14th European Signal Processing Conference (EUSIPCO 2006), Florence, Italy, September 4-8, 2006, copyright by EURASIP

WEIGHTED NONLINEAR PREDICTION BASED ON VOLTERRA SERIES FOR SPEECH ANALYSIS

Karl Schnell and Arild Lacroix

Institute of Applied Physics, Goethe-University Frankfurt
Max-von-Laue-Str. 1, D-60438 Frankfurt am Main, Germany
email: {Schnell, Lacroix}@iap.uni-frankfurt.de

ABSTRACT

The analysis of speech is usually based on linear models. In this contribution speech features are treated using nonlinear statistics of the speech signal. Therefore a nonlinear prediction based on Volterra series is applied segment-wise to the speech signal. The optimal nonlinear predictor can be determined by a vector expansion. Since the statistics of a segment is estimated a window function is integrated into the estimation procedure. Speech features are investigated representing the prediction gain between the linear and the nonlinear prediction. The analyses of speech signals show that the nonlinear features correlate with the glottal pulses. The integration of an appropriate window function into the prediction algorithm plays an important part for the results.

1. INTRODUCTION

The speech production is usually described by a linear model leading to linear prediction. However, nonlinear components also exist in the speech signal [1]. The voiced excitation is caused by vibrations of the vocal folds which can be described by a nonlinear oscillator; additionally nonlinear fluid dynamics are effective. Nonlinear systems and operators can be used for speech analysis [2],[3]. Here nonlinear components of the speech signal are estimated by nonlinear prediction. The nonlinear system of a Volterra series is used for the prediction. The estimation of the predictor can be achieved by an adaptive algorithm like LMS or RLS [4]. Another approach for the estimation is to minimize the prediction error of individual signal segments, which can be applied to coding [5] or speech generation [6]. In this contribution an approach is discussed minimizing the least square error of the prediction error of a signal segment by a vector expansion. To utilize the nonlinear prediction for extraction of features of speech segments a window function is integrated into the estimation procedure.

2. NONLINEAR PREDICTION

2.1 Prediction Based on Volterra Series

In the linear prediction a signal value $x(n)$ is estimated by a linear combination of last signal values $x(n-k)$ with $k>0$. In the case of a nonlinear predictor based on a Volterra system, the prediction considers products of last values, too. Without loss of generality systems are treated with the first and second order Volterra kernels only. The prediction error e for a signal x is defined as the difference between the actual value x and the estimated value $\hat{x}(n)$:

$$e = x - \hat{x}:$$

$$e(n) = x(n) - \sum_{k=1}^{N} h_1(k)\cdot x(n-k) - \sum_{i=1}^{M}\sum_{k=1}^{i} h'_2(i,k)\cdot x(n-k)x(n-i)\,. \tag{1}$$

In (1) the second-order kernel h_2 is assumed symmetrically so that the coefficients h'_2 are used instead of h_2 with $h'_2(i,k) = h_2(i,k)$ for $i=k$ and $h'_2(i,k) = 2\cdot h_2(i,k)$ for $i \neq k$ since the second sum in (1) ends with *k*=*i*. The coefficients of the predictor are estimated by a minimization of the least square error:

$$\sum\nolimits_n e(n)^2 \to \min\,.$$

Since the prediction error of a signal segment is considered, the use of a window function is useful. If the window function $w(n)$ is applied directly to the signal $x(n)$ the resulting weighted signal is $u(n) = w(n)\cdot x(n)$. For the calculation of a single error $u(n) - \hat{u}(n)$ corresponding to eq. (1) the last values $u(n-k) = w(n-k)\cdot x(n-k)$ have different weights, since $w(n-k)$ depends on *k*. This effect is stronger in the case of the second kernel, due to the products: $u(n-k)\cdot u(n-i) = \underline{w(n-k)\cdot w(n-i)}\cdot x(n-k)\cdot x(n-i)$. To solve this problem the window function has to be applied to the error $e(n)$ yielding the weighted error $e_w(n) = w(n)\cdot e(n)$. Applying to eq. (1) results in

$$e_w(n) = w(n)\cdot x(n) - \sum_{k=1}^{N} h_1(k)\cdot w(n)\cdot x(n-k) - \sum_{i=1}^{M}\sum_{k=1}^{i} h'_2(i,k)\cdot w(n)\cdot x(n-k)x(n-i)\,. \tag{2}$$

