THE NONLINEAR SIGNAL DECOMPOSITION IN VOICE RECOGNITION SYSTEM CONSTRUCTING

Alexander M. Krota and Polina P. Tkachova

* Institute of Engineering Cybernetics of the National Academy of Sciences of Belarus
6, Surganov Str., 220012, Minsk, Belarus
alxkrot@newman.basnet.minsk.by
Tel.: (375 17) 2842086, Fax.: (375 17) 2318403
bBelarusian State University,
4, Skoriny Av., 220050 Minsk, Belarus

Abstract. The nonlinear speech signal decomposition based on Volterra-Wiener functional series is described. The nonlinear filter bank structure is proposed for phonemes recognition solving.

Keywords: nonlinear signal decomposition, nonlinear filter bank structure, phoneme recognition.

Introduction
For speech recognition problems solving there are some paradigms and approaches. Among of them we can mention statistical approach based on hidden Markov models [1], [2], [3], [4], nonlinear dynamic method using neural networks [5], [6], algebraic approach [7], linguistic methods [8] and so on. However, in spite of some success in this problem solving there exist difficulties connecting with speech signal peculiarities. Speech signal by its nature has two aspects [4]. Firstly, speech of the person is defined by physical parameters, such as vocal tract length, glottal size and so on. Secondly, the speech producing is impossible without neural control of the articulations, which defines the personal learned abilities such as dialect or regional accents, pronunciation, speed and timing of the articulators. These two reasons find the reflection in speech signal nonlinearity, and it is necessary to implement nonlinear model.

This paper presents the nonlinear speech signal decomposition based on Volterra-Wiener functional series [9], [10]. It is shown the usage this nonlinear decomposition for nonlinear bank filters designing. It is proposed to solve the phoneme recognition problem by means of identification algorithm for these filters.

Linear and nonlinear decomposition of signal into series of functions and functionals
It is well known that music signal $y(t)$ in the time $t$ can be represented by means of Fourier series of the kind:

$$y(t)=\sum_{k=-\infty}^{\infty} Y(\omega_k) e^{\omega_k t}, \quad (1)$$

where $Y(\omega_k)$ are the coefficients of Fourier series as the following:

$$Y(\omega_k) = \frac{1}{T} \int_{0}^{T} y(t) e^{-\omega_k t} dt. \quad (2)$$

Here $T$ is interval of signal observation and $\omega_k = \frac{2\pi}{T} k$ is angular frequency. The relation (1) is described a linear decomposition of temporal function $y(t)$ into series from orthogonal Fourier functions $\exp(i\omega_k t) = \cos(\omega_k t) + i\sin(\omega_k t)$. On other hand, music signal as linear process, which is generated by linear dynamic system (LDS) exclusively has a full representation by means of linear series from orthogonal functions. Since a stationary LDS may be characterized by means of transfer function $H(\omega_k)$ the input signal $x(t)$ acting on LDS generates its output signal $y(t)$:

$$y(t) = \sum_{k=-\infty}^{\infty} H(\omega_k) X(\omega_k) e^{\omega_k t}. \quad (3)$$

As concerning speech signal, this signal is the product of strongly nonlinear dynamic systems (NDS), i.e. one is nonlinear process when its harmonic components have actions each other. In connection with this such signal can be represented by means of Volterra-Wiener series as the following [9], [10]:

$$y(t) = h_0 + \sum_{k=-\infty}^{\infty} H_1(\omega_k) X(\omega_k) e^{\omega_k t} + \sum_{k_1=-\infty}^{\infty} \sum_{k_2=-\infty}^{\infty} H_2(\omega_k\omega_{k_1}\omega_{k_2}) X(\omega_k) X(\omega_{k_1}) X(\omega_{k_2}) e^{\omega_k t + \omega_{k_1} t + \omega_{k_2} t} - \cdots$$

This work was supported partly by International Science and Technology Center (ISTC) under project B-95 and partly by University of Firenze, Italy

