Representation of a Voiced Speech Signal Model With Wigner-Ville Transform

Khawla Zammel¹, Noureddine Ellouze²

Signal, Image and Pattern recognition Lab
National School of Engineers of Tunis ENIT

Abstract—In this paper, we give a mathematical formulation of the Wigner Ville Transform of a voiced speech stationary model. The speech model is expressed as a sum of sinusoids with the harmonic frequencies. Identification of frequency coefficients from a stationary signal of real speech is operated. Then WVT term is compared with the transform of a real WVT voiced speech signal. Finally, identification of frequency coefficients from a stationary signal of real speech is operated.

Keywords—Energy distribution, Fourier Transform, Time-frequency analysis, Voiced speech model, Wigner Ville Transform.

I. INTRODUCTION

A signal is an information carrier, this information is time varying frequency, and analysing this signal need a time-frequency representation to show the behaviour of the energy in the time frequency plan [1].

Fourier transform and its squared modulus called spectrogram are straightforward tool for time frequency analysis but are parts of various methods of signal processing [7]. These distributions have limited resolution due to the correlation of the signal with time frequency atoms, and to overcome this limitation we need to define a distribution without waste of resolution.

The Wigner Ville transform (WVT) is a time frequency energy distribution that correlate the signal with itself which is a good way to outdo the problem of resolution but it shows another problem which is the interferences due to its bilinear nature. This distribution was first introduced by Wigner in 1932 in the quantum thermodynamics field and then used by Ville in 1948 in signal processing for improving the time-frequency distribution [12].

In this paper, we will work first on the parameters extraction and synthetise of the voiced speech signal using the sinusoidal model.

We will then express the WVT of the voiced speech model and compare it with the WVT of a real voiced speech signal, across which we will numerate a large number of WVT properties.

II. TIME-FREQUENCY ANALYSIS OF A VOICED MODEL SPEECH SIGNAL

Speech signal modelling is a mathematical representation that highlights some important aspect of the signal. This model was proposed by Portnoff [16] and was included in many other works as McAulay and Quatieri [15]. The idea begins with Fourier when he says that any periodic signal is a sum of a limited number of sinusoids k. from the short time Fourier transform only peaks are considered.

\[ x(t) = \sum_k \alpha_k e^{2\pi if_k t + \phi_k} \] (1)

The model parameters are, respectively, the amplitudes \( \alpha_k \), frequencies \( 2\pi f_k \), and the phase \( \phi_k \) [13].

This expression is one of many other sinusoidal modelling and it concerns only stationary signals, which mean that the model parameters are invariant in the considered frame [3].

The voiced signal is produced from a vibration of the vocal cords, in this case this model is well suited [2, 10].

After writing the model we need to extract the parameters from the real signal, for this we need to use the short time Fourier transform.

\[ F_s(t, f) = \int x(\tau)h(\tau - t)e^{-2\pi i\tau f} d\tau \] (2)

The signal used for parameters extracting and synthesis is a part of a vowel “a” of a woman, and the sampling frequency is 16,000 Hz.

The spectrum is computed using the STFT (2) with hamming window. The length of the window is equal to the signal.
The spectrogram, \( S_x(t,f) = \left| F_x(t,f) \right|^2 \), is represented using a hamming window with same length as the signal and 50% of overlapping.

Fig. 1. Signal: part of vowel "a".

From the spectrum we compute the amplitude of the local maxima from the spectrum \( \alpha_i \).

We also compute the fundamental frequency \( f_0 = 224 \text{Hz} \).

TABLE 1

<table>
<thead>
<tr>
<th>( \alpha ) Values</th>
<th>a1</th>
<th>a2</th>
<th>a3</th>
<th>a4</th>
<th>a5</th>
<th>a6</th>
<th>a7</th>
</tr>
</thead>
<tbody>
<tr>
<td>a1</td>
<td>1,4070</td>
<td>0,8153</td>
<td>1,8820</td>
<td>6,9970</td>
<td>14,6800</td>
<td>12,6300</td>
<td>7,3770</td>
</tr>
<tr>
<td>a8</td>
<td>1,5540</td>
<td>1,4200</td>
<td>0,9844</td>
<td>1,5580</td>
<td>1,8530</td>
<td>3,0120</td>
<td>0,9532</td>
</tr>
</tbody>
</table>

To synthetise the voiced speech signal we combine the previous parameters and (1) and then try to reduce the difference between the real and synthetised signal.

The idea is to find the factor which is the amplitude ration between the two signals.

The generated signal is:

Fig. 3. Reconstituted signal.

III. WIGNER VILLE TRANSFORM

A. Definition and Major Properties

The WVT \( W_x(t,f) \) of a time signal \( x(t) \) is expressed by [14]:

\[
W_x(t,f) = \int_{-\infty}^{\infty} x(t + \frac{\tau}{2})x^*(t - \frac{\tau}{2})e^{-2\pi i f \tau} d\tau. \tag{3}
\]

The WVT is a remarkable time-frequency transform as it gives an accurate energy distribution. This property is due to the correlation of the signal with itself using a translation in time and frequency and not using a window like in Fourier transform [8,12] because the windowing function has a blurry effect on the distribution [19].