The predictor coefficients are determined by minimizing the weighted error

$$\sum\nolimits_n e_w(n)^2 \to \min\,. \tag{3}$$

2.1.1 Vector Based Nonlinear Prediction

The prediction is applied to a segment of the speech signal therefore it is convenient to describe the signals by vectors.

For the prediction estimation the analyzed weighted signal $u(n) = w(n)\cdot x(n)$ is described by the vector

$$\boldsymbol{u} = \left(w(0)\cdot x(0), w(1)\cdot x(1),\ldots, w(K)\cdot x(K)\right)^{\mathrm{T}}$$

of length *L'*=*K*+1. The prediction error $e_w(n)$ contains last values $x(n-k)$ so that the definition of the vectors $\boldsymbol{u}_i$ containing the shifted signals with fixed weights is convenient:

$$\begin{aligned}\boldsymbol{u}_i &= (w(0)\cdot x(-i), w(1)\cdot x(1-i), \ldots \\ &\quad \ldots, w(-i)\cdot x(0), \ldots, w(K)\cdot x(K-i))^{\mathrm{T}}.\end{aligned} \qquad (4)$$

It has to be mentioned that $\boldsymbol{u}_i$ is not the shifted signal $u(n)$ since the weights w in (4) are independent of *i*. Furthermore for the description of products of last values the definition of the vectors $\boldsymbol{u}_{i,k}$ is defined

$$\boldsymbol{u}_{i,k} = \left(w(0)x(-i)x(-k),\ w(1)\,x(1-i)x(1-k),\ \ldots\right)^{\mathrm{T}}.$$

The estimation of the weighted signal values $u(n)$ can be described by the vector $\hat{\boldsymbol{u}}$:

$$\hat{\boldsymbol{u}} = \sum_{i=1}^{N} h_1(i)\cdot \boldsymbol{u}_i + \sum_{i=1}^{M}\sum_{k=1}^{i} h_2'(i,k)\cdot \boldsymbol{u}_{i,k}\ . \qquad (5)$$

By these definitions the prediction problem can be described by the vector equation $\boldsymbol{e}_w = \boldsymbol{u} - \hat{\boldsymbol{u}}$. Since the error depends on the number *N* of linear coefficients and *M* of nonlinear coefficients, the error $\boldsymbol{e}_w \to \boldsymbol{e}_w^{N,M}$ can be extended by the superscripts *N* and *M*:

$$\boldsymbol{e}_w^{N,M} = \boldsymbol{u} - \sum_{i=1}^{N} h_1(i)\cdot \boldsymbol{u}_i - \sum_{i=1}^{M}\sum_{k=1}^{i} h_2'(i,k)\cdot \boldsymbol{u}_{i,k} \qquad (6)$$

respectively

$$\boldsymbol{e}_w^{N,M} = \begin{pmatrix} e_w^{N,M}(0) \\ e_w^{N,M}(1) \\ \vdots \\ e_w^{N,M}(K) \end{pmatrix} = \begin{pmatrix} w(0)x(0) \\ w(1)x(1) \\ \vdots \\ w(K)x(K) \end{pmatrix} - h_1(1)\begin{pmatrix} w(0)x(-1) \\ w(1)x(0) \\ \vdots \\ w(K)x(K-1) \end{pmatrix} \ldots$$

$$- h_2'(1,1)\begin{pmatrix} w(0)x^2(-1) \\ w(1)x^2(0) \\ \vdots \\ w(K)x^2(K-1) \end{pmatrix} - h_2'(1,2)\begin{pmatrix} w(0)x(-1)x(-2) \\ w(1)x(0)x(-1) \\ \vdots \\ w(K)x(K-1)x(K-2) \end{pmatrix} \ldots$$

