Detection of the glottal closure by jumps in the statistical properties of the signal

R. Di Francesco, E. Moulines
CNET LAA/TSS 22301 Lannion Cedex, France

Abstract: Two new methods for the detection of the instant of glottal closure directly from the speech waveform are presented here. Both detect the abrupt changes in the short-term spectral characteristics of the speech signal within a pitch period caused by glottal events. The same statistical approach of sequential detection of events by hypothesis testing is used for this purpose. The first method is based on the maximization of a likelihood ratio, whereas the second one uses a divergence convexity test. Experiments on real speech demonstrate the improved robustness of these methods.

I. Introduction

The problem of estimating the instant of glottal closure (and opening) directly from the acoustic signal has been studied by many researchers [1] [2], and is crucial for several applications. For example, such an estimation is important for the design of a glottal source model for improving the quality of speech synthesis by rules. Another application has been found in very low bit rate speech coding [4].

Several approaches have been tried but a really satisfactory solution has yet to be found [1] [2] [4]. The methods can be divided roughly into two categories.

For the first category, the glottal waveform is fully described by a parametric model, usually derived from physiological measurements and/or inverse filtering experiments [5]. The parameters of the glottal source waveform are identified together with the vocal tract model so as to minimize a suitable distortion criterion between the original and the reconstructed speech signal [3] [4]. The instants of glottal closure and opening are then found directly from the set of estimated parameters. These methods suffer from the two following drawbacks:

- the formulation results in a highly non-linear problem. Consequently, the minimization is not straightforward and is computationally intensive,
- a parametric representation of speech production including a glottal model cannot be appropriate for every kind of speech sounds. For example, such model often fails to describe accurately nasals, voiced fricatives and liquids.

For the second class of methods, the detection is based on the short term evolution of the spectral characteristics of the speech signal.

They rely on a classical detection schema:

- a criterion is computed from the observed speech signal,
- according to one or several thresholds a decision is taken.

Many criteria have been studied: they usually rely on the assumption that the vocal tract is not excited during an interval within each pitch period, just after the instant of glottal closure where the main excitation occurs. In this small interval, the speech waveform is assumed to freely decay (sum of damped sinusoids, the decay being specified by the vocal tract resonances). Based on this analysis, Strube proposed a measure for the linear predictability of the signal over an interval. The covariance matrix associated with prediction order p has a small determinant if the signal samples within the frame are related by a linear equation. Strube showed that the instant of glottal closure corresponds to the beginning of the frame from which the logarithm of the determinant reaches a maximum. Wong proposed a different criterion based on the computation of a normalized total square error: a sequential covariance analysis is performed on frames lasting a fraction of the pitch period, total residual energy is then computed, and normalized by the total energy of the signal on the same analysis frame. Due to the assumed high predictability of the signal during the closed glottis interval, the residual energy is supposed to be much smaller than the energy of the signal in this interval. Thus, the criterion is supposed to jump to very small values just after the instant of the glottal closure, leading to the detection.

We have found that both methods perform reasonably well for high-quality input speech with low fundamental frequency and well defined closed phase. An example is shown in (Fig.1.a); the speech waveform was produced by a low-pitched male voice. Wong's normalized residual error criterion shows well defined minima, though, even in that case, the exact instant of glottal closure is rather difficult to determine. Moreover, in many situations, when the closed phase is of shorter duration, or in the case of high frequency or breathy speech, these methods do not provide unambiguous detection (Fig 1.b).

II. A sequential detection scheme

To overcome such a drawback, we define and compare two methods for improving the robustness of the estimation. As the methods presented above, they derive the instant of glottal closure and opening directly from the acoustical speech waveform. These new methods are based on the detection of abrupt changes in the short-time spectral characteristics of the speech signal.

The underlying production mechanism consists in two steady-states in each pitch-period: one state for an interval when the glottis is predominantly closed, and the other when the glottis is almost open. The glottal closure and opening are transient phenomena lasting a much shorter duration than the pitch period. The speech waveform is therefore represented by concatenating the response of two all-pole systems excited by white Gaussian noise. These responses alternate synchronously with closed and open glottis intervals and they are separated by transition zones centered around the glottal closure and opening, when the main vocal tract resonances occur.