MAVEBA 1999, Firenze, Italy
\[ -D_x \sum_{k_1=-\infty}^{\infty} H_x(\omega_{k_1},-\omega_{k_1}) + \sum_{k_1=-\infty}^{\infty} \sum_{k_2=-\infty}^{\infty} \sum_{k_3=-\infty}^{\infty} H_x(\omega_{k_1},\omega_{k_2},\omega_{k_3}) \times X(\omega_{k_1})X(\omega_{k_2})X(\omega_{k_3}) \times e^{j(\omega_{k_1}+\omega_{k_2}+\omega_{k_3})x} - 3D_x \sum_{k_1=-\infty}^{\infty} \sum_{k_2=-\infty}^{\infty} H_x(\omega_{k_1},-\omega_{k_1},\omega_{k_2}) \times X(\omega_{k_1}) + \ldots \quad (4) \]

Comparing (3) with (4) we can see that mentioned above linear decomposition is particular case of this nonlinear decomposition similarly LDS is specific case of NDS. It follows from (4) the stationary NDS represents in the form of parallel connection of one-, two-, ..., multi-dimensional stationary LDSs with input signals \( x(t_1), x_2(t_1,t_2) = x(t_1) x(t_2), \ldots, x_m(t_1,\ldots,t_m) = -x(t_1) x(t_m) \) respectively.

The Volterra-Wiener functionals on finite intervals and their identification algorithm based on measuring the Wiener kernels

When the Wiener method [9], [10] applies in practice for the NDS analysis it is necessary to tell about physical white noise, that is about the process with the finite spectrum which overlaps the bandwidth of the system under study. Moreover, when NDS model based on Volterra-Wiener series is realized on a computer discrete input \( x_n \) and output \( y_n \) signals and Wiener kernels \( h_m[n_1,\ldots,n_m] \) will have a finite duration in time; that is why some refinement of the relation (4) is required [11].

Let us write discrete Volterra-Wiener functionals for the discrete signals \( x_n \) and \( h_0, h_1[n], h_2[n_1,n_2], \ldots \) given on the finite time intervals, that is \( n, n_1, n_2 = 0, 1, \ldots, N-1 \) [11], [12].

To represent one-dimensional sequences \( y_n \) and \( x_n \) in with the finite length \( N \) in frequency domain let us consider discrete Fourier transform (DFT) [13]:

\[ X_k = \sum_{n=0}^{N-1} x_n w_n^{kn} \quad w_N = \exp \left( -j \frac{2\pi}{N} \right) \quad k=0,\ldots,N-1 \quad (5) \]

and also its multidimensional analogues, i.e. the \( m \)-dimensional DFT’s (\( m=2,3,\ldots \)) of the kind:

\[ H_m[k_1,\ldots,k_m] = \sum_{n_1=0}^{N-1} \ldots \sum_{n_m=0}^{N-1} h_m[n_1,\ldots,n_m] w_n^{kn_1+n_2+n_m} \quad (6) \]

Taking into account (15), (16) and using inverse DFT and its multidimensional analogues let us transform relation (14) to the form [11], [12]:

\[ y_n = h_0 + \frac{I}{N} \sum_{k_1=0}^{N-1} H_x[k_1] X_k w_n^{-(k_1)} + \frac{I}{N^2} \sum_{k_1=0}^{N-1} \sum_{k_2=0}^{N-1} H_x[k_1,k_2] X_k X_l w_n^{-(k_1+k_2)} - \frac{3}{N^2} \sum_{k_1=0}^{N-1} \sum_{k_2=0}^{N-1} \sum_{k_3=0}^{N-1} H_x[k_1,k_2,k_3] \times X_k X_l X_m w_n^{-(k_1+k_2+k_3)} + \ldots \quad (7) \]

We consider index \( N-k_1 \) in (17) with respect to modulo \( N \) [12], i.e. \((N-k_1) \mod N = (N-k_1)\). If \( x_n \) is \( N \)-pointed sample of stationary Gaussian white noise then

\[ h_0 = M[y_n] = \frac{1}{N} \sum_{n=0}^{N-1} y_n = \frac{1}{N} \sum_{n=0}^{N-1} \frac{1}{N} \sum_{n=0}^{N-1} Y_k w_n^{kn} = \frac{1}{N} \sum_{k=0}^{N-1} Y_k N \delta_{k,0} = \frac{1}{N} Y_0. \]

The identification scheme of discrete NDS is similar to Wiener’s circuit for determining \( m \)-order kernel [9]. This scheme may be present as follows: white Gaussian noise with zero mean and variance \( D_x \) is given by the inputs of unknown NDS and bank of \( m \) complex exponential filters. Then the output signals from the system and bank are multiplied and the result signal is averaged [10].