Equation (6) represents eq. (2) by a vector based description. From the equations (5) and (6) it can be seen that the optimal predictor $\hat{\boldsymbol{u}}$ is an expansion of $\boldsymbol{u}$ by the vectors $\boldsymbol{u}_i$ and $\boldsymbol{u}_{i,k}$. Therefore it is convenient to introduce new designations which are defined by:

$$\begin{aligned} &\boldsymbol{u}'_\lambda = \boldsymbol{u}_i \quad \text{for} \quad \lambda \le N \quad \text{and} \\ &\boldsymbol{u}'_\lambda = \boldsymbol{u}_{i,k} \\ &\quad \text{with} \quad \lambda = N + i(i-1)/2 + k \quad \text{for} \quad \lambda > N\ . \end{aligned} \qquad (7)$$

In this manner the coefficients $h_1(i)$ and $h_2'(i,k)$ are mapped also to the coefficients a_λ so that eq. (6) changes into

$$\boldsymbol{e}_w^{N,M} = \boldsymbol{u} - \sum_{\lambda=1}^{W} a_\lambda \cdot \boldsymbol{u}'_\lambda \quad \text{with} \quad W = N + M(M+1)/2\ .$$

For the prediction the vector $\boldsymbol{u}$ may be approximated as good as possible by linear combination of the vectors $\boldsymbol{u}'_\lambda$

$$\boldsymbol{u} = \sum_{\lambda=1}^{W} a_\lambda \boldsymbol{u}'_\lambda + \boldsymbol{e}_w^{N,M} \qquad (8)$$

for which the Euclidean norm $|\boldsymbol{e}_w^{N,M}|$ should be minimal.

2.1.2 Orthogonal Basis Transformation

The vectors $\boldsymbol{u}'_\lambda$ represent a basis for a vector space assuming that the vectors $\boldsymbol{u}'_\lambda$ are independent among each other. If a vector $\boldsymbol{u}'_\omega$ depends on the others the vector $\boldsymbol{u}'_\omega$ can be omitted from the procedure and the corresponding coefficient of the expansion can be set to zero. The optimal solution is the linear combination $\sum_\lambda a_\lambda \boldsymbol{u}'_\lambda$ representing the parallel part of $\boldsymbol{u}$ to the space with the basis $\{\boldsymbol{u}'_\lambda\}$ while the error vector represents the orthogonal part. The vectors $\boldsymbol{u}'_\lambda$ are not necessarily orthogonal among each other so that the basis of $\boldsymbol{u}'_\lambda$ is transformed into an orthogonal basis $\{\boldsymbol{v}_\lambda\}$ with the dot product $\langle \boldsymbol{v}_i, \boldsymbol{v}_k\rangle = 0$. This is performed by the Gram-Schmidt orthogonalization

$$\boldsymbol{v}_\lambda = \boldsymbol{u}'_\lambda - \sum_{i=1}^{\lambda-1} \frac{\langle \boldsymbol{u}'_i, \boldsymbol{v}_i\rangle}{|\boldsymbol{v}_i|^2}\cdot \boldsymbol{v}_i \quad \text{for} \quad \lambda = 1\ldots W\ .$$

2.1.3 Determination of the Optimal Coefficients

The vectors of the basis $\{\boldsymbol{v}_\lambda\}$ are orthogonal. Hence the optimal coefficients b_i in description of the basis $\{\boldsymbol{v}_\lambda\}$ can be easily obtained by

$$b_i = \frac{\langle \boldsymbol{u}, \boldsymbol{v}_i\rangle}{|\boldsymbol{v}_i|^2}\ , \qquad (9)$$

yielding an expansion of the vector $\boldsymbol{u}$ by the vectors $\boldsymbol{v}_i$.

