A Frobenius Norm Approach to Glottal Closure Detection from the Speech Signal

Changxue Ma, Yves Kamp, and Lei F. Willems

Abstract—**The detection of glottal closure instants has been a necessary step in several applications of speech processing, such as voice source analysis, speech prosody manipulation and speech synthesis. This paper presents a new algorithm for glottal closure detection that compares favorably with other methods available today in terms of robustness and computational efficiency. In this paper, we propose to use the singular value decomposition (SVD) approach to detect the instants of glottal closure from the speech signal. The proposed SVD method amounts to calculating the Frobenius norms of signal matrices and therefore is computationally efficient. Moreover, it produces well-defined and reliable peaks that indicate the instants of glottal closure. Finally, with the introduction of the total linear least squares technique, two proposed methods are reinvestigated and unified into the SVD framework.**

I. Introduction

THE process of speech production can be simply described by a source-filter model [1]. The filter can be characterized as linear [2], [3]. For voiced sounds, the source is situated in the vibrating vocal folds that modulate the air flow from the lungs and produce glottal pulses. The unvoiced sound source consists of the turbulent flow formed somewhere in the constricted vocal tract.

Present speech research shows a great interest in analyzing the voice sound period by period over an interval delimited by two successive instants of glottal closure. For the sake of simplicity, we call the instants of glottal closure in the speech signal the epochs. Determination of the epochs plays an important part in applications, such as the analysis of the glottal pulse waveform by using inverse filtering to extract speaker characteristics [4], [5], [7], prosody manipulations of speech sounds by means of the PSOLA technique [6], and speech synthesis and speech coding [4], [7].

During the past decades, several epoch detection methods have been proposed for the speech signal. One such method is to detect the discontinuities of the differentiated speech signal [8]. It is a simple and effective technique for very clean vowels with sharp glottal closures, but as it is a high-pass filter operation, it is thereby understandably sensitive to the noise excitations in sounds like voiced frictives and contaminating noise. At present, the following two methods are better known because they can produce a reliable glottal closure detection. The first, proposed by Strube, calculates determinants of autocovariance matrices and produces satisfactory detection of the epochs. However, it cannot easily be normalized [10]. The second approach, proposed by Wong, Markel and Gray [11], directly makes use of the speech production model with a clearly defined glottal pulse. Here the epoch is defined as the minimum of the total linear predictive coding (LPC) residual energy calculated from rather short analysis frames. Unfortunately, total LPC residual energy tends to be noisy and therefore needs to be smoothed.

Many epoch detection methods, among which are the two important methods of Strube and Wong *et al.* are, in essence, based on the idea that the linear prediction model fits better and, consequently, its prediction error is smaller within a short segment (less than one pitch period) of the speech signal containing no excitations [9]–[13]. When the instant of glottal closure or main excitation is included in the data segment, the linear prediction model does not fit the data well and the prediction error will be large. These large prediction errors are indications of the instants of glottal closure.

The main contribution of this research is to establish a framework of the epoch detection, to compare the results from different approaches and, finally, to propose a new singular value decomposition (SVD) approach to the epoch detection problem. This approach leads to a better formulation and has clear advantages over the methods of Strube and Wong *et al.*, as it is computationally very efficient and robust against noise. The resulting measure has a dimension of energy and can be easily normalized and thresholded. We are also able to show the relationship between Strube's method, Wong's method, and our SVD approach and its advantages.

In the next section, we introduce our approach more explicitly with a brief description of the singular value decomposition (SVD), the linear least squares (LLS), and the total linear least squares (TLLS) techniques. In the third section, we propose our new SVD-based approach for epoch detection. Finally, our method is compared with two others and examples are given.

II. The SVD as Unifying Framework for Epoch Detection

Epoch detection has often been based on a source-filter model of the speech production. In either parametric or statistical approaches, the all-pole system assumption is usually made [12], [13]. The source of the system is assumed to have an open glottis portion and a closed glottis portion in each pitch period of a voiced sound. Although the shape of the glottal pulse depends on the type of phonation, the rate of transition

Manuscript received July 17, 1991; revised April 15, 1993. The associate editor coordinating the review of this paper and approving it for publication was Dr. David Nahamoo.

The authors are with the Insititute for Perception Research, Eindhoven University of Technology, Eindhoven, The Netherlands.

IEEE Log Number 9215234.

from the closed to the open glottis portion is generally slower than that from the open to the closed glottis portion, and thus the main excitation occurs at the instant of the glottal closure [16]. The differentiation of the main excitation results in a very sharp pulse at the instant of glottal closure. Epoch determination from the speech signal is based on the fact that there is strong and abrupt change of the glottal flow at the instant of glottal closure. A vocal tract can be approximately modeled by a linear time invariant system that imposes a linear relation on the speech samples when no excitation is present. Thus, when the system parameters are well estimated, the deviation with respect to this linear relationship should be small in the closed glottis region and large at the instants of glottal closure.