Let us calculate DFT-image of kernel \( h_1[n] \). First obtain signal \( y_n^{(1)} \) from the output of known system that is filter with complex exponential impulse response \( w_N = \exp \left( -j \frac{2\pi}{N} \right) \):
\[\begin{align*}
M[Z^{(1)}_{n}] = & M[y_n y_{n}^{(1)F}] = h_0 M[X_{n}^*] w_{N}^{k_n} + \\
+ & \frac{1}{N} \sum_{k_1=0}^{N-1} H_I[k_1] M[X_{n_1}^* X_{n_1}^*] w_{N}^{n_1(k_1-k_1)} + \\
+ & \frac{1}{N^2} \sum_{k_1=0}^{N-1} \sum_{k_2=0}^{N-1} H_2[k_1, k_2] M[X_{n_1}^* X_{n_2}^*] w_{N}^{n_1(k_1-k_2)-n_2(k_2-k_2)} - \\
- & \frac{3D_2}{N^3} \sum_{k_1=0}^{N-1} H_1[k_1; N-k_1, k_1] M[X_{n_2}^* X_{n_2}^*] w_{N}^{n_2(k_2-k_2)} + \ldots. (9)
\end{align*}\]

It was shown in [10], [12] that if \( \{x_n\} \) is the sample of white Gaussian noise, then their coefficients \( X_k \) are \( N \)-pointed sample of Gaussian noise, that is the properties are true:

\[\begin{align*}
M[X_k] = & 0; \\
M[X_k X_{k_2}] = & D_X \delta_{k_k N-k_2}; \\
M[X_k X_{k_2} X_{k_3}] = & 0; (10) \\
M[X_{k_1} X_{k_2} X_{k_3} X_{k_4}] = & D_X^2 \left[ \delta_{k_1, N-k_2} \delta_{k_2, N-k_4} + \delta_{k_2, N-k_1} \delta_{k_4, N-k_3} \right] + \\
& + \delta_{k_1, N-k_2} \delta_{k_2, N-k_4} \delta_{k_4, N-k_3} \ldots,
\end{align*}\]

where \( D_X = \frac{1}{N} \sum_{k=0}^{N-1} X_k^2 \) - is the variance of partial population of coefficients \( X_k \) and \( D_X = N \delta_X \). Taking into account (10) we transform (9) to the form

\[\begin{align*}
M[y_n y_{n}^{(1)F}] = & \frac{D_X}{N} H_I[k] + \\
+ & \frac{D_X^2}{N^3} \left[ \sum_{k_1=0}^{N-1} H_I[k_1; N-k_1, k_1] + \sum_{k_2=0}^{N-1} H_I[k_2; N-k_2, k_2] + \\
& + \sum_{k_3=0}^{N-1} H_2[k, k_2; N-k_2, k_3] \right] - \\
+ & \frac{3D_2}{N^3} \sum_{k_1=0}^{N-1} H_1[k_1; N-k_1, k_1] M[X_{n_2}^* X_{n_2}^*] w_{N}^{n_2(k_2-k_2)} + \ldots. (11)
\end{align*}\]

Symmetry property of kernel \( H_3[k_1, k_2, k_3] \) was used for deduction (11). On the other hand, using (8) we can calculate the mean of partial population:

\[\begin{align*}
M[y_n y_{n}^{(1)F}] = & \frac{1}{N} \sum_{n=0}^{N-1} X_n^* w_{N}^{n k_n} + \frac{1}{N} \sum_{n=0}^{N-1} Y_{n} w_{N}^{n m a}. (12)
\end{align*}\]

Taking into account (11), (12) we have that [11], [12]

\[H_I[k] = \frac{y_k X_k^*}{N D_x} = \frac{Y_k X_k^*}{D_x} \tag{13} \]

According to relation (13) \( H_I[k] \) is the sample of transfer function \( H_I(\omega) \) for the stationary LDS, identified on the basis of stationary white noise \( \{x_n\} \).

