above, comes only from the nonlinear parameter identification. Calculations can be reduced by carefully balancing the nonlinear/linear combination. This means that it is possible to reduce the number of linear components by properly selecting the nonlinear components when compared to traditional linear models.

The model presented here can be used in frame-by-frame adaptive parameterization speech coding, and it provides a stable filter and function coefficient estimation method. The parameter identification is fast and the calculation overhead comes only from the nonlinear parameter identification compared to traditional linear filter analysis methods. The inner structure of the nonlinear and linear blocks can be selected quite freely with only few practical limitations.

ACKNOWLEDGMENT

We would like to thank Professor Tarmo Lipping from Tallinn Technical University, Estonia, for his useful suggestions and improvements.

REFERENCES

[1] G. Kubin, "Nonlinear processing of speech," in *Speech Coding and Synthesis*, W. Kleijn and K. Paliwal, Eds., pp. 557–610, Elsevier Science B.V., Amsterdam, The Netherlands, November 1995.

[2] J. Thyssen, H. Nielsen, and S. Hansen, "Non-linear short-term prediction in speech coding," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '94)*, pp. 185–188, Adelaide, Australia, April 1994.

[3] J. Schroeter and M. Sondhi, "Speech coding based on physiological models of speech production," in *Advances in Speech Signal Processing*, S. Furui and M. Sondhi, Eds., pp. 231–268, Marcel Dekker, New York, NY, USA, 1992.

[4] J. Fackrell, *Bispectral analysis of speech signals*, Ph.D. thesis, Department of Electronics and Electrical Engineering, University of Edinburgh, Edinburgh, Scotland, September 1996.

[5] P. Mergell and H. Herzel, "Modelling biphonation—the role of the vocal tract," *Speech Communication*, vol. 22, pp. 141–154, 1997.

[6] T. Miyano, A. Nagami, I. Tokuda, and K. Aihara, "Detecting nonlinear determinism in voiced sounds of Japanese vowel /a/," *International Journal of Bifurcation and Chaos*, vol. 10, no. 8, pp. 1973–1979, 2000.

[7] M. Banbrook, S. McLaughlin, and I. Mann, "Speech characterization and synthesis by nonlinear methods," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 1, pp. 1–17, 1999.

[8] F. Martínez, A. Guillamón, J. Alcaraz, and M. Alcaraz, "Detection of chaotic behaviour in speech signals using the largest Lyapunov exponent," in *Proc. IEEE 14th International Conference on Digital Signal Processing (DSP '02)*, pp. 317–320, Santorini, Greece, July 2002.

[9] B. Townshend, "Nonlinear prediction of speech," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '91)*, pp. 425–428, Toronto, Canada, May 1991.

[10] W. Wokurek, "Time-frequency analysis of the glottal opening," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '97)*, pp. 1435–1438, Munich, Germany, April 1997.

[11] P. Maragos, T. Quatieri, and J. Kaiser, "Speech nonlinearities, modulations, and energy operators," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '91)*, pp. 421–424, Toronto, Canada, May 1991.

[12] J. Hansen, L. Gavidia-Ceballos, and J. Kaiser, "A nonlinear operator-based speech feature analysis method with application to vocal fold pathology assessment," *IEEE Transactions on Biomedical Engineering*, vol. 45, no. 3, pp. 300–313, 1998.

[13] N. Ma and G. Wei, "Speech coding with nonlinear local prediction model," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '98)*, pp. 1101–1104, Seattle, Wash, USA, May 1998.

[14] A. Kumar and A. Gersho, "LD-CELP speech coding with nonlinear prediction," *IEEE Signal Processing Letters*, vol. 4, no. 4, pp. 89–91, 1997.

[15] M. Faúndez-Zanuy, F. Vallverdú, and E. Monte, "Nonlinear prediction with neural nets in ADPCM," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '98)*, pp. 345–349, Seattle, Wash, USA, May 1998.

[16] F. Díaz-de-Maria and A. Figueiras-Vidal, "Nonlinear prediction for speech coding using radial basis functions," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '95)*, pp. 788–791, Detroit, Mich, USA, May 1995.

