INTERSPEECH 2009 BRIGHTON

A Novel Method for Epoch Extraction from Speech Signals

Lakshmish Kaushik [1], *Douglas O'Shaughnessy*[1]

[1] Speech Communication group, INRS, Montreal, Canada

Kaushik@emt.inrs.ca, dougo@emt.inrs.ca

Abstract

This paper introduces a novel method of speech epoch extraction using a modified Wigner-Ville distribution. The Wigner-Ville distribution is an efficient speech representation tool with which minute speech variations can be tracked precisely. In this paper, epoch detection/extraction using accurate energy tracking, noise robustness, and the efficient speech representation properties of a modified discrete Wigner-Ville distribution is explored. The developed technique is tested using the Arctic database and its epoch information from an electro-glottograph as reference epochs. The developed algorithm is compared with the available state of the art methods in various noise conditions (babble, white, and vehicle) and different levels of degradation. The proposed method outperforms the existing methods in the literature.

Index Terms: Epoch extraction, Wigner-Ville distribution, Glottal closure instance, Electro-Glottograph.

10.21437/Interspeech.2009-45

1. Introduction

In various applications of speech, the epoch of voiced speech is a very useful speech parameter. Epoch is defined as an instant where there is significant excitation due to glottal vibrations in the vocal tract. The vocal folds vibrate during the production of voiced sounds and speech is characterized by a substantial instantaneous increase in signal energy due to the closure of the vocal folds at the end of each glottal cycle. This instantaneous change can be represented as an impulse-like excitation [4]. Glottal pulse inverse filtering [1, 5] to find the glottal impulse locations is a major step in speech analysis.

For performing speech analysis, the frequency response of the vocal tract and glottal pulse, representing the excitation source, has to be accurately found. Using epoch information, the fundamental frequency (f_0) can be obtained. Utilizing epoch information we can perform pitch variation, prosody modification, time-scale modification, pitch mark extraction for speech applications like speech recognition, text-to-speech synthesis, voice conversion, etc. It can also be used in speaker tracking and speaker verification for obtaining speaker specific information. Speech coding efficiency can also be improved using epoch information.

Presently, there are various algorithms which perform epoch extraction. Since 1975 many notable algorithms have been developed by Yegnanarayana et al, using glottal wave-shapes [10], LPC residual, group delay function, Hilbert envelope [7] and resonance filters [9]. Another well known approach is DYPSA [6] by Naylor et al, using energy, phase slope and zero crossings of speech. Though there are various methods to extract epochs, they are either not very accurate in finding epoch positions or perform poorly in the presence of noise. Hence there is a requirement for an algorithm that is robust to noise and also obtains accurate epoch positions.

This paper presents a new algorithm using a modified Wigner-Ville distribution (WVD) to extract epochs from speech signals. WVD is an elegant speech representation that can track spectral and energy changes over time efficiently. The Discrete-Time WVD (DWVD) [3] and its modified version are used in the present work. DWVD is an efficient energy tracker, is robust to noise and can represent speech of shorter duration with a high resolution time-frequency representation (TFR) in comparison to other TFRs like the spectrogram.

The outline of the paper is as follows. Section 2 introduces the WVD and its variants. The relevance of WVD for epoch extraction is explained. Section 3 explains the epoch extraction algorithm and implementation in detail. Database specifications, comparison and analysis of results with other techniques for different noise conditions are presented in section 4 and Section 5 is a conclusion.

2. Wigner-Ville Distribution

WVD was developed in the context of quantum mechanics. The wave function (a probabilistic amplitude function that maps possible states in a particular space of a system to complex numbers) in Schrodinger's equation is very difficult to calculate as stated by Heisenberg's uncertainty principle (HUP). Hence, Eugene Wigner developed a quasi-probabilistic distribution in phase-space as a substitute for the wave-function, which is not affected by HUP [2]. Later in 1980 Claasen and Mecklenbräuker published a series of articles depicting elegant TFR properties of WVD [3]. An important property of WVD is that it can generate high resolution TFRs of non-stationary signals. Interestingly in our case, speech is either quasi-periodic or non-stationary, which is added to a non-stationary noise signal giving rise to a hybrid signal. Hence a tool like WVD that can efficiently exploit the above characteristic is advantageous. The WVD of a real signal $s(t)$ is given by [2]:

$$WVD_s(x;t,\omega) = \int_{-\infty}^{\infty} e^{-j\omega\tau}\, x\left(t+\frac{\tau}{2}\right) x^*\left(t-\frac{\tau}{2}\right) d\tau \quad (1)$$

where $x(t)$ is the analytic signal associated with the real signal $s(t)$. On observation, equation (1) is a Fourier transform (FT)-like operation. The equation (1) is equivalent to taking the Fourier transform of the analytic representation obtained from the Hilbert transform of the input signal $s(t)$.

2.1. Discrete Wigner-Ville Distribution

In windowed digital speech processing a discrete time equivalent of equation (1) is used. A $2N$ point DWVD equation is reduced to an N point formulation, which is given by [2],

$$W_s(k) = 2\sum_{n=0}^{N-1} \alpha(n) W_N^{nk} \quad (2)$$

where $\alpha(n) = K(n) + K(N+n)$, which looks like an N-point DFT. $K(n)$ is given by

$$K(n) = \begin{cases} x(n) & 0 \le n \le N-1 \\ x(n-2N) & N+1 \le n \le 2N-1 \\ 0 & Otherwise \end{cases} \quad (3)$$

3. Speech Epoch Extraction

Epoch is applicable only for voiced speech content. The vocal chords vibrate only during the production of voiced speech. Hence epoch detection is not done on unvoiced speech units. Epoch extraction (EE) from speech can be carried out in the time domain, spectral domain or other parametric domains. The presented technique uses TF coefficients obtained by a modified DWVD. The following sections consist of a detailed explanation of the algorithm and the properties of DWVD that are beneficial for EE. The algorithm consists of 3 steps namely, (a) Windowing, (b) DWVD Generation and (c) Epoch Extraction.

3.1. Windowing

The entire experimentation is carried out using CMU's Arctic speech database [8]. The wave files are sampled at 32 kHz. Data is windowed using a 5 msec window (N) with 50% overlap. This means, in every frame, 2.5 msec of new information is obtained. But for efficient epoch extraction 5 msec of each data frame is again divided into frames of 2.5 msec with 50% overlap each. All processing is performed on the 2.5 msec frame. It is also experimentally advantageous because the DWVD is very sensitive to energy changes, and the smaller the window length the better energy tracking capability, which results in an increased accuracy.

Selection of window length is independent of sampling frequency. By rigorous experimentation it is found that a 2.5 msec speech frame with 50% overlap is appropriate for the best epoch extraction results. The possible human pitch range varies from as low as 60 Hz (16 msec) in males to 500 Hz (2 msec) in children. If frame length is equal to the minimum pitch possible then the accuracy increases. Irrespective of the sampling frequency the window length is selected to be 2.5 msec with 50% overlap. The number of samples in a frame is parametric, hence can be scaled depending on the sampling frequency.

3.2. DWVD Coefficients

The DWVD representation of speech is advantageous due to many reasons. DWVD is the only distribution that can satisfy all six properties for an ideal Joint-TFR [3]. Hence it allows a high degree of flexibility and accuracy in representing signal variation with high resolutions without any compromise in signal representation, even if the signal is a continuous aperiodic non-stationary signal. DWVD and its variants are also robust to noise [11]. DWVD can track energy variations with high resolution even using short range data.

Figure 1 shows a maximum resolution DWVD for a speech frame. The window length and band resolutions are fixed in a manner which suits EE. To maximize accurate EE the time and band resolutions should be maximum. Hence the band resolution is selected as one sample and time resolution is the window length, which results in the maximum resolution. The fig shows that the speech coefficients are concentrated in the first few columns of coefficients irrespective of the level of SNR [11]. Hence DWVD is advantageous in noisy conditions.