To detect such events, a Generalized Maximum Likelihood approach is first investigated. This approach is off-line and computationally intensive. It is assumed that the observation is the succession of two (unknown) Gaussian AR processes modeling the two physical steady-states of closed (resp. open) glottis interval. The problem is to determine the instant of the transition between the models. We use a sequential maximum likelihood method: for each possible transition time r, two consecutive models (until r, and starting at r respectively) are identified. A single model is simultaneously identified under the hypothesis of no change. The likelihood ratio $Z(r)$ between the alternatives "one change at r" and "no change" is then computed. It provides a test signal used in a subsequent detection procedure.

A second method is then studied: glottal events are located by detecting jumps in the divergence between a short-term PDF (1 or 2 msec of speech signal) and a long-term PDF (up to one pitch period). These two PDF's are estimated under gaussian AR assumptions: the short-term model by a sliding window covariance algorithm and the long-term model by a growing window Burg adaptive procedure. During the steady-states, the shape of divergence function is shown to be convex, whereas during the transient states, the divergence jumps down, leading to the detection of glottal events.
III. Generalized Maximum Likelihood (GML)

A) Construction of the test signal

A window of small fixed size 2N is sequentially shifted on one sample. At each step we consider an observation vector of size 2N made of the speech samples between time \( t - N \) and time \( t + N - 1 \). We choose between two hypotheses:

- either \( H_0(t) \) no statistical change occurs during the observation interval and the speech signal is the realization of a single gaussian autoregressive (AR) process on \( [t - N, t + N - 1] \)
- or \( H_1(t) \) a single change occurs at time \( t \). The signal can then be described by a first gaussian AR process between time \( t - N \) and \( t - 1 \) followed by a second (and strictly different) gaussian AR process between \( t \) and \( t + N - 1 \).

More precisely,

\[
H_0(t) \quad x(t) = \sum_{i=1}^{2N} a_i x(t - k) + \sigma_i^2 n(t) \quad \forall t \in [t - N, t + N]
\]

\[
H_1(t) \quad x(t) = \sum_{i=1}^{N} a_i x(t - k) + \sigma_i^2 n(t) \quad \forall t \in [t - N, t - 1]
\]

\[
x(t) = \sum_{i=1}^{N} a_i x(t - k) + \sigma_i^2 n(t) \quad \forall t \in [t, t + N - 1]
\]

where \( n(t) \) is a white noise process with unit variance. To test \( H_0(t) \) against \( H_1(t) \), the likelihood ratio of the observation vector is computed [8].

\[
Z(t, \theta_0, \theta_1) = \frac{p(Y; \theta_1)}{p(Y; \theta_0)}
\]

where \( \theta = (a, \sigma) \) is the AR filter parameter vector and \( \sigma^2 \) is the excitation variance, for each underlying gaussian AR model. Moreover:

\[
Y_1 = (y(t - N), y(t - N + 1), \ldots, y(t - 1))
\]

\[
Y_0 = (y(t), y(t + 1), \ldots, y(t + N - 1))
\]

\[
Y_2 = (y(t - N), y(t - N + 1) \ldots, y(t, \ldots, y(t + N - 1))
\]

are the observation vectors. The parameters of the models have to be estimated for each time \( t \). A maximum likelihood estimation (MLE) has been retained for this purpose [6]. The use of MLE for parameter estimation can be justified for large data records by asymptotic unbiased and minimum variance properties. For shorter data records, the MLE of the AR parameters does not attain the Cramer-Rao lower bound: in this case, the use of the MLE is justified by the good accuracy it affords in comparison with other possible estimators. With this ML estimator, the test signal becomes:

\[
Z(t) = \max_{\theta_0, \theta_1} \frac{p(Y_1; \theta_1) p(Y_2; \theta_2)}{p(Y_1; \theta_0) p(Y_2; \theta_0)}
\]

B) Parameters selection and detection

The use of a fixed window length 2N is inappropriate because of the variability of the pitch period value. The window length is adapted to the local pitch period \( T \). In the experiments, this ratio has been set to 1/3, but this value doesn't seem to be critical. The pitch period is computed by a standard time domain pitch detection method (a normalized cross-correlation followed by a dynamic programming based pitch-tracker [7]).

The selection of the model order is an important task. Too low an order results in a smoothed estimate and a loss in resolution, while too large an order causes spurious spectral information and general statistical instability. An appealing solution is to monitor the model order by using an estimate provided by the Akaike Information Criterion (AIC).

The AIC is computed for the model \( \theta \) and is averaged by a first-order low pass filter with a time constant of about 30 ms (appropriate rounding are made to obtain odd integer value). The maximum value for the predictor order in AIC calculation is chosen to be less than one third of the frame size in order to prevent instability. The model order \( p \) is set to the mean value of the AIC criterion.