[17] M. Birgmeier, H.-P. Bernhard, and G. Kubin, "Nonlinear long-term prediction of speech signals," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '97)*, pp. 1283–1286, Munich, Germany, April 1997.

[18] M. Birgmeier, "A fully Kalman-trained radial basis function network for nonlinear speech modeling," in *Proc. IEEE International Conference on Neural Networks (ICNN '95)*, pp. 259–264, Perth, Australia, November–December 1995.

[19] J. Gómez and E. Baeyens, "Identification of multivariable Hammerstein systems using rational orthonormal bases," in *Proc. 39th IEEE Conference on Decision and Control (CDC '00)*, vol. 3, pp. 2849–2854, Sydney, Australia, December 2000.

[20] J. Gómez and E. Baeyens, "Identification of nonlinear systems using orthonormal bases," in *Proc. IASTED International Conference on Intelligent Systems and Control (ISC '01)*, pp. 126–131, Tampa, Fla, USA, November 2001.

[21] E. Bai, "An optimal two-stage identification algorithm for Hammerstein-Wiener nonlinear systems," *Automatica*, vol. 34, no. 3, pp. 333–338, 1998.

[22] D. Westwick and R. Kearney, "Identification of a Hammerstein model of the stretch reflex EMG using separable least squares," in *Proc. 22nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS '00)*, pp. 1901–1904, Chicago, Ill, USA, July 2000.

[23] L. S. H. Ngia and J. Sjöberg, "Nonlinear acoustic echo cancellation using a Hammerstein model," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '98)*, pp. 1229–1232, Seattle, Wash, USA, May 1998.

[24] L. Ljung, *System Identification: Theory for the User*, Prentice-Hall, Englewood Cliffs, NJ, USA, 1987.

[25] G. Golub and C. Van Loan, *Matrix Computations*, North Oxford Academic, Oxford, UK, 1983.

[26] R. Tsay, "Nonlinearity tests for time series," *Biometrika*, vol. 73, no. 2, pp. 461–466, 1986.

[27] D. Keenan, "A Tukey nonadditivity-type test for time series nonlinearity," *Biometrika*, vol. 72, no. 1, pp. 39–44, 1985.

[28] J. Tukey, "One degree of freedom for nonadditivity," *Biometrics*, vol. 5, pp. 232–242, 1949.

[29] J. Turunen, P. Loula, and J. Tanttu, "Effect of adaptive nonlinearity in speech coding," in *Proc. 2nd WSEAS International Conference on Signal, Speech and Image Processing (ICOSSIP '02)*, pp. 3401–3406, Koukounaries, Skiathos Island, Greece, September 2002.

[30] J. Turunen, J. Tanttu, and P. Loula, "New model for speech residual signal shaping with static nonlinearity," in *Proc. 7th International Conference on Spoken Language Processing (ICSLP '02)*, pp. 2145–2148, Denver, Colo, USA, September 2002.

[31] B. Ninness and F. Gustafsson, "A unifying construction of orthonormal bases for system identification," *IEEE Transactions on Automatic Control*, vol. 42, no. 4, pp. 515–521, 1997.

Jari Turunen received his M.S. and Licentiate of Technology degrees in 1998 and 2000, respectively, from Tampere University of Technology. Currently he is preparing his Ph.D. dissertation in telecommunication and speech processing.

Juha T. Tanttu was born in Tampere, Finland, on November 25, 1957. He obtained his M.S. and Ph.D. degrees in electrical engineering from Tampere University of Technology in 1980 and 1987, respectively. From 1984 to 1992, he held various teaching and research positions at the Control Engineering Laboratory of Tampere University of Technology. He currently holds professorship of information technology at Tampere University of Technology, Pori.

Pekka Loula received his M.S. and Ph.D. degrees in information technology in 1987 and 1994, respectively, from Tampere University of Technology. Currently he holds a telecommunication professorship at Tampere University of Technology, Pori. He is the Author of over 100 publications in the field of telemedicine, telecommunication, and signal processing. His current research interests cover topics such as IP-based networks, broadband telecommunication, QoS aspects, and telecommunication applications.