Consider a voiced speech segment and its corresponding epochs generated by the differenced Electro-Glotto-Graph (EGG) signal as shown in figure 2. The epoch locations are in the negative parts (or lower part) of the speech signal $s(n)$, assuming the signal has no bias as given in equation (4). For calculating the epoch locations, speech values below zero are considered.

$$g(n) = s(n) < 0 \qquad (4)$$

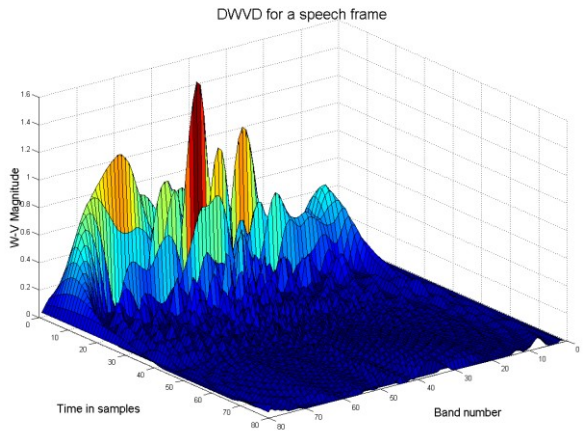

Figure 1: *DWVD coefficients of a speech frame in maximum resolution (band number = time in samples = window length)*

The generated DWVD of a data frame has both positive and negative values. For the purpose of EE only positive values are considered, as negative values do not contribute meaningfully because the negative values are obtained from the imaginary part of the spectrum due to conjugate symmetry. When the noise content increases it gets distributed over the entire range. It can be observed that the speech region is concentrated in a small range as seen in Figure (1). Just adding noise does not affect the distribution greatly.

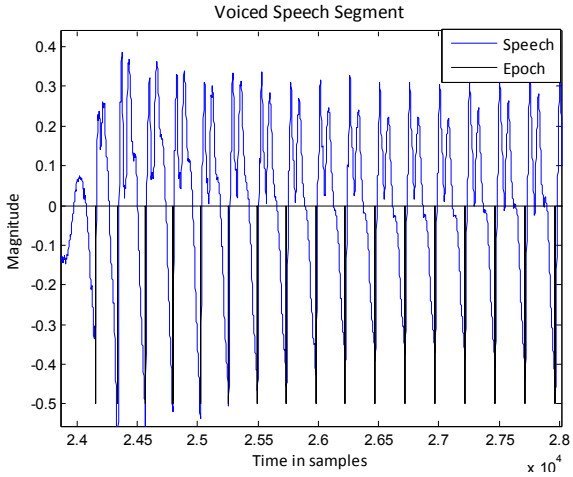

Figure 2: *Epoch locations in a voiced speech segment*

3.3. Algorithm

Epoch Extraction adopts a top-down approach. Here epochs are determined by localizing the search region for the epoch. The steps of the algorithm are as follows.

3.3.1. DWVD coefficients generation

Given a speech frame of 2.5 msec, maximum resolution DWVD coefficients are obtained. For efficient epoch extraction perform Discrete Wigner-Ville transform for absolute of the negative speech values ($|g(n)|$) as explained in section 3.2. Considering only negative values will reduce computational complexity and increase the accuracy of epoch detection. DWVD coefficients of dimension NxN for each speech frame are obtained by

$$DWVD_{frame}\ (i) = \sum_{i=1}^{N}\sum_{j=1}^{N}\left|DWV_{transform}\ (|g(n)|)\right| \qquad (5)$$

3.3.2. DWVD summation and Epoch extraction

Given the DWVD coefficients for a speech frame, perform the absolute sum of all the coefficients as given by equation (5). Figure 3 shows a voiced speech segment and the corresponding DWVD curve variation over time. From figure 3, epochs can be localized as follows,

a) Search for a local maximum in the DWVD curve in the voiced part of the speech. The window length in the present case is 80 samples with 40 samples overlap, as we have considered a 2.5 msec window (it is parametric, and can be scaled depending on the sampling frequency). The presence of a local maximum in the DWVD curve of Figure 3d is an indication of the presence of an epoch in the vicinity of the data window in Figure 3a.
b) The next step is to localize the search for the epoch. Consider the index of the local maximum as I. The window length being 80 samples with 40 samples overlap there is a probability that the exact epoch is located in the vicinity of ± 40 samples of the local maximum of DWVD. Search for a local maximum in the modified speech data in Figure 3b over the index of $I \pm 40$ samples.
c) The index of the local maximum obtained during the search in modified speech is the epoch location.