Therefore, the amount of deviation from the linear prediction is a primary criterion used in different epoch detection approaches. The largest deviation is expected to happen at the instants of glottal closure. The question, however, of how to extract the linear predictability or how to identify the linear relation from the speech signal has a significant influence on the quality of the detection schemes. Moreover, speech sounds are dynamic in nature and the source-filter model of the speech production is inevitably accompanied by the presence of unknown disturbances, parameter variations, and other uncertainties. Therefore, the linear model will only hold approximately and the solution will depend on the error criterion used. In practice, a particular solution is obtained by imposing additional constraints on the problem, such as least squares, maximum likelihood or $l1$-norm, and, accordingly, a variety of estimation schemes are utilized [14]. Among the most popular estimation schemes for linear relation from noisy data, are the LLS and the TLLS schemes. As we shall see, the SVD method can provide a unifying framework in identifying linear relations from data, making the formulation of the problem explicit and guaranteeing the robustness of the numerical solutions. In these estimation schemes, the data matrix or measurement data are in fact modified to meet linear relations imposed on the data. In other words, the data matrix is decomposed into the sum of two matrices, one of which consists of the linear dependent column vectors and another that consists of error elements. The constraints mentioned above are used because they approximately meet the physical requirements of the problem and produce a tractable mathematical solution.

A. Singular Value Decomposition (SVD)

The SVD method has been often used in many applications of digital signal processing. The SVD of a certain data matrix allows a particularly robust separation of signal and noise and is very effective in dealing with noisy data [17]–[22].

Consider a sequence of measurements or observation vectors, consisting of segments of a speech signal obtained by advancing a rectangular window of length $p+1$ samples one sample further successively. The following data matrix can then be formed

$$S = \begin{pmatrix} s_{p+1} & s_p & s_{p-1} & \cdots & s_1 \\ s_{p+2} & s_{p+1} & s_p & \cdots & s_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ s_{p+m} & s_{p+m-1} & s_{p+m-2} & \cdots & s_m \end{pmatrix}. \quad (1)$$

We shall assume that $m \geq p+1$ and that the data matrix has full column rank, i.e., $p+1$. Under these assumptions, it is well-known [26] that there exist orthogonal matrices $U = (\overline{u}_1, \overline{u}_2, \cdots, \overline{u}_m)$ and $V = (\overline{v}_1, \overline{v}_2, \cdots, \overline{v}_{p+1})$ such that

$$S = \sum_{i=1}^{p+1} \sigma_i \overline{u}_i \overline{v}_i^t \quad (2)$$

where $U^tU = V^tV = VV^t = I_{p+1}$, and $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{p+1} > 0$ where σ_i are called the singular values, the superscript t denotes matrix transposition, and I_{p+1} is an identity matrix of order $p+1$. The column vector $\overline{u}_i$ of the matrix U is a normalized eigenvector associated with the eigenvalue σ_i^2 of matrix SS^t. In the same manner, the column vector $\overline{v}_i$ is a normalized eigenvector associated with the eigenvalue σ_i^2 of matrix S^tS. Equation (2) is called the SVD.

It is clear that the SVD method decomposes a data matrix into the sum of $(p+1)$ rank one matrices. The matrix S^tS is the autocovariance matrix of the speech signal and its determinant can be rewritten as

$$\det(S^tS) = \prod_{i=1}^{p+1} \sigma_i^2. \quad (3)$$

Moreover, the Frobenius norm of a matrix $S = \{s_{ij}: 1 \leq i \leq m, 1 \leq j \leq p+1\}$ is defined as

$$\|S\|_F = \left(\sum_{i=1}^{m} \sum_{j=1}^{p+1} s_{ij}^2 \right)^{1/2}. \quad (4)$$

On the other hand, $\|S\|_F^2 = Tr(S^TS)$, where Tr denotes the trace of a matrix. In view of (2) and the orthogonality of matrices V and U, one readily observes that the Frobenius norm can also be expressed in terms of the singular values as (see [26])

$$\|S\|_F = \left(\sum_{i=1}^{p+1} \sigma_i^2 \right)^{1/2}. \quad (5)$$

Hence, the Frobenius norm of S is the square root of the sum of its squared singular values.

B. Strube's Method for Epoch Detection

Let $\overline{s}_i$ denote the ith column vector of matrix S. In the absence of excitation, the linear filter model of order p imposes a linear dependence between the vectors $\overline{s}_1, \overline{s}_2, \cdots, \overline{s}_{p+1}$. Consequently, the determinant of the matrix S^tS as a function of time will increase sharply when the speech segment covered by the data matrix (1) contains an excitation, and it will decrease when this speech segment is excitation-free. Therefore, the determinant value can be used as a way to detect the location of epochs in the signal. This is, in essence, Strube's method for the detection of epochs [10], which in view of (3) is equivalent to computing the product of all squared singular values of matrix S. The Cholesky factorization of S^tS provides, however, an efficient recursive scheme to actually perform this calculation [10].

