A Maximum Likelihood Approach to the Detection of Moments of Maximum Excitation and its Application to High-Quality Speech Parameterization

Ranniery Maia†, Yannis Stylianou†, Masami Akamine‡
†Toshiba Research Europe Limited, Cambridge Research Laboratory, Cambridge, UK
‡Toshiba Corporation, Corporate Research and Development Center, Kawasaki, Japan
ranniery.maia@crl.toshiba.co.uk

Abstract
This paper presents an algorithm to detect moments of maximum excitation (MME) in speech. It assumes a model in which speech can be represented as a sequence of pulses located at the MME convolved with a time-varying minimum-phase impulse response. By considering that in the glottal cycle speech concentrates more energy at the MME than at other instants, the locations and amplitudes of the excitation pulses are determined through maximum likelihood estimation. The suggested approach provides a fully automatic and consistent method for the detection of MME in speech without relying on ad hoc procedures which usually do not work well across different speech styles without a required amount of adjustments. Experiments with speech parameterization, in the context of complex cepstrum analysis and synthesis, have shown that the proposed MME-based processing can improve signal to error reconstruction ratio up to 10%, when compared to the use of glottal closure instant estimations provided by a well-known algorithm.

Index Terms: speech analysis, pitch marking, epoch detection, speech modeling, speech parameterization

1. Introduction
Epoch extraction from speech is an important process in which the goal is to isolate or identify glottal cycles of human speech. This sort of information is important for many speech applications, such as speech synthesis and coding. The identification of the glottal cycles can also be used in methods to identify speech pathologies. Epochs usually indicate moments of maximum excitation (MME) in speech, and in some cases are closely related to glottal closure instants (GCIs), which are moments where the glottis is fully closed and air pressure is built-up in the lungs to be released for the next cycle of voiced speech. Although different terms such as GCI, pitch marks, pitch period onset times, or instants of significant excitation can be used by many authors, generally these entities tend to represent the same variables in the context of epoch extraction. In this work, the term MME is used. Usually, if speech is analyzed at the MME then a robust estimation of the factors that compose the source-filter model of speech production, such as vocal tract and glottal excitation amplitude and phase responses, can be obtained [1]. The separation of the different factors of the speech production mechanism is an essential step, for instance, in order to achieve good quality in statistical parametric synthesizers [2].

The literature shows a vast number of publications on MME detection, e.g. [3–11]. Basically, the methods can be divided into families regarding which information is input, such as speech or linear prediction residual, and in which domain they process such information, for instance time-domain through the autocorrelation or frequency domain through the group delay functions. The diversity of existing algorithms for MME detection has motivated some researchers to contribute with some papers whose purpose have been focused on evaluating some of these methods under certain conditions such as noisy speech [9], or different voice styles [12, 13]. These evaluations are usually conducted in quantitative terms, where a ground truth is assumed. For that, usually MME extracted from Electroglossograph (EGG) are regarded as the real MME and measures including: identification accuracy, false alarm rate, mean squared error between the positions of real and estimated MME, and missing rate, are among the measures utilized [9, 12, 13]. A common aspect among most of these approaches is that the ad hoc procedures they include perhaps explains how their performances differ considerably in each condition and speech style. This paper presents a method to detect MME in voiced speech. It assumes that speech can be represented as the convolution of an excitation signal represented as a pulse train, with pulses located at the MME, convolved with a time-varying minimum-phase impulse response. By assuming such model, each pulse location is then determined by maximum likelihood estimation (MLE) of the positions and amplitudes of the pulses, given natural speech and minimum-phase impulse responses. The algorithm works under the assumption that around the MME speech concentrates more energy, and consequently have more quantitative impact on the mean squared error (MSE) between natural and speech reconstructed by the assumed model. Although the method is not specifically based on the minimum MSE (MMSE) criterion, it can be demonstrated that the MMSE is a special case of the proposed maximum likelihood criterion when the noise added to the model is Gaussian with zero mean and variance one. Therefore, the MSE between natural and reconstructed speech is incorporated in the log likelihood function. Based on the log likelihood cost as a function of the excitation pulses, a six-step procedure is proposed in order to derive MME from speech. The advantages of the proposed method is that is fully automatic, requiring a minimum amount of ad hoc procedure. The algorithm uses the following information as input: speech, \( F_0 \), cepstrum and impulse response orders, and assumption regarding the noise of the model. \( F_0 \) information is only used in the last steps of the procedure to remove MME indications from the unvoiced regions, and to detect half pitch problems together with the log likelihood function.

This paper is organized as follows. Section 2 presents the assumed speech model and derives the log likelihood function dependent on the MME; Section 3 describes the six-step procedure utilize to detect MME from speech using MLE; some experiments are shown in Section 4; and the conclusions are in Section 5.
2. Maximum likelihood approach to detect MME

2.1. Assumed speech model

Here we assume that speech, $s(n)$, can be represented as the convolution of an excitation signal, $e(n)$, and a time-varying minimum-phase impulse response, $h(n)$, added by a noise component $w(n)$,

$$s(n) = e(n) * h(n) + w(n).$$  \hspace{1cm} (1)

Figure 1 illustrates the assumed model. Using matrix notation, (1) can be represented as

$$s = He + w,$$  \hspace{1cm} (2)

where

$$s = \begin{bmatrix} s(0) & \cdots & s(N-1) & \cdots & 0 \end{bmatrix}^T,$$  \hspace{1cm} (3)

$$H = \begin{bmatrix} g_0 & \cdots & g_{N-1} & \cdots & 0 \end{bmatrix},$$  \hspace{1cm} (4)

$$g_n = \begin{bmatrix} 0 & \cdots & 0 & h_n^T & 0 & \cdots & 0 \end{bmatrix}^T,$$  \hspace{1cm} (5)

$$h_n = \begin{bmatrix} h_n(0) & \cdots & h_n(M) \end{bmatrix}^T,$$  \hspace{1cm} (6)

$$e = \begin{bmatrix} e(0) & \cdots & e(N-1) \end{bmatrix}^T,$$  \hspace{1cm} (7)

$$w = \begin{bmatrix} w(0) & \cdots & w(N-1+M) \end{bmatrix}^T,$$  \hspace{1cm} (8)

with $N$ and $M$ being respectively the number of samples of $s(n)$ and impulse response order of $h(n)$. $h_n$ contains the impulse response of $H(z)$ at the $n$-th sample position.