2.1.4 Inverse Basis Transformation

Since the original basis is $\{\boldsymbol{u}'_\lambda\}$ the coefficients b_i of the basis $\{\boldsymbol{v}_i\}$ have to be transformed into coefficients a_i of the basis $\{\boldsymbol{u}'_\lambda\}$. This transformation can be performed by the matrix $\boldsymbol{\Phi}$ with $\boldsymbol{a} = \boldsymbol{\Phi}\cdot\boldsymbol{b}$. The vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ contain the coefficients b_i and a_i. The matrix $\boldsymbol{\Phi}$ is defined by the vectors $\boldsymbol{\varphi}_i$:

$$\boldsymbol{\Phi} = \left(\boldsymbol{\varphi}_1, \boldsymbol{\varphi}_2, \ldots, \boldsymbol{\varphi}_W\right)^{\mathrm{T}}\ .$$

The matrix $\boldsymbol{\Phi}$ can be determined recursively starting with the identity matrix $\boldsymbol{\Phi} = \boldsymbol{I}$ with $\boldsymbol{I}(i,k) = \delta(i,k)$. The next steps are performed by:

$$\boldsymbol{\varphi}_i := \boldsymbol{\varphi}_i - \sum_{k=1}^{i-1} \frac{\langle \boldsymbol{u}_i, \boldsymbol{v}_k\rangle}{|\boldsymbol{v}_k|^2}\cdot \boldsymbol{\varphi}_k \quad \text{for} \quad i = 1\ldots W\ .$$

The final step is to map the coefficients a_i to the coefficients $h_1(i)$ and $h_2'(i,k)$ which is the inverse of the mapping (7).

2.1.5 Analysis of Signal Segments

Since in eqs. (2), (6) signal values outside of the frame appear, represented by negative arguments *n* of $x(n)$, the vector lengths *L'*=*K* +1 are truncated in such a way that only values inside of the analyzed segment appear in the vectors with $K = L - \max(N,M) - 1$; *L* is the segment length.

2.2 Analysis of Test Signals

Test signals are analyzed to evaluate the integration of the window function. For that purpose a test signal is generated with the inverse system

$$y(n) = x(n) + \sum_{k=1}^{N} h_1(k) \cdot y(n-k) + \sum_{i=1}^{M}\sum_{k=1}^{M} h_2(i,k) \cdot y(n-k)y(n-i) \qquad (10)$$

of the prediction error system. The inverse system (10) has a purely recursive structure, whereas the prediction error system (2) is nonrecursive. In the following a second order recursive Volterra system is used with the coefficients:

$$h_1(i) = \begin{pmatrix} 1 \\ 0.2 \\ -0.4 \end{pmatrix}, \; h_2(i,k) = \begin{pmatrix} 0 & 0.1 & -0.15 \\ 0.1 & 0.4 & 0.2 \\ -0.15 & 0.2 & 0 \end{pmatrix}. \qquad (11)$$

The excitation of the recursive system is an impulse train; the pulse period is 100 samples. The output *y* of the recursive system is analyzed by the nonlinear prediction with a window function w_s of a squared Hann window

$$w_s(k) = \left(0.5\left(1 - \cos(2\pi k / K)\right)\right)^2 \qquad k = 0 \ldots K \,.$$

Besides of the window function w_s, an asymmetric window function w_a is defined:

$$w_a(k) = 1 \qquad k = 0 \ldots K/4 - 1$$

$$w_a(k) = \left(0.5\left(1 + \cos\left(\tfrac{\pi(k - K/4)}{K - K/4}\right)\right)\right)^2 \qquad k = \tfrac{K}{4} \ldots K.$$

This window is sensitive towards changes on its left side.

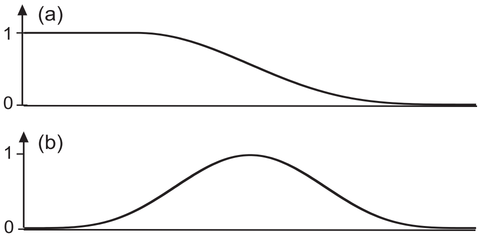

Figure 1 – Window functions: (a) asymmetric window w_a, (b) symmetric window w_s

The integration of the window function is realized in two different cases denoted by I and II. The order of the predictor is the same as the order of the recursive Volterra system.