Despite all properties that make WVT a powerful tool [11], the application of this transform is limited by the existence of interferences (from the quadratic aspect of the transform).

B. WVT of the Speech Signal Model

In this part, we will use the WVT with a voiced speech signal model to get the self and cross-terms expressions.

WVT of a speech signal can be written, based on (3) and the signal \( x(t) = \sum_k \alpha_k e^{2\pi i f_k t} \):

\[
W_x(t,f) = \sum_p \sum_q \alpha_p \alpha_q e^{2\pi i (p-q) f_0} \delta(f - f_0 \frac{p+q}{2}). \tag{4}
\]

Equation (4) can be written as:

\[
W(t,f) = \sum_k W_{kk}(t,f) + \sum_{k \neq p} W_{kp}(t,f).
\]
When \( p = q \) \( W_s(t, f) = \sum_k \alpha_k^2 \delta(f - kf_0) \). \hspace{1cm} (5)

When \( p > q \) as \( W_y^* (t, f) = W_p(t, f) \) we write
\[
W(t, f) = \sum_k W_{kk}(t, f) + \sum_{p>q} 2\Re W_{kp}(t, f).
\]

\[
W_s(t, f) = \sum_{p>q} 2\alpha_p \alpha_q \cos(2\pi(p-q)f_0t) \delta(f - f_0 \frac{(p+q)}{2}). \hspace{1cm} (6)
\]

Where \( W_{kk}(t, f) \) are self-terms energy distribution of \( x_k(t) \) and \( W_{kp}(t, f) \) are cross-terms of WVT.

When two or more sinusoids from the signal interact, cross terms are born. Their existence leads to interpretation problems. But from the previous equation not only their presence but the fact that they overlap the self-terms.

Same work, using the expression with the phase:
\[
x(t) = \sum_k \alpha_k e^{2\pi j f_k t + i\phi_k}.
\]

\[
W_s(t, f) = \sum_p \sum_q \alpha_p \alpha_q e^{2\pi i(p-q)f_0t} \delta(f - f_0 \frac{(p+q)}{2}).
\]

When \( p = q \) \( W_s(t, f) = \sum_k \alpha_k^2 \delta(f - kf_0) \).

When \( p > q \)
\[
W_s(t, f) = \sum_{p>q} 2\alpha_p \alpha_q \cos(2\pi(p-q)f_0t) f_0 t
\]
\[
+ (\phi_p - \phi_q) \delta(f - f_0 \frac{(p+q)}{2}).
\]

We will combine the values of the parameters with the formula found to represent the WVT of the voiced speech model.

The formula used are (5) and (6). The WVT representation is:

\[
\text{Fig.4. Wigner Ville Transform of the voiced speech model.}
\]

The Wigner Ville Transform of the real signal:

\[
\text{Fig.5. Wigner Ville Transform of the real signal.}
\]

C. Analysis

From the previous representation we can see that:

- The self-terms in the equation (5) are located in \( kf_0 \);
- The cross-term in the equation (6) are located in \( f_0 \frac{(p+q)}{2} \);
- Some of the cross-terms overlap the self-terms, which leads to a problem of energy analysis;
- Amplitude variation is due to the product \( \alpha_p * \alpha_q \);
- Presence of periodicity variation and negative values due to the cosine in (6);
- Energy distribution is accurate whereas it is blurred in the spectrogram.

The difference between the two representation is due to the real aspect of the signal which is non stationary. To correct this problem, we need to adopt a model with correction.
IV. CONCLUSION

In this paper, we discussed a comparison between two representation using the WVT, the first one represents the energy distribution of a sinusoidal model of a voiced speech signal (we extracted the parameters of a voiced frame of speech and used it to represent the WVT of the model) and the second is the WVT a real signal.

Through this comparison we studied the WVT from a theoretical point of view, by writing this formula we were able to separate the self-terms and cross terms, making the WVT easier to read. Other aspects of this transform were mentioned like the energy accuracy, amplitude and periodicity.

Perspective

Several studies worked on reducing the resulting interferences after the WVT by the use of a smoothing function [9, 5, 4, 17].

In future works we will develop the smoothing process using the WVT of the window as a smoothing function which leads to the spectrogram.

The spectrogram of the signal $x(t)$ can be estimated by computing the squared magnitude of $S(t, f)$:

$$S(t, f) = |S(t, f)|^2 \quad \text{But also} [6]$$

$$S(t, f) = \int \int W_x(\tau, v)W_y(\tau - t, v - f)d\tau df.$$  

We will focus also on the WVT of non-stationary signals as a generalization of this work [18, 7].

REFERENCES