Figure 3 shows detailed representations of different stages in the epoch extraction using DWVD. Analyzing Figure 3d, it can be seen that the DWVD curve is consistent with the variations in the speech signal. The results obtained are highly accurate and are also consistent in adverse noisy conditions. Another advantage of this technique lies in the fact that DWVD is inherently is robust to noise [11].

4. Evaluation of Extracted Speech Epochs

4.1. Database generation

For the purpose of this experiment CMU's Arctic Speech Database is used. This database has 1132 phonetically balanced speech sentences from three speakers (2 males and 1 female) which are recorded in a controlled environment and the sampling frequency is 32 kHz [8]. The Arctic database also contains a corresponding Electro-Glottograph (EGG) signal recorded using a laryngograph simultaneously with the speech recording. The larynx-to-microphone delay is determined to be 0.7 msec and a time alignment of the EGG signal is performed accordingly to compensate for the delay.

In this experiment we have considered the entire Arctic speech database of 3 speakers. The voiced regions were extracted from the speech sentences using boundary information of voiced speech segments from EGG. Totally there are 792,249 epochs in the voiced regions in the entire database.

4.2. Noise addition

For the database above, different noises are added with various levels of signal degradation. This was performed using the CMU's NOISEX-92 noise database. Three noises, namely babble, white and vehicle noise backgrounds, are considered. Using the above, the speech signals are degraded to various Signal to Noise Ratio (SNR) levels.

4.3. Parameters evaluated

Many parameters of epoch detection can be evaluated in different signal and noise scenarios using the CMU's Arctic Speech Database and NOISEX-92 database. In the presented work following parameters are evaluated.

a) *Identification rate (IDR)*: The percentage of larynx cycles for which exactly one epoch is detected.
b) *Miss rate (MR)*: The percentage of larynx cycles for which no epoch is detected.
c) *False alarm rate (FAR)*: The percentage of larynx cycles for which more than one epoch is detected.
d) *Identification accuracy (IDA)*: The standard deviation of the identification error (The time error between the detected epoch locations and the reference epoch). The smaller the value of IDA the higher is the accuracy of identification.

