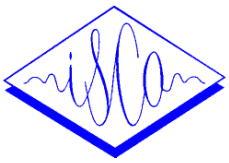

4th European Conference on
Speech Communication and Technology
EUROSPEECH '95
Madrid, Spain, September 18-21, 1995

AN IMPROVED EPOCH DETECTION ALGORITHM BASED ON SINUSOIDAL MODELLING OF SPEECH

M. Larreategui, F.J. Ancin and R.A. Carrasco
School of Engineering, Staffordshire University,
Beaconside, PO Box 333, Stafford STD18 ODF, U.K.
E-mail: mikel@bss10a.staffs.ac.uk

ABSTRACT

Reliable and automatic estimation of the Glottal Closure Instant (GCI), also referred to as epoch or pitch-event, has increasingly become an essential requirement in many speech processing applications, such as high-quality speech synthesis [1][2] and speech coding [3]. Although several methods exist [4][5][6], GCI detection still remains a difficult task. In this paper a robust pitch-event detection algorithm based on the sinusoidal modelling of speech is proposed. The improvement in performance over other well-known epoch detectors demonstrates the powerfulness of the sinusoidal technique for GCI determination.

1. INTRODUCTION

Sinusoidal analysis has already proven to be a suitable technique in many speech processing areas such as speech synthesis [2], modification [7], coding [3] and enhancement [8]. In this paper a robust epoch detector algorithm using the sinusoidal technique is proposed. This algorithm is based on the sinusoid-based phase model proposed by McAulay in [3]. The original aim of the algorithm in [3] was to improve the coding efficiency of the sine-wave phases and reduce the size of the parameter set for low-rate coding applications. However, it has also proven to be a suitable technique for pitch-event estimation. In our research work, appropriate modifications have been carried out in order to obtain an improved and more robust pitch-epoch detector.

The outline of the paper is as follows. In section 2, the mathematical description of the original sinusoid-based epoch detector (OSED) is given. Section 3 explains the appropriate modifications (e.g., the Hilbert envelope) to attenuate the problems existing in the OSED method. Then, in section 4 the results of the evaluation tests are described. Finally, in the last section some conclusions are given.

2. EPOCH DETECTION USING SINUSOIDAL TECHNIQUE: THE OSED ALGORITHM

On a short-time basis, the speech waveform $s(n)$ can be modelled in terms of the sinusoidal model as follows [3]

$$s(n) = \sum_{l=1}^{L} A_l \exp[j(n\omega_l + \theta_l)] \quad -N/2 \le n \le N/2 \qquad (1)$$

where the amplitude, frequency and phase of the sine waves (A_l, ω_l, θ_l) correspond to the peaks of the magnitude of the short-time Fourier transform and L is the number of sine waves. In order to have well-resolved peaks in the Fourier spectrum, the length N of the current analysis frame must be at least twice the local pitch period.

During steady voicing, the excitation waveform will consist of a sequence of pitch pulses separated by the local pitch period. Given an analysis frame of length $N+1$, let the pitch pulses or epochs be located at n_1, n_2, ... n_K, where $-N/2<n_i<N/2$ and K be the number of pitch pulses within the analysis frame. In the context of the sine wave model, a pitch pulse occurs when all the sine waves add coherently, that is, are in phase [3]. If a pitch pulse occurs at $n_o \in [n_1, ..., n_K]$, then the excitation signal can be modelled as

$$\hat{e}(n) = \sum_{l=1}^{L} a_l \exp[j(n-n_o)\omega_l + \beta\pi] \quad -N/2 \le n \le N/2 \quad (2)$$

where β is either 0 (for positive pulses) or 1 (for negative ones). Passing this excitation signal through the system function $H_s(\omega)$ (vocal tract), the speech signal at its output becomes

$$\hat{s}(n) = \sum_{l=1}^{L} a_l |H_s(\omega_l)| \exp[j(n\omega_l + \hat{\theta}_l)] \qquad (3a)$$

where $$\hat{\theta}_l = -n_o\omega_l + \arg H_s(\omega_l) + \beta\pi \qquad (3b)$$

For estimating the system function $H_s(\omega)$, either homomorphic processing or linear prediction analysis can be used assuming that the vocal tract is a minimum-phase system.