2.2.1 Prediction with Windowed Signal

In case I the window function w_s is applied directly to the analyzed segment $y' = y(p, p+1, \ldots p+L-1)$ resulting in the weighted signal $y'_w(k) = w_s(k) \cdot y'(k)$. Then y'_w is analyzed by the nonlinear prediction with $w(k) = 1$ in eq. (2) since the windowing is already applied to the signal. By this procedure, the original coefficients cannot be determined correctly by the weighted segment y'_w. The estimated coefficients are for example

$$\hat{h}_1^{\mathrm{I}}(i) = \begin{pmatrix} 1 \\ 0.16 \\ -0.51 \end{pmatrix}, \; \hat{h}_2^{\mathrm{I}}(i,k) = \begin{pmatrix} 0.39 & -0.92 & 1.78 \\ -0.92 & 2.18 & -0.24 \\ 1.78 & -0.24 & -1.34 \end{pmatrix}$$

with start index *p* = 140 and segment length *L*=130. For an arbitrary start index *p* with segment length *L*=130 the averaged coefficient errors ε_1 and ε_2 between the original and the estimated coefficients of first and second order, respectively, are

$$\varepsilon_1 = \mathrm{mean}\left(\textstyle\sum_k (h_1(k) - \hat{h}_1(k))^2\right) = 0.05$$

$$\varepsilon_2 = \mathrm{mean}\left(\textstyle\sum_{i,k} (h_2(i,k) - \hat{h}_2(i,k))^2\right) = 16.2$$

2.2.2 Prediction with Windowed Error

In case II the nonlinear prediction with $w = w_s$ in eq. (2) is applied to the segment y'. In this way the estimated coefficients can be determined correctly in comparison to case I; in case II the averaged coefficient errors ε_1 and ε_2 are smaller than 10^{-8}. The estimated coefficients are correctly estimated for any segmentation values of *p* and *L* on condition that a pulse of the excitation is in the range of the analyzed segment. If *L* is greater than the period of the impulse train, this condition is always fulfilled.

The analysis of test signals generated by a noisy excitation shows also the advantage of the procedure of case II in comparison to case I. Overall the analyses of test signals show that the integration of the window function into the prediction by eq. (2) yields correct results, whereas applying the window function directly to the analyzed signal leads to incorrect estimation results; the deviation depends on the window function.

3. ANALYSIS OF SPEECH

3.1 Speech Features

Speech features can be defined by nonlinear prediction characterising the nonlinearity of the speech. The coefficients $h_1(i)$ represent linear components of the speech whereas the coefficients $h'_2(i,k)$ represent nonlinear components. $F^{N,M}$ is the logarithmic ratio of the weighted prediction error without and with nonlinear components:

$$F^{N,M} = \log\left(\frac{\left|\boldsymbol{e}_w^{N,0}\right|}{\left|\boldsymbol{e}_w^{N,M}\right|}\right).$$

The consideration of the nonlinear coefficients $h'_2(i,k)$ can be seen from the superscript *M*. Since the nonlinear coefficients contribute to a decrease of the prediction error, the feature $F^{N,M}$ has positive values and represents the prediction gain by the nonlinear components.