C. Wong's Approach to Epoch Detection and LLS

The source-filter model used for LPC is based on the assumption that the vocal tract can be approximated by an all-pole filter of order p. Accordingly, the first column of the data matrix S in (1) is assumed to be a linear combination of the other columns, and any deviation from this particular linear dependence is attributed to the excitation produced by the source. This viewpoint is expressed by the following set of equations

$$S\begin{pmatrix}1\\-a_1\\-a_2\\\vdots\\-a_p\end{pmatrix}=\begin{pmatrix}e_{p+1}\\e_{p+2}\\e_{p+3}\\\vdots\\e_{p+m}\end{pmatrix} \tag{6}$$

where e_n is the nth sample of the glottal excitation wave, and a_i the ith predictor coefficient. The least squares solution of the equations above can be obtained from

$$S^tS\begin{pmatrix}1\\-a_1\\-a_2\\\vdots\\-a_p\end{pmatrix}=\begin{pmatrix}E_1\\0\\0\\\vdots\\0\end{pmatrix} \tag{7}$$

where $E_1=\Sigma_{i=p+1}^{m+p}e_i^2$ is the LPC residual energy that can be computed by

$$E_1=\frac{\det(S^tS)}{\det(S^tS)_{11}} \tag{8}$$

where $(S^tS)_{11}$ denotes the principal submatrix obtained by removing the first row and column in matrix S^tS. The epoch detection proposed by Wong, Markel, and Gray [11] is essentially based on the minimum of this normalized residual energy E_1. In practice, the LPC residual energy E_1 is sequentially calculated from the speech samples covered by a short analysis window. When the analysis window includes a glottal pulse, the residual energy will first increase and then sharply decrease (in principle to zero) when the window just leaves the glottal pulse. However, these minima may not be well-defined in real speech due to the facts that the LPC model does not perfectly fit the speech samples and the speech samples are corrupted by noise. Owing to the poor prediction of the vocal tract resonances, the residual does not become zero after the glottal pulse and the minima may not correspond to the instants of the glottal closure. This has been demonstrated in [5] and [15], and further discussion about this method will be developed in the following sections.

Finally, let us also observe that the residual energy can also be expressed in terms of the SVD of matrix S. Indeed, (7) yields

$$\begin{pmatrix}1\\-a_1\\-a_2\\\vdots\\-a_p\end{pmatrix}=(S^tS)^{-1}\begin{pmatrix}E_1\\0\\0\\\vdots\\0\end{pmatrix} \tag{9}$$

and in particular $\alpha E_1=1$ where α is the element in position (1, 1) of $(S^tS)^{-1}$. From the SVD of S, as given in (2), one finds that $(S^tS)^{-1}=\Sigma_{i=1}^{p+1}\sigma_i^{-2}\overline{v}_i\overline{v}_i^t$ and thus $\alpha=\Sigma_{k=1}^{p+1}v_{1k}^2\sigma_k^{-2}$, which finally gives

$$E_1=\frac{1}{\Sigma_{k=1}^{p+1}\frac{v_{1k}^2}{\sigma_k^2}} \tag{10}$$

where v_{1k} are the elements of the first row of the matrix V. If the smallest singular value σ_{p+1} is significantly smaller than the others, the equation above is approximately

$$E_1\approx\frac{\sigma_{p+1}^2}{v_{1,p+1}^2}. \tag{11}$$

The LPC solution is in fact a particular case in a whole family of estimation schemes for linear relation between noisy data. Indeed, p other estimations can be conceived, similar to LPC, but where each column of S in turn is considered to be a linear combination of the p remaining columns. This set of estimations is known as the LLS family [23].

Let E_i denote the residual energy of the ith LLS solution where the column i of the data matrix S is considered to be a linear combination of the other columns. In view of the singular value decomposition (2) of S one obtains as shown hereabove

$$E_i=\frac{1}{\sum_{k=1}^{p+1}\frac{v_{ik}^2}{\sigma_k^2}} \tag{12}$$

and (10) corresponds thus to the case $i=1$ in the LLS family.

D. Total Linear Least Squares (TLLS)

Each of the LLS solutions considered above can be looked as a modification of the original data matrix S such that a rank reduction from $p+1$ to p results. For instance, the LPC (or first LLS) equations (6) can be rewritten as

$$\begin{pmatrix}s_{p+1}-e_{p+1}&s_p&\cdots&s_1\\s_{p+2}-e_{p+2}&s_{p+1}&\cdots&s_2\\s_{p+3}-e_{p+3}&s_{p+2}&\cdots&s_3\\\vdots&\vdots&\vdots&\vdots\\s_{p+m}-e_{p+m}&s_{p+m-1}&\cdots&s_m\end{pmatrix}\cdot\begin{pmatrix}1\\-a_1\\-a_2\\\vdots\\-a_p\end{pmatrix}=\begin{pmatrix}0\\0\\0\\\vdots\\0\end{pmatrix}. \tag{13}$$

With the equation in this form, one can thus consider that the LPC solution achieves the rank reduction by modifying only the first column of S while all other columns remain unchanged [23], [24]. Similarly, the ith LLS solution can be interpreted as modifying only the ith column. In speech signals, however, all data in the matrix S could be contaminated by noise or deviations from the model. In addition, the same elements in the first column of the matrix S occur in other columns as well and should therefore also be changed. Even in pitch synchronous speech analysis [5], the closed-glottis portions of speech, which are often considered excitation free, deviate from the linear model because of noise and nonlinearity. Therefore, it may be unrealistic to modify one

column only in order to fit the linear production model, and it would be more reasonable to modify all elements of the matrix S. This is the point of view adopted by the TLLS, which modifies all data columns. In other words, every element of the data matrix can be changed or perturbed in order to fit the linear relation model.