The value of epoch n_o must be such that $\hat{s}(n)$ is as close as possible to $s(n)$. One way to estimate n_o is to find the minimum of the squared error (MSE) over n_o between $\hat{s}(n)$ and $s(n)$,

$$\varepsilon(n_o) = \sum_{n=-N/2}^{N/2} |s(n) - \hat{s}(n; n_o)|^2 \quad -N/2 \le n_o \le N/2 \qquad (4)$$

Combining equations (1-4) leads to the following simplified error function[3]

$$\rho(n_o) = \sum_{l=1}^{L} A_l^2 \cos[\theta_l + n_o\omega_l - \Phi_s(\omega_l) + \beta\pi] \qquad (5)$$

where $\Phi_s(\omega)$ is the phase arg $H_S(\omega)$ of the system transfer function $H_s(\omega)$. Evaluating Equation (5) over $n_o \in [-N/2, N/2]$ leads to a set of maximum peak values n_i, each of which corresponds to one pitch pulse or epoch $n_i \in [n_1, ..., n_K]$.

The epoch detector described above presents two major problems: firstly, identification of epochs (maximum peak values) from the error function $\rho(n_o)$ could still be difficult due to the presence of several secondary peaks. An example of this problem is shown in Figure 1: the epoch pulses, marked by arrows, correspond to the maximum peaks in $\rho(n_o)$. However,

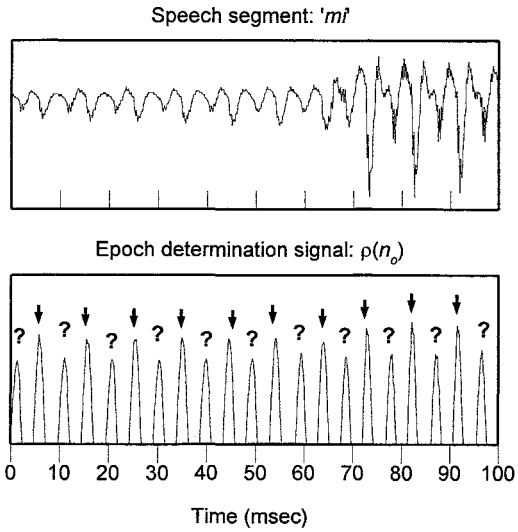

Figure 1. Ambiguities in the original sinusoid-based epoch detection algorithm.

there are also some non optimal peaks (question marks) whose amplitudes are considerable. This makes it difficult, in certain cases, to distinguish between the actual epoch pulse and subpulses. The second problem concerns the sensitivity of the OSED algorithm to noise. The performance of the algorithm is degraded when input speech is corrupted by noise.

3. THE IMPROVED SINUSOID-BASED EPOCH DETECTOR (ISED)

To alleviate the effects of the problems described above, we propose to apply some additional signal processing techniques to the original algorithm. Specifically, in our new algorithm, a Hilbert envelope is used to reduce the ambiguity between epochs and secondary pulses. Also, for improving the performance of our epoch detector in a noisy environment, low-pass filtering of the input speech signal is carried out. The result of these additional techniques, explained below, is a more robust epoch detection algorithm called *improved sinusoid-based epoch detector* (ISED) [9].

3.1 Hilbert Envelope

The Hilbert transform has already been used in various epoch detector algorithms, such as the maximum-likelihood epoch detector [5] and epoch filtering of LP residual technique [4]. The objective is to emphasise the contrast between the actual epoch pulse and subpulses in the error function $\rho(n_o)$ to reduce ambiguities. The Hilbert envelope $f(n_o)$ of the function error $\rho(n_o)$ is calculated as follows [3]:

$$f(n_o) = \left[\rho^2(n_o) + \rho_H^2(n_o)\right]^{1/2} \tag{6}$$

where $\rho_H(n_o)$ is the Hilbert transform of $\rho(n_o)$. It is known that the Hilbert transform can be identified as a filter with the transfer function [3]

$$H(\omega) = \begin{cases} -j & 0 < \omega < \pi \\ 0 & \omega = 0, \pi \\ j & -\pi < \omega < 0 \end{cases} \tag{7}$$

The Hilbert transform of $\rho(n_o)$ is then

$$\rho_H(n_o) = \sum_{l=1}^{L} A_l^2 \sin[\theta_l + n_o\omega_l - \Phi_s(\omega_l)] \tag{8}$$