3.2 Analysis of Speech Signals

The speech signal is segmented into overlapping segments, which are analyzed individually. Applying the nonlinear prediction to each segment yields the corresponding predictor coefficients $h_1(i)$, $h'_2(i,k)$, and the speech feature $F^{N,M}$. To measure the features quasi continuously in time the displacement of the segments is chosen to one sample. Therefore the sequence of the estimated features of the segments represents a feature signal $F^{N,M}(n)$. The figures 2-4 show analyses of stationary speech whereas fig. 5 shows the analysis of a sound transition. The sampling rate of the analyzed speech signals is 16 kHz. For the interpretation of the features the analyzed speech signal and the corresponding LPC-residual signal are shown represented by curves (a) and (b). The LPC-residual is obtained from a standard linear prediction of order 30. One reason for the observed pulses in the LPC-residual is the glottal closure per period. The curves of the feature signals $F^{N,M}$ are shifted in a consistent way that the center of the symmetric windows and the left side of the asymmetric windows correspond to the values of the analyzed speech signal and LPC-residual. Figures 2 and 3 show the analyses of the vowel /a:/ and /e:/, respectively. The effect of the segment length can be seen from the curves (d)-(g) in fig. 2 and (d),(e) in fig. 3 representing the feature $F^{4,4}$ generated by the use of the symmetric window with different segment lengths *L*. The smaller the segment length, the finer is the time resolution. A comparison wit the LPC-residual shows that the features with small segment lengths can describe finer events like second pulses in the periods; this can be seen especially in fig. 3 where the second pulses are marked by arrows. The maxima of $F^{4,4}$ correspond to the main peaks of the LPC-residual caused by glottal closures and to the second peaks. The feature $F^{16,1}$ with the asymmetric window is more sensitive to the main pulses and less sensitive to the second pulses. In figure 3 it can be seen that the second impulses are missing in (c) in comparison to $F^{4,4}$ in (d). This behaviour is helpful to detect the pitch pulses if additional pulses occur in the LPC-residual. Figure 4 shows the analyses of consonants. It can be seen that the LPC-residual consists of many pulses per period whereas the feature signal $F^{16,1}$ show mostly one dominant pulse per period indicating the glottal closure. The analysis results show that events can be detected in a very fine time resolution by the use of the asymmetric window; the feature $F^{16,1}$ turned out to be effective. Figure 5 shows the analysis of a transition from the plosive /d/ to the vowel /a:/. The region of interest is here the start of the voiced excitation. The first glottal impulse can be observed from the first period of the speech signal by curve 5 (a); the start of the first period is marked by an arrow. The corresponding initial glottal closure can be seen by a peak of the feature signal $F^{16,1}$ with the asymmetric window function. Here the LPC-residual shows no indication for this event.

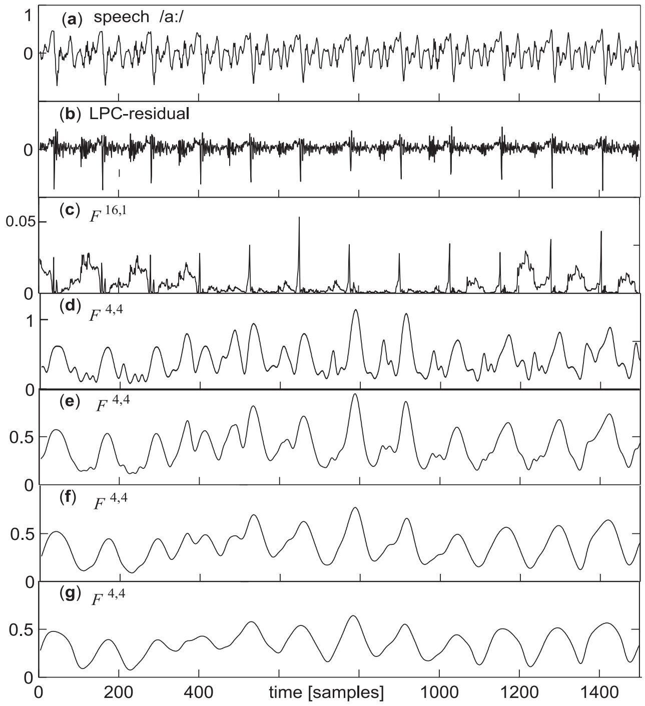

Figure 2 – Analysis of vowel /a:/: (a) analyzed speech signal, (b) corresponding residual by linear prediction. (c) Feature $F^{16,1}$ with segment length 180 and asymmetric window w_a, (d)-(g) Feature $F^{4,4}$ with the symmetric window w_s and different segment lengths: Length 100 for (d), length 115 for (e), length 180 for (f), and 200 for (g)