Let $\hat{S}$ be a perturbed matrix and let $\|S - \hat{S}\|_F$ be the perturbation energy. The TLLS solution to the linear relation model is then obtained by modifying matrix S into $\hat{S}$ such that the following set of equations

$$\sum_{i=0}^{p} \hat{s}_{n-i}\,\alpha_i = 0 \tag{14}$$

where index n runs from $p+1$ to $m+p$, or equivalently

$$\begin{pmatrix} \hat{s}_{p+1} & \hat{s}_p & \cdots & \hat{s}_1 \\ \hat{s}_{p+2} & \hat{s}_{p+1} & \cdots & \hat{s}_2 \\ \hat{s}_{p+3} & \hat{s}_{p+2} & \cdots & \hat{s}_3 \\ \vdots & \vdots & \vdots & \vdots \\ \hat{s}_{p+m} & \hat{s}_{p+m-1} & \cdots & \hat{s}_m \end{pmatrix} \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_p \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \tag{15}$$

is exactly solvable and the perturbation energy is minimized. We thus want to find $\hat{S}$ such that rank $(\hat{S}) \leq p$ and $\|S-\hat{S}\|_F$ is minimized. The solution to this problem is well-known [25], [26] and obtained by deleting in (2) the contribution of the smallest singular value, i.e., σ_{p+1}, assuming for simplicity that $\sigma_p > \sigma_{p+1}$. Thus

$$\hat{S} = \sum_{i=1}^{p} \sigma_i\, \overline{u}_i \overline{v}_i^t. \tag{16}$$

The perturbation error $\|S-\hat{S}\|_F = \sigma_{p+1}^2$, and the linear model is $\overline{v}_{p+1}$ since, in view of the orthogonality of the vectors $\overline{v}_i$, the following holds

$$\hat{S}\overline{v}_{p+1} = 0. \tag{17}$$

Finally, the residual error signal is given by

$$S\overline{v}_{p+1} = \sigma_{p+1}\overline{u}_{p+1}. \tag{18}$$

If, instead of the approach described above, we delete from (2) the contribution of a different singular value $\sigma_i \neq \sigma_{p+1}$, then we obtain a different perturbed matrix $\hat{S}_i$, namely, $\hat{S}_i = \Sigma_{j\neq i}\sigma_j \overline{u}_j \overline{v}_j^t$ for which a different linear model holds, i.e., $\hat{S}_i\overline{v}_i = 0$, but with a relatively larger perturbation error $\|S \neq \hat{S}_i\|\|_F = \sigma_i^2$. Each singular value thus measures the deviation from some corresponding linear model and the sum $(p+1)^{-1}\Sigma_{i=1}^{p+1}\sigma_i^2$ can be considered as an "average" of the deviation from any linear model.

III. The Use of the Frobenius Norm for Epoch Detection

The new criterion for epoch detection proposed here is the arithmetic mean of the squared singular values, namely $C = 1/(p+1)\Sigma_{i=1}^{p+1}\sigma_i^2$. Although it does not seem possible to establish a rigorous and direct connection between the numerical value of C and the instant of glottal closure, the theoretical argument given below shows why this criterion makes sense. Additional support is provided by a series of experiments presented at the end of this section and further developed in Section IV. These experimental results show that the maxima of C indeed correlate with the instants of glottal closure.

We start from (12) for the residual energy of the ith LLS solution, which is rewritten as

$$\frac{1}{E_i} = \frac{1}{\sigma_i^2}\sum_{k=1}^{p+1} v_{ik}^2 \frac{\sigma_i^2}{\sigma_k^2}. \tag{19}$$

Since $(\sigma_i^2/\sigma_k^2) \leq (\sigma_1^2/\sigma_{p+1}^2)$ one has

$$\frac{1}{E_i} \leq \frac{1}{\sigma_i^2}\frac{\sigma_1^2}{\sigma_{p+1}^2}\sum_{k=1}^{p+1} v_{ik}^2 \tag{20}$$

and, in view of the orthogonality of the matrix V, which implies $\Sigma_{k=1}^{p+1} v_{ik}^2 = 1$, the latter inequality reduces to

$$\sigma_{p+1}^2 \leq \sigma_i^2 \leq \frac{\sigma_1^2}{\sigma_{p+1}^2} E_i. \tag{21}$$

On the other hand, the inequality between geometric and arithmetic means yields

$$\left(\prod_{i=1}^{p+1}\sigma_i^2\right)^{(1/p+1)} \leq \frac{1}{p+1}\sum_{i=1}^{p+1}\sigma_i^2. \tag{22}$$

Considering (21) and (22) we finally get

$$\begin{aligned}\left(\prod_{i=1}^{p+1}\sigma_i^2\right)^{(1/p+1)} &\leq C = \frac{1}{p+1}\sum_{i=1}^{p+1}\sigma_i^2 \\ &\leq \frac{\sigma_1^2}{\sigma_{p+1}^2}\frac{1}{p+1}\sum_{i=1}^{p+1} E_i.\end{aligned} \tag{23}$$

This double inequality provides the rationale for the new criterion. Indeed, C lies between an upper and a lower bound, both of which can be considered as measuring the deviation of the speech data from the linear dependence model. The lower bound $(\Pi_{i=1}^{p+1}\sigma_i^2)^{1/p+1}$ is in fact Strube's criterion in view of (3). On the other hand, except for the scaling factor $\sigma_1^2/\sigma_{p+1}^2$, the upper bound is the arithmetic mean of the residual energies E_i associated with each of the LLS solutions, and it can thus be considered as an "average" deviation of the data from linear dependence. By definition, these lower and upper bounds will both increase in the open phase of the glottal pulse and will both decrease in the closed glottis portion when the linear dependence between the columns of the data matrix is better realized. Consequently, one can reasonably expect that the new criterion C will follow a similar behavior in view of the fact that it lies between these bounds.