One way to make $f(n_o)$ more pulse-like is to use average-value subtraction, which allows the null signal between pulses to be obtained

$$\hat{f}(n_o) = \begin{cases} 0 & \text{if } f(n_o) < \bar{f} \\ f(n_o) - \bar{f} & \text{if } f(n_o) \geq \bar{f} \end{cases} \tag{9.a}$$

where $$\bar{f} = \sum_{n_o=-N/2}^{n_o=N/2} f(n_o) \ /(N+1) \tag{9.b}$$

The epoch pulses are now estimated from the peaks in the following expression

$$\rho'(n_o) = \rho(n_o)\, \hat{f}(n_o) \tag{10}$$

3.2 Lowpass Filtering

From Equation (5), it is shown that the $\rho(n_o)$ signal depends on the spectral amplitude A_l in a quadratic way. In order to minimise the effect of a noisy environment in the $\rho(n_o)$ signal, the high frequency region, dominated mainly by the noise spectral component, is discarded in our research work. This is carried out by means of simple lowpass filtering. The cutoff frequency has been set to 0.25 × f_s (2.5 kHz for f_s= 10 kHz). The filter must have a linear phase characteristic (FIR filter) so that the phases of the spectral components in Equation (5) (θ_l and $\Phi_s(\omega_l)$) are not affected. The bandwidth of the transition band can be specified by the order of the FIR filter. In our algorithm, the lowpass filter contains 50 taps.

4. COMPARISON AND RESULTS

In order to study the performance of our proposed algorithm, a comparative study has been done with four well-known techniques for GCI estimation. Specifically, these techniques are as follows: the Epoch Filtering Linear Prediction Residual algorithm (EFLPR)[4], Maximum-likelihood Epoch Detector (MLED)[5], original Sinusoid-based Epoch Detection algorithm (OSED)[3], and the Autocovariance Method (AUTO)[6]. In Table 1, the analysis parameters of the evaluated epoch detectors are shown. The sampling frequency f_s is 10 KHz.

Figure 3 shows an illustrative example of the performance of the five epoch detectors for the speech segment /bo/. Obviously, this example is not conclusive but it gives us an approximate idea about the performance of the algorithms.

	Frame Rate	Frame Length	LPC Order	Lowpass Filter	FFT
AUTO	1	20	12	0.25×f_s	
EFLPR	128	256	12	0.25×f_s	256
MLED	128	256	12	0.25×f_s	256
OSED	128	257	12	0.25×f_s	1024
ISED	128	257	12	0.25×f_s	1024

Table 1. Analysis parameters of the five epoch detectors.

4.1 Data Base for Evaluation

For the evaluation test, an appropriate data base is required to span the range of pitch and types of utterance. Our data base contains 50 voiced speech segments uttered by 4 male and 3 female speakers. Each of the segments is 200 msec long.

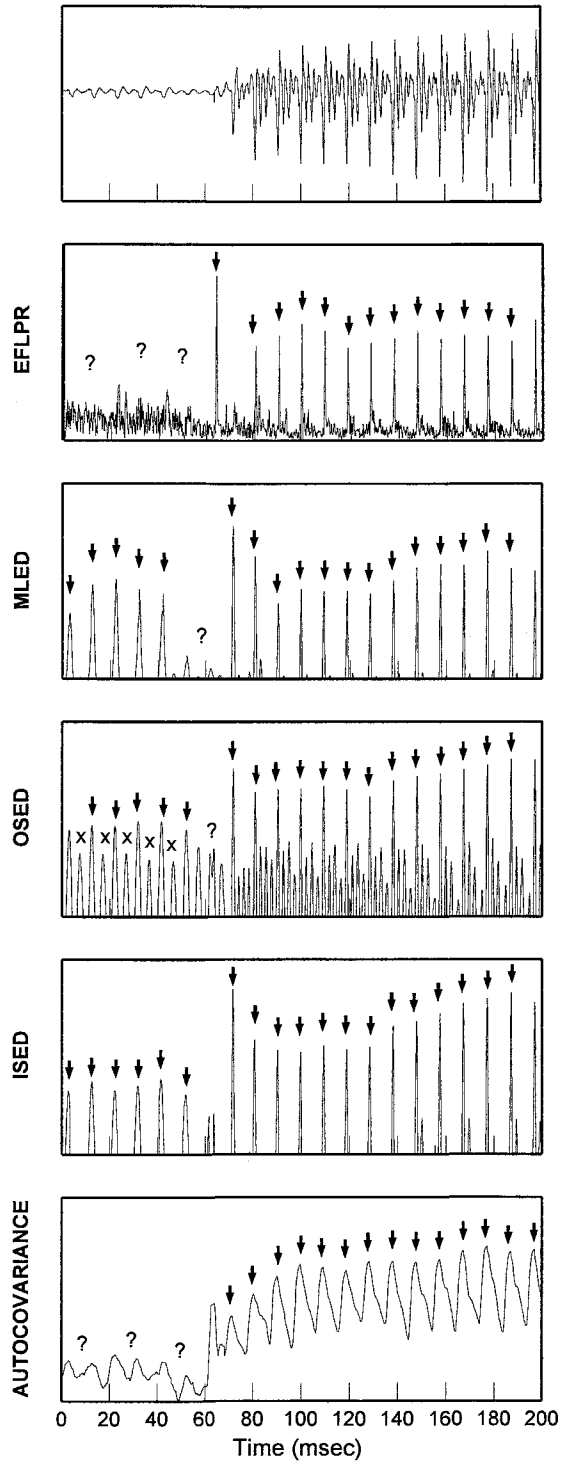

Figure 3. Graphical comparison of the five epoch detectors. The first graph shows the original speech segment (/bo/). The plots from 2 to 6 illustrate the performance of the EFLPR, MLED, OSED, ISED and, finally, the autocovariance method, respectively.