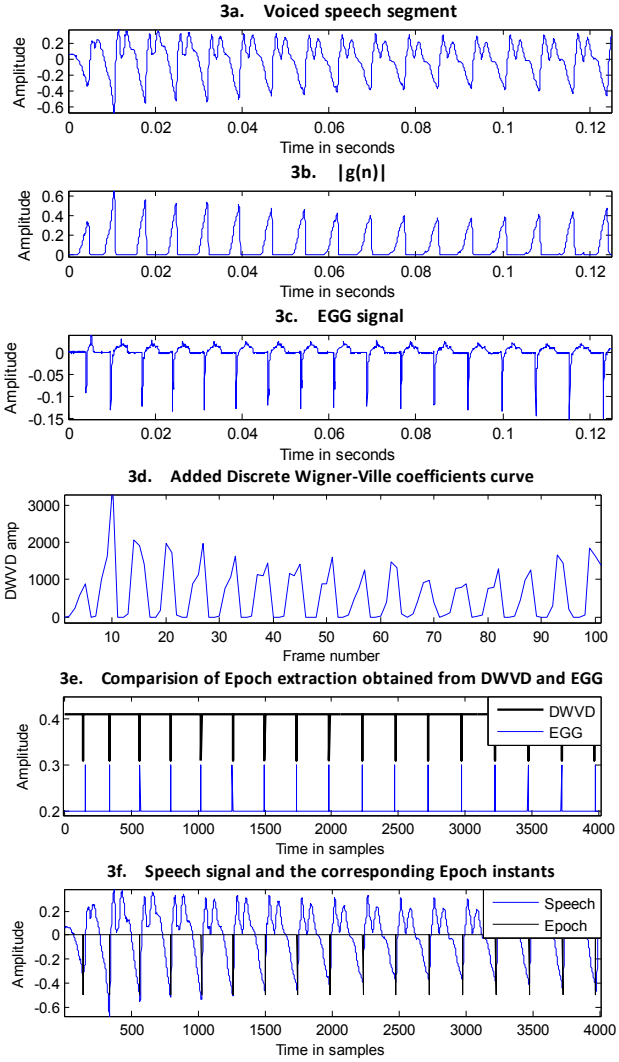

Figure 3: *Plot of different stages of algorithm and its result is compared with the EGG from the Arctic database. (a) Considered voiced speech segment. (b) Modified voiced speech segment. (c) Differenced EGG signal of the speech segment from the Arctic database. (d) Summed DWVD curve. (e) Comparison of epochs extracted from the algorithm and from the differenced EEG signal of* 3c. *(f) Speech segment along with its extracted epoch positions.*

4.4. Results and Tabulation

The DWVD-based epoch extraction algorithm is tested on all the sentences of the three speakers of CMU's Arctic Database [8]. The results are presented and analyzed in this section. As seen in Figure 3e, extracted epoch locations are highly accurate when it is compared with the differenced EGG signal from the

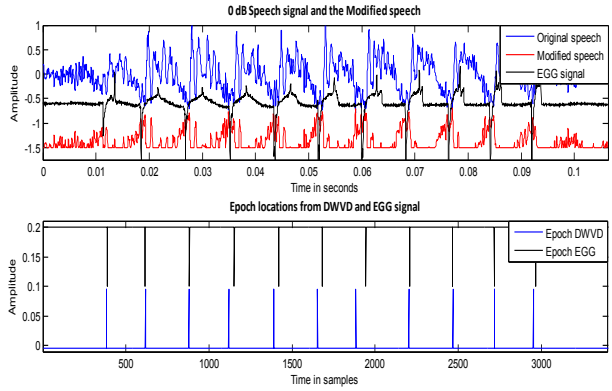

Figure 4: *Plot of extracted epochs from a voiced speech signal segment of 0 dB SNR affected by babble noise*.

Figures (3e & 3c) are consistent in various conditions tested. From Figure 3f it can be observed that the extracted epoch locations are in conformity with the required results. Figure 4 shows an example of epoch extractions in a highly noisy condition. Here the speech signal is affected by babble noise and the SNR of the signal is 0 dB. Even in such extremely noisy condition the algorithm is very robust and the accuracy of the results obtained is good. It is observed that even at low SNRs, as low as 0dB, the DWVD-based algorithm has not missed any epochs.

The DWVD based epoch extraction algorithm is compared with two other widely used methods namely,

a. Hilbert Envelope based method (HE-based) [7]
b. DYPSA Algorithm based method (DYPSA) [6]

Table 1: *Comparison of results of epoch detection using Hilbert envelope based method and DYPSA Algorithm DWVD based method in various noise conditions*

Environment		**HE-based**				**DYPSA**			
Noise	SNR (dB)	IDR (%)	MR (%)	FAR (%)	IDA	IDR (%)	MR (%)	FAR (%)	IDA msec
White	20	84.56	1.58	13.86	0.686	92.12	1.41	6.47	0.738
White	15	82.26	1.9	15.85	0.761	85.33	1.24	13.43	0.841
White	10	79.45	2.39	18.16	0.864	75.95	1.09	22.96	0.957
Babble	20	86.73	1.54	11.73	0.674	96.42	1.8	1.79	0.621
Babble	15	84.88	1.77	13.35	0.743	96.14	1.82	2.05	0.647
Babble	10	82.51	2.17	15.32	0.842	95.48	1.78	2.74	0.69
Vehicle	20	89.75	1.4	8.85	0.584	96.67	1.76	1.57	0.589
Vehicle	15	89.58	1.39	9.03	0.585	96.6	1.78	1.62	0.596
Vehicle	10	89.25	1.37	9.38	0.591	96.64	1.76	1.61	0.597