One observes incidentally that a similar argument also holds for the individual singular values in view of the double inequality (21). Indeed, the lower bound σ_{p+1}^2 measures the deviation from the linear model for the TLLS solution and the upper bound is, within a scaling factor, the residual energy E_i of the ith LLS solution. Both can be expected to increase when

TABLE I
COMPARISON OF THE COMPUTATIONAL COST FOR THE THREE METHODS ($m > p$)

	Initial Cost	Time Updating
New criterion	mp	$2p$
Strube's criterion	$mp^2 + p^3/3$	$mp + p^3/3$
Wong's criterion	$mp^2 + p^2$	$mp + p$

the excitation is present, i.e., when the data do not comply with the linear model. It is then not surprising that the experimental observations presented in the next section support the fact that each singular value increases in the open glottis portion of the signal. Finally, it should be noted that (23) provides a tighter lower bound than would result from straightforward addition of the inequalities (21) over index i, since σ_{p+1}^2 is of course smaller than the geometric mean of the squared singular values.

It should be stressed that the new criterion is very efficient from a computational point of view. First, (4) and (5) show that the numerical value of C can be obtained simply by calculating the Frobenius norm of the data matrix S without the need of actually performing a singular value decomposition. Moreover, our criterion can easily be updated when a new sample comes in the observation window. The sequential computation of the Frobenius norm of the matrix reduces to adding the sum of the squared entries of the last row of the matrix and to subtracting the sum of the squared entries of the first row of the preceeding matrix. The initial computation of the Frobenius norm of the data matrix (1) is $O(mp)$, and updating this norm when the observation window advances by one sample requires $O(2p)$ operations. For the methods of Strube and Wong *et al.*, however, one has first to compute the product S^TS, which needs mp^2 operations, and the cost of updating this product is mp. Calculation of det (S^TS) in Strube's method requires $O(p^3/3)$ operations, and no economical updating formula is known. For the method of Wong *et al.*, the computational cost of the total prediction error E_1 is $O(p^2)$, while its time update needs $O(p)$ operations. These results are summarized in Table I, and taking into account that $m > p$, it shows that the new criterion is significantly more efficient both in initial phase and in time updating.

IV. EXPERIMENTAL RESULTS AND COMPARISON

The arguments presented above are now corroborated by experimental evidence and comparisons are made between the methods of Strube and Wong *et al.* Since an electroglottograph (EGG) reflects the vibratory motion of vocal folds and has been used as a tool to validate speech processing algorithms [27], we will use the EGG signal as a reference for comparison. We will also use waveforms of differentiated glottal pulses as references. The glottal pulse is the excitation source for synthetic vowels and is generated according to the LF-model [28]. In the experiments, the speech signals were sampled at a sampling frequency of 10 kHz and preemphasized with a filter $(1 - 0.9z^{-1})$. The analysis interval was 30 samples long in total and the prediction order p was 10.

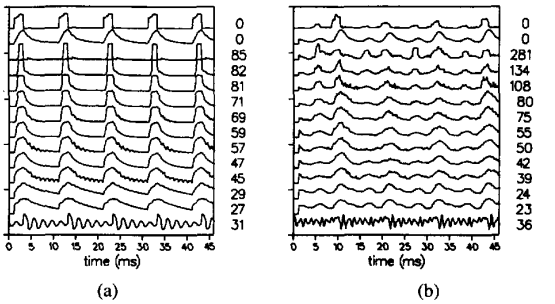


Fig. 1. From top to bottom, the total LPC residual energy, the sum of squared singular values, and 11 singular value curves obtained from a) a synthetic vowel /a/ with impulse excitation and b) a natural vowel /a/. The singular values are ordered and scaled (indicated by the numbers in the figures), the smallest one on the top (the third curve in the figure), the speech waveform on the bottom.

We decide to locate the time position of the maxima in the Frobenius norm of the signal segment at time $t = p + 1$, given that the signal segment extends from $t = 1$ to $t = m + p$, where m is the number of the equations and p is the order of the linear model. The reason for this is that a maximum in the Frobenius norm appears when the excitation point just enters the first row of the data matrix S. This can be seen from (1): when s_{p+1} is the excitation point, and when the analysis interval shifts further, there will be fewer rows of the data matrix that contain the excitation point. Thus, beyond $t = p + 1$, the perturbation energy starts to decrease. As a consequence, the maxima have been delayed with respect to the speech signal. The amount of the delay is equal to the number m of prediction equations and this delay has been compensated for in all the figures.

Fig. 1 shows the time evolution of 11 singular values obtained from a synthetic vowel /a/ with impulse excitation (Fig. 1(a)) and for a natural vowel /a/ (Fig. 1(b)). The speech signal is displayed on the bottom of the figure. For comparison, the total LPC residual energy of Wong's criterion is shown on the top of the figure. The time functions of the 11 singular values are scaled in the display and ordered such that the smallest singular value is the uppermost (i.e., third curve from the top of the figure). The sum of all the squared singular values is shown just above the curve for the smallest singular value. It can be seen that all singular values exhibit local maxima when the analysis interval just comes across the excitation; the new criterion being the arithmetic mean of the squared values, which also shows maxima that coincide with the occurrences of the glottal pulses. From Fig. 1(a) and 1(b), one sees also that the peaks of the new criterion coincide with the sharp down-going edges in Wong's criterion.