A great variety of voiced sounds is included, such as nasals (/m/,/n/), voiced fricatives (/z/,/f/), voiced stops (/b/,/d/), vowels and diphthongs. Finally, the pitch period ranges from 85 Hz to 340 Hz.

4.2 Definition of Errors

In oder to carry out an objective comparison test, two kinds of errors are defined, namely, *non-detected-epoch* error *and false-epoch* error. These two errors are obtained as follows: firstly, for each of the utterances in the data base, a series of epoch locations was manually obtained. We denote the series of epoch as $p_m(n)$ where n goes from 1 to N_m, and N_m is the number of epochs in the utterance. The subscript m stands for *manual*. Next, each of the utterances was used as input to each of the five epoch detectors and a set of epoch locations was obtained as the output. We denote the series of epochs from the ith epoch detector (i=1, ..., 5) as $\{p_i(n), n= 1, ..., N_i\}$ (in general $N_i \neq N_j \neq N_m$). The next step is to match the epoch in $p_i(n)$ with those in $p_m(n)$. In this matching two kinds of errors occur:

1) *non-detected-epoch error*: this happens when for a given epoch in $p_m(n)$ (say $p_m(l)$) there is no match between any of the epochs in $p_i(n)$. This means that the algorithm i was not able to detect an epoch around $p_m(l)$.

2) *false-epoch error*: this occurs when there exists an epoch in $p_i(n)$ which is not matched to any of the manually obtained epochs, that is, $p_m(n)$. This means that a big secondary or spurious peak in the output signal of the ith algorithm was detected and erroneously considered as an epoch.

4.3 Test1: Clean Data Evaluation

These two kinds of errors was performed on the entire data base and the major results are presented in Figure 3. The figure shows the normalised *non-detected-epoch* and *false-epoch* errors for each of the pitch detection algorithms. It can be seen that the best performance corresponds to the ISED algorithm since it produces the lowest value in both the *non-detected-epoch* and *false-epoch* errors. The OSED algorithm produces almost the same *non-detected-epoch* error as the ISED method (4.4 %). However, the selection of the proper epoch peaks in the output of this algorithm is difficult due to the existence of many spurious and secondary peaks. This problem is reflected in its large value of the normalised *false-epoch* error (22.6 %).

The MLED is the second best algorithm in the evaluation test. Both normalised *non-detected-epoch* and *false-epoch* errors are similar (10.3 % and 12.5 %, respectively) and twice as large as in the ISED method. The worst performance corresponds to the EFLPR and autocovariance techniques. This indicates that the outputs of these algorithms are too ambiguous for detecting correct epochs.

4.4 Test2: Contaminant-Noise Influence Test

For this test, all the segments in the data base were corrupted with different level of Gaussian noise (SNR from 0 to 21 dB). Then, for each SNR level, the *non-detected-epoch* and *false-epoch* errors were computed (see Figure 4). Obviously, all the

curves fall down when increasing the SNR value. From Figure 4 it can be appreciated that the most robust method is the ISED algorithm whereas the worst performance corresponds to the EFLPR and autocovariance methods. A medium performance is obtained by the MLED and the OSED algorithms.

4.5 Discussion

From the above named tests, the following conclusions are derived: in the first place, the autocovariance matrix determinant evaluation method does not work with all sorts of voiced speech segments. Some segments cause great difficulty in determining GCI's (specially, high-pitched and nasal segments). Besides, this method is computationally expensive.

An alternative method is the EFLPR algorithm. This algorithm works well for most clean vowel signals. However, for voiced speech segments which do not possess manifest discontinuities in the derivatives of the glottal airflow it fails. Thus, restrictions of clean data and of certain vowel signals are imposed on each application.

The main problem with the OSED algorithm is that many non-optimal peaks appear around the actual epoch pulse and thus, ambiguities occur. However, by using the Hilbert transform and lowpass filtering, as in the ISED algorithm, the ambiguity is further reduced and the performance and robustness of the algorithm are improved.