The DFT-image of kernel \( h_2[n_1, n_2] \) may be calculated in an analogous manner. In this case signal \( y_n^{(2)F} \) is the product of outputs for two complex exponential filters, and either is described by the relation (6), that is \( y_n^{(2)F} = X_{n_1}^* X_{n_2}^* w_{n(k_1+k_2)} \). As a result we have [11], [12]:

\[\begin{align*}
H_2[k_1, k_2] = & \frac{Y_{k_1} X_{k_1}^* X_{k_2}^*}{2N D_x^2} - \frac{h_0}{N} \delta_{k_1, N-k_2}. (14)
\end{align*}\]

We can also show that DFT-image of kernel \( h_3[n_1, n_2, n_3] \) is [12]:

\[\begin{align*}
H_3[k_1, k_2, k_3] = & \frac{Y_{k_1} X_{k_1}^* X_{k_2}^* X_{k_3}^*}{6D_x^2} \\
& - \frac{N(H[k_1]) \delta_{k_1, N-k_1} + H[k_1] \delta_{k_1, N-k_1} + H[k_1] \delta_{k_1, N-k_1}}{6D_x}. (15)
\end{align*}\]

In this paper we use the identification scheme by Wiener kernel measuring for NDS testing by the white noise on finite interval [11], [12] and also the analogous identification scheme for other types testing signals (see, for example Ref. [14], [15]).

The synthesizer and recognizer of phonemes of Belarusian language based on nonlinear decomposition

The obtained nonlinear decomposition (4) may be used for identification of group of phonemes (say, sonorous phonemes of Belarusian language) by means of mentioned m-order multidimensional nonlinear filters.

But, at first let us give some preliminary information concerning a phonetic structure of Belarusian language [16].

For recognition the following classification is used. All the phonemes of belarusian language are divided into two groups: the first group has vocal (vowel) ones, the second group has consonant ones.
The vocal phonemes are again divided into labial once ( , ) and nonlabial once ( , , I{} ). (The sound is not considered as individual phoneme since in belarusian language it meet only after hard consonants and is modification of the phoneme I).

The consonants (due to difficulties in their recognition) are classified on the basis of the two approaches [16].

According to the first approach (account is taken into part in forming consonant phonemes of voice and noise, their formation place and method) the seven groups are separated [16]:

1. Noise, plosive, and media: , , , , *
2. Noise, plosive, and surd: , *, , *
3. Noise, fricative, and media: , *, *, *
4. Noise, plosive, and surd: , *, *, *, *
5. Noise, affricates, and media: Z, Z*
6. Noise, affricates, and surd: *, *
7. Sonorous: J {}, *

In accordance with the second (account is taken into pronunciation method) the two blocks involving the ten groups are extracted [16].

The labial (the block A):
1. Labial-labial, hard: , , { }.
2. Labial-labial, round: , *, *.
4. Labial-dental, round: *.
The lingua (the block B):
6. Front, dental, round: *, *, Z*, *, *
7. Front, alveolar, hard: , , , .
8. Middle, round: J{}.

According to the first approach we using the nonlinear filter structure consisting 7 nonlinear Volterra-Wiener filters. Each from them may be stimulated by a different testing signals (for example, by white noise, by colored noise, by tone, by tone plus noise etc.) depending on their position in the first scheme. Then each phoneme (there are from 3 to 10 ones for each corresponding Volterra-Wiener filter) may be recognized as one from orthogonal functionals based on mentioned identification scheme (see sect. 3 this paper).

The identification schemes are built for each type of testing signals by analogy with (8)-(15). As the final result, each phoneme corresponds to oneself functional, i.e. certain Wiener kernel and concrete type of testing signal. But first approach has essential lack because seven (sonour) group consists 12 phonemes that it makes difficult to use the identification scheme in practice.

That is why we apply the second approach for more reliable phonemes recognition. According to second approach we have a recognizer (synthesizer) in the form of the nonlinear bank filters consisting 10 Volterra-Wiener filters which may include from 1 to 6 nonlinear multidimensional (m-order) filters (or functionals). Thus, by increasing the number of channels number in the nonlinear filter structure (or the of testing signals) we increase a probability of phoneme recognition, generally speaking.

It is important to mention that both approaches to phoneme recognition do not exclude each other and are using jointly in phoneme recognition problem solving.

References