The instant of glottal closure can be determined *a priori* via the EGG signals for natural vowels or the excitation waveform for synthetic vowels. To compare with the EGG signals, we show in each panel of Fig. 2, from top to bottom, the results of Wong's method, Straube's method, the new method, the EGG waveform, and the speech waveform, for a natural vowel. The delay of the speech waveform with respect to the EGG signal caused by the propagation of the speech waveform from the glottis to the microphone (mounted on

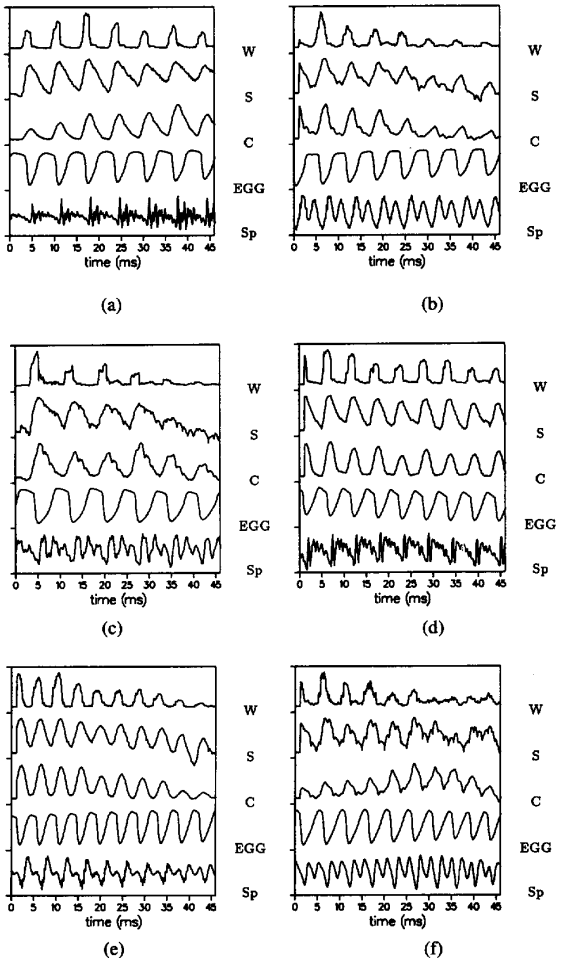


Fig. 2. Comparison of the results obtained by Wong's method (W), Strube's method (S), the new method (C), and the electroglottograph (EGG). Sp corresponds to the speech waveform. (a), (b), and (c) show results obtained from natural vowels /a,i,u/, uttered by a male, respectively; (d), (e), and (f) show results obtained from vowels /a,i,u/, respectively, uttered by a female.

the Rothenberg flowmask) is compensated for in the figures. Fig. 2(a)–(c) give the results of the three methods for natural vowels /a/, /i/, and /u/, uttered by a male. Fig. 2(d)–(f) show the results from natural vowels /a/, /i/, and /u/, uttered by a female. Curve C for the new criterion shows clearly-defined peaks for all vowels. For most of the vowels, one can observe that Wong's method produces sharp down-going edges that correspond to the local maxima in both Strube's criterion and the new criterion. Compared with the EGG signals, we see that the epochs determined by the three methods coincide with the sharp down-going transitions in the EGG waveforms. One sees, however, that for the female vowel /u/ the criteria of Strube and Wong *et al.* are not reliable for the determination of epochs.

To use the excitation waveform as a reference, we show in Fig. 3(a) and (b) the results obtained from two synthetic vowels /a/ and /u/, respectively. Each panel shows, from top to bottom, the results of Wong's method, Strube's method, the new method, the waveform of differentiated glottal pulses,

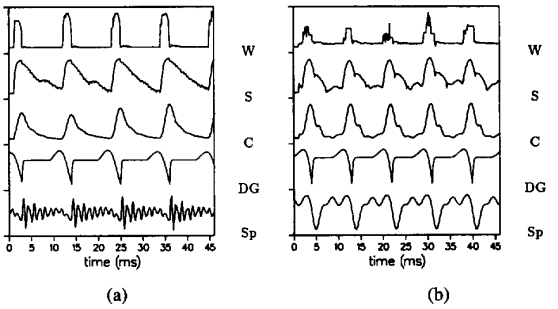


Fig. 3. Comparison of the results obtained by Wong's method (W), Strube's method (S), the new method (C), and the differentiated glottal pulses (DG). (a) and (b) show results obtained from synthetic vowels /a,u/, respectively.

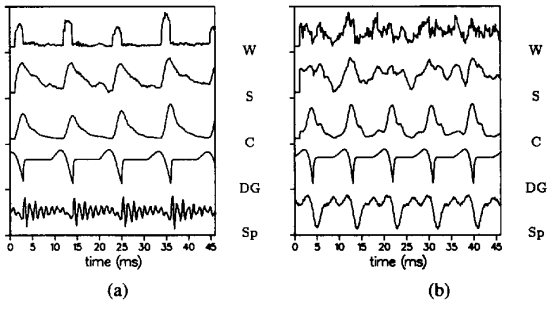


Fig. 4. (a) and (b). Results obtained from the same vowels /a,u/ used in Fig. 3, but with white-noise signals added; SNR = 20 dB.

and synthetic vowels /a/ and /u/. From the waveform of differentiated glottal pulses we determine exactly the instants of glottal closure. It can be seen that the new criterion and Strube's criterion produce clear peaks at the main excitation instants.