Finally, the MLED algorithm presents a good performance and robustness against noise, although the overall performance is slightly worse than the ISED method. However, the MLED is still an attractive technique since it is less computationally expensive and simpler to implement than the ISED method.

5. CONCLUSIONS

In summary, these tests show that the best performance corresponds to the ISED algorithm and therefore, it is demonstrated that the sinusoidal technique is a suitable and powerful signal processing tool for epoch detection.

REFERENCES

[1] E. Moulines and F. Charpentier, "Pitch-Synchronous Waveform Processing Techniques for Text-to-Speech Synthesis Using Diphones", *Speech Communication, no. 9, pp 453-67, 1990*

[2] M. Larreategui and R. A. Carrasco, "A Novel Algorithm Based on Sinusoidal Modelling of Speech for Text-to-Speech System", *EUPSICO '94, vol. 1, pp 12-15, Edinburgh, 1994.*

[3] R. J. McAulay and T. F. Quatieri, "Low-Rate Speech Coding Based on the Sinusoidal Model", *in Advances in Speech Signal Processing (S. Furui and M. M. Sondhi, eds.), pp. 165-208, N. Y.: Markel Dekker, 1992.*

[4] T. V. Ananthapadmanabha and B. Yegnanarayana, "Epoch Extraction from Linear Prediction Residual for Identification of Closed Glottis Interval", *IEEE Trans. on ASSP, vol. 27, no. 4, August 1979.*

[5] Y. M. Cheng and D. O'Shaughnessy, "Automatic Algorithm Estimation of Glottal Closure Instant", *IEEE Trans. on Acoust., Speech and Signal Processing, vol. 37, no. 12, Dec. 1989.*

[6] H. W. Strube, "Determination of the Instant of Glottal Closure from the Speech Wave", *J. Acoust. Soc. Amer., vol. 56, pp. 1625-29, Nov. 1974.*

[7] T. F. Quatieri and R. J. McAulay, 'Shape Invariant Time-Scale and Pitch Modification of Speech', *IEEE Trans. of Signal Processing, vol. 40, no. 3, pp 497-510, March 1992.*

[8] T. F. Quatieri and R. J. McAulay, "Phase Coherence in Speech Reconstruction for Enhancement and Coding Applications", *ICASSP '89, Glasgow, pp 207-10, 1989.*

[9] M. Larreategui, F. J. Ancin and R. A. Carrasco, 'An Improved Pitch-Event Detection Algorithm Based on Sinusoidal Modelling of Speech', *IEE Colloquium on "Speech and Image Processing", London, U.K., digest no. 1995/091, 2 May 1995.*

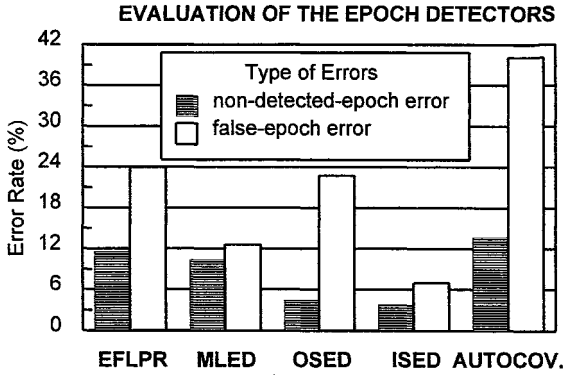

Figure 3: performance of the five epoch detectors for the clean segments in the data base.

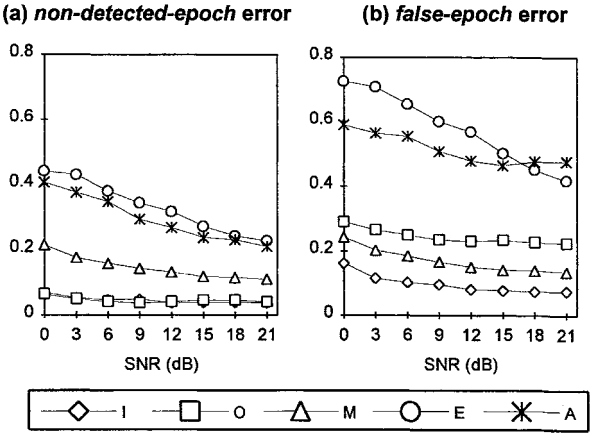

Figure 4: performance of the epoch detectors for the contaminant-noise influence test. The left and right graphs show, respectively, the *non-detected-epoch error* rate and *false-epoch error* rate for different SNR noise contamination (0 to 21 dB). In the box above, 'I' stands for ISED, 'O' for OSED, 'M' for MLED, 'E' for EFLPR and 'A' for Autocovariance methods. The 'y' axis represents normalised error.