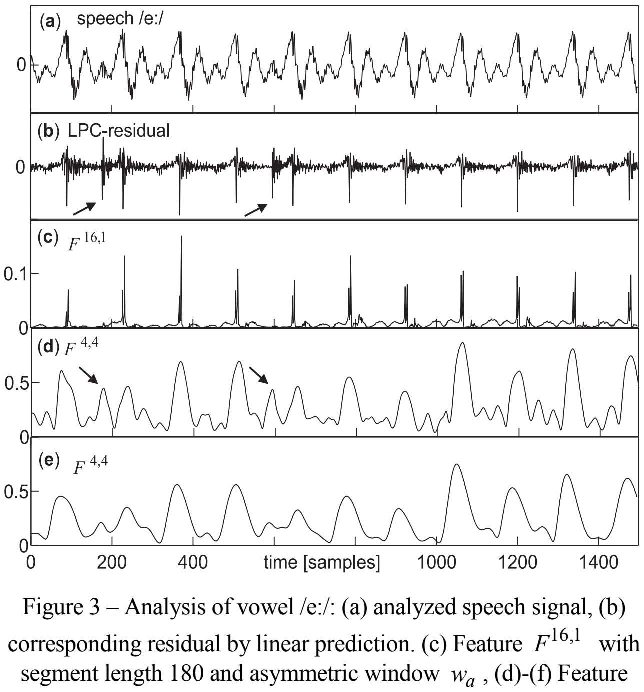

Figure 3 – Analysis of vowel /e:/: (a) analyzed speech signal, (b) corresponding residual by linear prediction. (c) Feature $F^{16,1}$ with segment length 180 and asymmetric window w_a, (d)-(f) Feature $F^{4,4}$ with the symmetric window w_s and different segment lengths: Length 115 for (d) and length 180 for (e)