In Fig. 4, we illustrate the effectiveness of the three algorithms in the noisy condition. The speech signal was corrupted by additive white noise with an SNR of 20 dB. The same vowels /a/ and /u/ used in Fig. 3 are used here. One sees from Fig. 4(a) for the vowel /a/ that the instants of glottal closure can be reliably obtained from all three methods. From Fig. 4(b) of the vowel /u/, it is rather difficult to detect the glottal closure instants from the criteria of Strube and Wong *et al.* One sees from Fig. 4 that the new criterion is robust in the noisy condition.

Strube, in fact, uses the logarithm of the determinant of the autocovariance matrix to determine the epochs. In view of (3), the actual criterion he uses is thus $\Sigma_{i=1}^{p+1} \log \sigma_i^2$, whereas our criterion is $\Sigma_{i=1}^{p+1} \sigma_i^2$. Consequently, the dynamics of the singular values is nonlinearly compressed in Strube's method with the consequence that the peaks will be less prominent.

We have seen from (11) that Wong's criterion is approximately determined by the time function of the smallest singular value. We observe from Fig. 1 that, as a function of time, the smallest singular value tends to exhibit flat tops and bottoms, with sharp transitions. Therefore, both the maxima and the minima are not well defined. However, the sharp down-going transitions from Wong's criterion coincide with the location of the epoch.

V. Conclusion

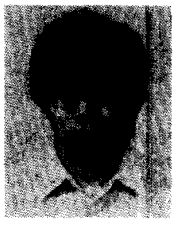
In summary, a new epoch detection technique is proposed in which only the Frobenius norm of the linear predictive matrix need be computed. The sequential computation of the Frobenius norm of the matrix is reduced to just the addition of the sum of the squared entries of the last row of the matrix and the subtraction of the sum of the squared entries of the first row of the preceding matrix. The determination of the instants of glottal closure by calculating the Frobenius norm of the matrix of the speech signal shows clear advantages over the methods of Strube and Wong *et al.* in computational efficiency and noise sensitivity. Finally, all three methods are interpreted in the unifying framework of singular value decomposition.

Acknowledgment

The authors are very grateful to the anonymous reviewers whose comments greatly helped in improving the original version of this paper. We thank also B. Cranen, Department of Language and Speech, Nijmegen University, the Netherlands, for providing us with the speech along with the EGG data.