Once the test signal is computed, the detection of glottal closure point is fairly easy, because the test is most of the time unambiguous. A dynamic programming based peak picking algorithm working on the differentiated test signal, with a search path constrained by the local period is sufficient to process reliably continuous speech (in the case of high quality linear-phase recording) [8].

IV. Divergence convexity detection (CVX)

The previous method has been designed for off-line applications, and assumes the knowledge of the pitch period value. We now present a method which has been designed for on-line applications in the very constraining context of real-time speech coding. It does not require any predetermination of the pitch value and is based on an easily computable test function derived from the input signal by linear filtering, and very few multiplications. In order to obtain robustness (e.g. to speaker variability, background noise,...) the number of adjustable parameters is very small. A suitable set of parameters has been assessed on a large database, and fixed for all further experiments. The other main point is that this method is directly connected to Information Theory, and has therefore a general interpretation.

The divergence test has been first applied in speech as a tool for the segmentation of the acoustical signal into phonemic units [9] [10]. More recently, a segmentation system also based on the divergence, but operating at a different time-scale, (i.e. smaller than the pitch interval versus a few pitch periods in [10] and with a different decision process has been designed for application to variable rate coding of speech [12]. We will show below how this second statistical method can be used for the detection of the instant of maximum glottal excitation.
The test compares sequentially two gaussian AR models: a long-term model \( \theta(t) = (a, c, \sigma) \) identified by an adaptive Burg procedure on a growing window [0,1] (the forgetting factor \( \lambda \)) from the last detected closure at time 0 to the current time \( t \), and a short term model \( \theta'(t) = (a', c', \sigma) \) estimated on a small fixed-size sliding window (typically 2 ms) \([t-2, t] \) by a standard covariance method [15].

As opposed to the GML algorithm we do not use joint probability densities estimated on the observations but consecutive one-sample models: \( \theta(t) \), \( i = 0, 1 \), yields a prediction \( \hat{y}'(t) \) of the signal \( y(t) \), and a probability distribution \( P_t' \) on the signal considered as a realization of a random process:

\[
y'(t) = \hat{y}'(t) - y(t)
\]

\[
\sigma_t^2 = \text{Var}(y(t) - \hat{y}'(t)) = \text{Var}(c_t')
\]

\[
dP_t'(y) = \frac{1}{\sigma_t'} e^{-\frac{1}{2} \frac{y(y')}{\sigma_t'}}
\]

We can now compute the a posteriori likelihood ratio between the models \( \theta(t), \ i = 0, 1 \) and subtract its mean expectation under the hypothesis of stability, i.e. under \( P_0 \). This expectation is in fact the directed divergence between \( P_0 \) and \( P_1 \) (which is non-negative and vanishes if and only if the PDF's are equal [11]):

\[
E_P \left[ \log \frac{dp_t'}{dp_t} \right] = D(P_0 || P_1)
\]

The resulting quantity is the increment for the divergence test:

\[
w(t) = \log \frac{dp_t'}{dp_t} - D(P_0 || P_1)
\]

(3)

An easily computable expression is obtained for the gaussian case:

\[
w(t) = \frac{c_t' - y(t) - \hat{y}(t)}{\sigma_t'}
\]

(4)

with \( c_t' = \text{Var}(y(t) - \hat{y}'(t)) \) the standard deviation of the forward prediction error, \( \sigma_t = \sigma_t' \) the ratio of the standard deviations.

The divergence test is defined as the cumulative sum \( W(t) \) of the increment \( w(t) \) between the instants \( s = 0 \) where the last jump has been detected and the current time \( t \). Two classical properties of this test [9] are:

\[
E_P [w(t)] = 0
\]

(5)

\[
E_P [w(t)] = -D(P_0 || P_1) - D(P_0 || P_1) = -I(P_0, P_1) < 0
\]

(6)

where \( I(P_0, P_1) \) is the mutual information between distributions \( P_0 \) and \( P_1 \), [11], which is always non-negative. Thus, the theory predicts that:

- \( W(t) \) stays close to zero in the case of a stable statistical behavior. Then \( P_0 \) is the best description of the signal and (5) yields a zero expectation for the first derivative \( w(t) \) of \( W(t) \). - \( W(t) \) jumps down if a shock (or instability) occurs, because in this case \( P_0 \) becomes better than \( P_1 \). Then (6) yields a negative expectation for the first derivative \( w(t) \) of \( W(t) \).