Environment		**DWVD-based**			
Noise	SNR(dB)	IDR (%)	MR (%)	FAR (%)	IDA(msec)
White	20	99.2	0.47	0.33	0.406
White	15	98.77	0.56	0.67	0.504
White	10	98.14	0.98	0.88	0.569
Babble	20	98.94	0.51	0.55	0.423
Babble	15	98.33	0.68	0.99	0.518
Babble	10	97.8	1.03	1.17	0.590
Vehicle	20	99.18	0.44	0.38	0.413
Vehicle	15	98.9	0.6	0.5	0.499
Vehicle	10	98.26	0.94	0.8	0.546

From the Table 1, we can observe that the DWVD-based algorithm is superior when compared to all the other algorithms. If we observe the values of IDR, the accuracy has increased by a good margin. For example, consider the case of a signal affected by 10 dB white noise; has increased by an average of more than 20%, which is remarkable. Consider IDA values for the same case. The average error in the deviation from the exact locations of epochs is just 9.14 samples, whereas in other cases it is, for the entire database, it is 14 samples. From analysis it is found that more than 95% of the miss rate is due to those laryngeal cycles that occur at the end of speech segments. Still this performance is better than any other technique even at the end of speech segments, because it can track energy variations very efficiently. Another important observation is that the average identification error is very much less in comparison to other methods, which means that the distance between the estimated epoch locations and reference epoch from the EGG signal is much less. The results of the algorithm for signals affected by lower SNR (as low as 0 dB) are not reported in the present paper. It has been tested and the accuracies are found to be very promising. From figure 4 it can be observed that the algorithm is very robust even at low SNRs.

5. Conclusion

In this paper a Discrete Wigner-Ville distribution-based speech epoch extraction algorithm is presented. The algorithm developed proves to be robust and efficient in a variety of noise conditions tested. The advantage of the algorithm lies in the fact that it is very simple yet can produce such accurate results. The algorithm developed is tested against the state of the art algorithms in the literature and the results are found to be very encouraging.

6. Bibliography

[1] D. Veeneman and S. Be Ment, "Automatic glottal inverse filtering from speech and electro-glottographic signals," *IEEE Trans. Signal Process.*, vol. SP-33, no. 4, pp. 369–377, Apr. 1985.

[2] E.P. Wigner, "On the quantum correction for thermodynamic equilibrium", Phys. Rev. 40 (June 1932) 749-759.

[3] T. A. C. M. Claasen and W. F. G. Mecklenbräuker, "The Wigner distribution a tool for time-frequency signal analysis. Part I, II & III", Phillips Journal of Research, vol 35, 1980.

[4] O'Shaughnessy, Douglas. Speech communication. IEEE press, 2000.

[5] H. W. Strube, "Determination of the instant of glottal closures from the speech wave," *J. Acoust. Soc. Amer.*, vol. 56, pp. 1625–1629, 1974.

[6] P. A. Naylor, A. Kounoudes, J. Gudnason, and M. Brookes, "Estimation of glottal closure instants in voiced speech using the DYPSA algorithm," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 15, no. 1, pp.34–43, Jan. 2007.

[7] K. S. Rao, S. R. M. Prasanna, and B. Yegnanarayana, "Determination of instants of significant excitation in speech using Hilbert envelope and group-delay function," *IEEE Signal Process. Lett.*, vol. 14, no. 10, pp. 762–765, Oct. 2007.

[8] J. Kominek and A. Black, "The CMU Arctic speech databases," in *5th ISCA Speech Synthesis Workshop*, Pittsburgh, PA, 2004, pp. 223–224.

[9] Sri Rama Murty K. and Yegnanarayana B., ". Epoch Extraction From Speech Signals", Nov 2008, *IEEE Trans. ASSP*, Vol. 16, pp. 1602-1613.

[10] T. V. Ananthapadmanabha and B. Yegnanarayana, "Epoch extraction of voiced speech," *IEEE Trans. Acoust., Speech, Signal Process.*, vol.ASSP-23, no. 6, pp. 562–570, Dec. 1975.

[11] Lakshmish Kaushik, Douglas O'Shaughnessy, "Voice activity detection using modified Wigner-Ville distribution", *Interspeech 2008*,Brisbane Australia, September 2008, pp. 2574-2578.