References

[1] J. L. Flanagan, *Speech Analysis, Synthesis and Perception,* 2nd ed. New York: Springer-Verlag, 1972.
[2] J. D. Markel and A. H. Gray, *Linear Prediction,* 2nd ed. The Hague, The Netherlands: Mouton, 1970.
[3] J. Makhoul, "Linear prediction: a tutorial review," *Proc. IEEE,* vol. 63, no. 4, pp. 561–580, Apr. 1975.
[4] J. H. Eggen, "A glottal-excited speech synthesizer," IPO Ann. Progress Rep. 24, 1989.
[5] H. Kuwabara, "A pitch-synchronous analysis/synthesizer system to independently modify formant frequencies and bandwidth for voiced speech," *Speech Commun.,* vol. 3, no. 3, pp. 211–220, Dec. 1984.
[6] F. Charpentier and E. Moulines, "Pitch-synchronous wave form processing techniques for text-to-speech synthesis using diphones," in *Proc. EUROSPEECH-89,* vol. 2, 1989, pp. 13–19.
[7] P. Hedelin, "High quality glottal LPC-vocoding," in *Proc. Int. Conf. Acoust., Speech, Signal Processing* (Tokyo), 1984, pp. 465–468.
[8] T. V. Ananthapadmanabha and B. Yegnanarayana, "Epoch extraction of voiced speech," *IEEE Trans. Acoust., Speech, Signal Processing,* vol. ASSP-23, no. 6, pp. 562–570, Oct. 1975.
[9] T. V. Ananthapadmanabha and B. Yegnanarayana, "Epoch extraction from linear prediction residual for identification and closed glottis interval," *IEEE Trans. Acoust., Speech, Signal Processing,* vol. ASSP-27, no. 4, pp. 309–319, Aug. 1979.
[10] H. W. Strube, "Determination of the instant of glottal closures from the speech wave," *J. Acoust. Soc. Am.,* vol. 56, no. 5, pp. 1625–1629, Nov. 1974.
[11] D. Y. Wong, J. D. Markel, and A. H. Gray, Jr., "Least squares glottal inverse filtering from the acoustic speech waveform," *IEEE Trans. Acoust., Speech, Signal Processing,* vol. ASSP-27, no. 4, pp. 353–362, 1979.
[12] Y. M. Cheng and D. O'Shaughnessy, "Automatic and reliable estimation of glottal closure instant and period," *IEEE Trans. Acoust. Speech, Signal Processing,* vol. 37, no. 12, pp. 1805–1815, Dec. 1989.
[13] E. Moulines and R. Di Francesco, "Detection of the glottal closure by jumps in the statistical properties of the speech signal," *Speech Commun.,* vol. 9, nos. 5/6, pp. 401–418, Dec. 1990.
[14] J. Lansford and R. Yarlagadda, "Adaptive L_p approach to speech coding," in *Proc. Int. Conf. Acoust., Speech, Signal Processing,* 1988, pp. 335–338.
[15] J. N. Larar, Y. A. Alsaka, and D. G. Childers, "Variability in closed phase analysis of speech," in *Proc. Int. Conf. Acoust., Speech, Signal Processing,* 1985, pp. 29.2.1–29.2.4.
[16] E. B. Holmberg, R. E. Hillman, and J. S. Perkell, "Glottal airflow and transglottal air pressure measurements for male and female speakers in soft, normal, and loud voice," *J. Acoust. Soc. Am.,* vol. 84, no. 2, pp. 511–529, Aug. 1988.
[17] J. Vandewalle and B. De Moor, "A variety of applications of singular value decomposition in identification and signal processing," in *SVD and Signal Processing,* F. Deprettere, Ed. New York: North Holland, 1988, pp. 43–91.
[18] H. C. Andrews and C. L. Patterson, "Singular value decomposition and digital image processing," *IEEE Trans. Acoust., Speech, Signal Processing,* vol. ASSP-24, no. 1, pp. 26–53, Feb. 1976.
[19] V. C. Klema and A. J. Laub, "The singular value decomposition: its computation and some applications," *IEEE Trans. Automat. Contr.,* vol. AC-25, no. 2, pp. 164–176, Apr. 1980.
[20] B. S. Atal, "Efficient coding of LPC parameters by temporal decomposition," in *Proc. Int. Conf. Acoust., Speech, Signal Processing,* 1983, pp. 81–84.
[21] D. W. Tufts and R. Kumaresan, "Estimation of frequencies of multiple sinusoids: making linear prediction perform like maximum likelihood," *Proc. IEEE,* vol. 70, no. 9, pp. 975–989, Sept. 1982.
[22] B. S. Atal, "A model of LPC excitation in terms of eigenvectors of the autocorrelation matrix of the impulse response of the LPC filter," in *Proc. Int. Conf. Acoust., Speech, Signal Processing,* 1989, pp. 45–48.
[23] B. De Moor and J. Vandewalle, "A unifying theorem for linear and total linear least squares," *IEEE Trans. Automat. Contr.,* vol. 35, no. 5, pp. 563–566, May 1990.
[24] S. Van Huffel and J. Vandewalle, "The total least squares technique: computation, properties and applications," in *SVD and Signal Processing,* F. Deprettere, Ed. New York: North-Holland, 1988, pp. 189–207.
[25] G. H. Golub and C. F. Van Loan, "An analysis of the total least squares problem," *SIAM, J. Numer.,* vol. 17, no. 6, pp. 883–893, Dec. 1980.
[26] G. H. Golub and C. F. Van Loan, *Matrix Computation.* Baltimore: Johns Hopkins University Press, 1983.
[27] A. K. Krishnamurthy and D. G. Childers, "Two-channel speech analysis," *IEEE Trans. Acoust., Speech, Signal Processing,* vol. ASSP-34, no. 4, pp. 730–743, Aug. 1986.
[28] H. Fujisaki and M. Ljungqvist, "Proposal and evaluation of models for the glottal source waveform," in *Proc. Int. Conf. Acoust., Speech, Signal Processing,* 1986, pp. 1605–1608.



Changxue Ma was born in Hubei, China, on March 15, 1958. He received the B.Sc. degree in 1982 and the M.Sc. degree in 1986, both from Wuhan University, China, and the Ph.D. degree in 1992 from the Eindhoven University of Technology, Eindhoven, the Netherlands, all in electrical engineering.

He joined the Institute for Television and Audio Research, Beijing, China, in 1982, and was involved in research work on television transmitting systems. In 1986 he became affiliated with the Department of Computer Science at Wuhan University, where he engaged in computer vision research. From 1988 to 1992 he was a Research Assistant at the Institute for Perception Research, Eindhoven, the Netherlands, where he engaged in research work in speech coding, speech processing, speech perception, and signal processing. He is now with INRS Telecommunications, Canada, working on speech recognition. His interests include signal processing, speech modeling, speech coding, and speech recognition.

Yves Kamp received the Engineer degree in electrical and mechanical engineering in 1959 and the Doctoral degree in Applied Sciences in 1966 from the University of Louvain, Belgium.

From 1961 to 1967 he was Professor of Electrical Engineering at the University Lovanium (Republic of Zaire). In 1967 he joined the Philips Research Laboratory, Belgium, where he became Head of the research group on Advanced Information Processing. He is presently Advisor for the Institute of Perception Research, Eindhoven (the Netherlands) and part-time professor at the Catholic University of Louvain (Belgium). His research activity has touched on several areas of applied mathematics, among which are stability of multivariable systems, fast algorithms for signal processing, spectral estimation, speech recognition, and neural networks.

Lei F. Willems received the Engineer degree in electrical engineering in 1961 from the Delft University of Technology, the Netherlands.

From 1963 to 1991 he worked at the Philips Research Laboratory, Eindhoven, the Netherlands, where he was affiliated with the Institute for Perception Research. Presently, he is Advisor to the Institute for Perception Research. His research activity was in the field of speech signal processing, mainly pitch detection, speech manipulation, and speech synthesis.