Now, another key property of the divergence can be used: its convexity [11]. Experiments show that Basseville's test has a convex shape inside the pitch interval and jumps down at the end of this interval. This can be justified by some mathematical considerations on the adaptation process [12]. Let us give a qualitative interpretation of this piecewise convexity of \( w(t) \). Voiceless speech is the succession of a strong glottal excitation (typically a pulse) and a weaker stable one (comparable to white noise). This pulse enters the memory of both models at the same time, say \( t_0 \). The short-term model \( \theta(t) \) has a sliding window of small length \( \eta \) (e.g. 2 ms), then at time \( t_0 \) the pulse exists in the memory of \( \theta(t) \) but stays in the memory of \( \theta'(t) \) with some forgetting factor (e.g. \( \lambda = 0.9 \)). At this exit time \( \theta(t) \) switches back to the description of a bounded stationary signal, whereas \( \theta'(t) \) still takes the pulse into account: the models strongly disagree, and the divergence test \( W(t) \) jumps down. Now as time goes by the pulse is exponentially damped in \( \theta(t) \)'s memory, and therefore the slope \( w(t) \) measuring the disagreement decreases in absolute value, yielding this piecewise convex shape of \( W(t) \) we observe inside voiced speech.

V. Comparison on real speech data

Both methods have been assessed on large speech database: 16 kHz high-quality linear phase recording for the (GML) method, with an application to speech synthesis [8], and a phonetically balanced set of 8 kHz clean and noisy sentences, for (CVX) with an application to speech coding [12].

The behavior of the algorithms are illustrated in Figs.1-3. Figs. 1 and 2 correspond to speech uttered by a low-pitch male speaker. Fig.1 shows a transition between two oral vowels [a]-[a], where 2 a transition between a nasal consonant and a back vowel [u]-[u]. Figs.3 correspond to the transition [a]-[a]-[a] uttered by a female speaker. The signal has been recorded in an anechoic room with a nearly linear-phase microphone, digitized at 44.1 kHz, filtered by a linear phase FIR filter, and decimated to 16 kHz for the GML processing, and again lowpass-filtered and decimated to 8 kHz for the CVX computations.

The first derivative of the log-likelihood (GML) leads most of the time to the clear detection of the glottal closure, marked by pitch synchronous negative peaks. It is interesting to notice that the GML performs equally well on the transition [u]-[a]-[a] than on [o]-[ka]-[u] (whereas the Wong's criterion fails to detect any glottal event on this signal as can be seen in Fig.2). The interpretation of the positive peaks is more difficult. They are apparently correlated to the instant of the maximum glottal opening, where the acoustical coupling between sub-glottal and supra-glottal cavities reaches its maximum.

The CVX test curves plotted in logarithmic amplitude are the divergence test \( W(t) \), its first and second derivative \( w(t) = \dot{W}(t) \), and \( w'(t) = \ddot{W}(t) \). The original speech signal is shown on top. The convexity of \( W(t) \) as a function of time is observed through \( \dot{W}(t) \), which is clearly jumping at the ends of the pitch interval (glottal closure). The shape is very precise for [a]-[ka]-[a], whereas the detection is less accurate for [u]-[ka]-[u] from the second derivative, whereas it is still clear from the first derivative \( w(t) \). A wider range of speech data has been segmented in [4] with the same convexity detection method reinforced by a simple time-domain pitch detection procedure.

VI. Conclusion

Two new procedures for the detection of glottal closure have been described and compared on real speech. Both methods are based on a sequential likelihood comparison between two models of the speech signal. They seem to perform better than previous methods. The computational load required by the Generalized Maximum Likelihood (GML) is much larger than for the Convexity Jump (CVX) method, which just needs two simple filtering operations to get \( w(t) \) and can therefore be implemented in real time on a DSP. Now the results given by the GML algorithm are very accurate. For off-line applications where the pitch period value is available from another method, the GML should be preferred, whereas for real time processing the CVX method is more adequate.
From top:
- Woon's normalized residual error criterion (sampling frequency 16 kHz, model order 10, preemphasis 0.95, analysis frame 46 samples, stabilized covariance)
- First derivative of the log-likelihood ratio (GML)
- Signal waveform (diphone /a/)
- Divergence test
- First logarithmic derivative of the divergence
- Second logarithmic derivative of the divergence

Fig. 2: idem Fig. 1 for diphone /m/.

Fig. 3: Logatom /a''-y-R' signal
- divergence
- first derivative
- second derivative (of divergence)

References