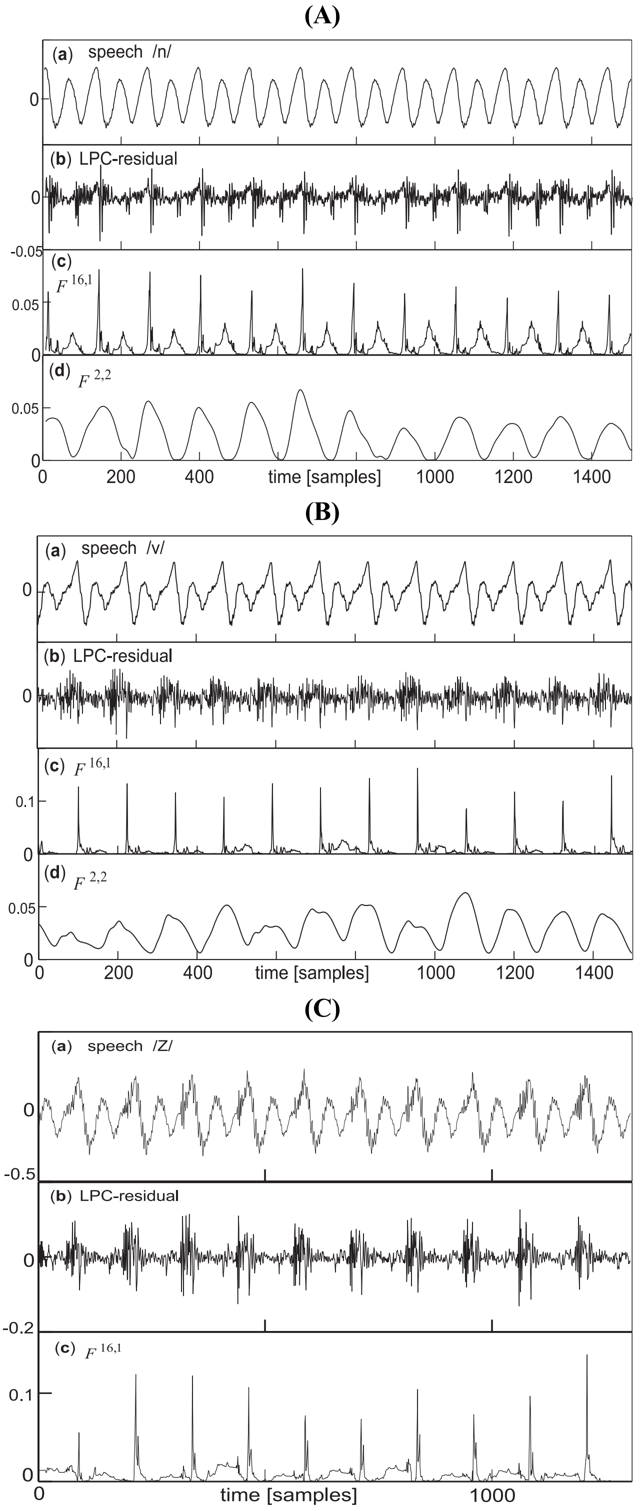

Figure 4 – Analysis of consonants: (A) nasal /n/, (B) voiced fricative /v/, (C) voiced fricative /Z/; (a) analyzed speech, (b) corresponding residual by linear prediction, (c) corresponding feature signal $F^{16,1}$ with asymmetric window w_a, (d) corresponding feature signal $F^{4,4}$ with symmetric window w_s

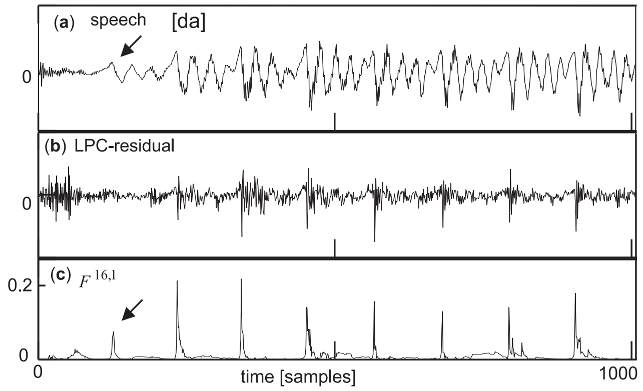

Figure 5 – Analysis of sound transition [da:]: (a) analyzed speech, (b) corresponding LPC-residual, (c) corresponding feature signal $F^{16,1}$ with asymmetric window w_a, arrows mark first glottal event

4. CONCLUSIONS

The nonlinear prediction of speech is performed by a vector expansion. For short-term analyses a window function is useful. Test signals show that applying the window function directly to the analyzed signal leads to incorrect results. To obtain correct results the window function should be applied to the prediction error in contrast to the first-mentioned procedure. Speech features are investigated representing the prediction gain improved by an inclusion of nonlinear components of a Volterra series. The impact of the type of the window function has been shown considering a symmetric and an asymmetric window. The usage of an asymmetric window function results in feature signals which consist of pulses correlating to the glottal closures. Examples demonstrate the informational content of nonlinear statistics of speech obtained by weighted nonlinear prediction. The results are relevant to the understanding of the speech production process and algorithms of feature analysis.

REFERENCES

[1] M. Faundez et al., "Nonlinear Speech Processing: Overview and Applications," in *Int. J. Control Intelligent Syst.*, vol. 30, no. 1, pp. 1–10, 2002.

[2] P. Maragos, T. Quatieri, and J. Kaiser, "Speech Nonlinearities, Modulations, and Energy Operators," in Proc. *ICASSP*, 1991, pp. 421-424.

[3] L. Atlas and J. Fang, "Quadratic Detectors for General Nonlinear Analysis of Speech," in *Proc. ICASSP*, 1992 vol. II, pp. 9–12.

[4] E. Mumolu, A. Carini, and D. Francescato, "ADPCM With Non Linear Predictors," in *Proc. EUSIPCO*, 1994, pp. 387–390.

[5] J. Thyssen, H. Nielsen, and S. D. Hansen, "NON-LINEAR SHORT-TERM PREDICTION IN SPEECH CODING," in *Proc. ICASSP*, 1994 vol. I, pp. 185–188.

[6] K. Schnell and A. Lacroix, "Modeling Fluctuations of Voiced Excitation for Speech Generation Based on Recursive Volterra Systems," contribution in *Nonlinear Analyses and Algorithms for Speech Processing,* LNAI Vol. 3817, pp. 338-347, Springer 2005.