2.2. Likelihood of speech given excitation

If $w$ is assumed to be a normally distributed error with mean 0 and variance $R$, then the likelihood of the speech vector, $s$, given the excitation vector, $e$, and impulse response matrix, $H$, and error covariance, $R$, becomes

$$p(s \mid H, e, R) = \frac{1}{\sqrt{(2\pi)^N|R|}} e^{-\frac{1}{2}[s-He]^T R^{-1}[s-He]}.$$  \hspace{1cm} (9)

Considering that $H$ and $R$ are given, then vector $e$ is the sole responsible for the likelihood function in (9). By assuming that $e$ has only $Z$ non-zero samples, as illustrated in Figure 1, then the product $He$ can be written as

$$He = \sum_{z=0}^{Z-1} a_z g_{pz},$$  \hspace{1cm} (10)

where $\{a_0, \ldots, a_{Z-1}\}$ and $\{p_0, \ldots, p_{Z-1}\}$ are respectively the amplitudes and positions of the $Z$ non-zero samples of $e(n)$. By substituting (10) into (9), and taking the logarithm, we obtain a log likelihood function in terms of the excitation signal

$$\mathcal{L}(\lambda) = -\frac{N}{2} \ln (2\pi) - \frac{1}{2} \ln |R| - \left[s - \sum_{z=0}^{Z-1} a_z g_{pz}\right]^T R^{-1} \left[s - \sum_{z=0}^{Z-1} a_z g_{pz}\right],$$  \hspace{1cm} (11)

where $\lambda = \{p_0, a_0, \ldots, p_{Z-1}, a_{Z-1}\}$.

2.3. Maximum likelihood estimation of $\lambda$

If we assume that $R$ is not dependent on $a_z$, the $z$-th pulse amplitude $\hat{a}_z$ which minimizes (11) can be found by making $\partial \mathcal{L} / \partial a_z = 0$, which results in

$$\hat{a}_z = \frac{g_{pz}^T R^{-1} g_{pz}}{g_{pz}^T R^{-1} s - \sum_{z=0}^{Z-1} a_z g_{pz}}.$$  \hspace{1cm} (12)

By substituting (12) into (11), and considering the terms which depend only on the $z$-th pulse, the best position $\hat{pz}$ can be derived as

$$\hat{pz} = \arg \max_{p_z \in \mathcal{P}} \left[ g_{pz}^T R^{-1} \left(s - \sum_{z=0}^{Z-1} a_z g_{pz}\right) \right]^2.$$  \hspace{1cm} (13)

where $\mathcal{P}$ is a codebook of samples in which the best position $\hat{pz}$ must be retrieved from. It can be noticed that equations (12) and (13) resemble the form to determine an optimal excitation in a family of speech coders known as multipulse excited linear prediction [14]. In fact, (12) and (13) indicate that the pulse amplitudes and locations are being determined through the minimization of the MSE between natural, $s(n)$, and reconstructed speech signal, $\hat{s}(n) = h(n) * e(n)$. The inverse of the covariance matrix, $R^{-1}$, can be interpreted in this case as the effect of an error weighting filter, which is usually used in the family of speech coders based on the analysis-by-synthesis principle.

3. Procedure to determine MME based on ML of $\lambda$

Based on the ML framework as presented in Section 2 we can develop a procedure to determine the MME, which is illustrated in Figure 2. In the following, the main steps performed in Figure 2 are explained with details.

3.1. Short-term MME detection

In the short-term MME detection, the speech signal is analyzed over $T$ short frames shifts. For each frame $t = \{0, \ldots, T-1\}$, the universe of pulse positions $\mathcal{P}$ is set to $\mathcal{P} = \{tK, \ldots, t(K+1) - 1\}$, where $K$ is the number of samples per frame. After defining the frame size, or in other words after defining the universe of locations $\mathcal{P}$, for each position $p_z \in \mathcal{P}$ an amplitude is obtained as

$$a_n = \frac{g_{nz}^T R^{-1} s}{g_{nz}^T R^{-1} g_n}, \quad n \in \mathcal{P}.$$  \hspace{1cm} (14)
After that, a cost function related to each position in $\mathcal{P}$ is obtained as
\[
F(n) = \begin{cases} 
\frac{(g_\mathcal{P}^n R^{-1}s)^2}{g_\mathcal{P}^n g_{\mathcal{P}_n}}, & a_n > 0, \\
0, & a_n \leq 0.
\end{cases}
\] (15)

Finally, the position, if it exists, is given by
\[
\hat{p}_z = \arg\max_{p_z \in \mathcal{P} | a_z > 0} F(p_z).
\] (16)

### 3.2. Pulse elimination based on minimum distance

The pulse elimination step utilizes a minimum distance $D = 2K + 1$, where $K$ is the number of samples per frame, which is below the range of human pitch periods. It makes use of the short-term analysis to eliminate unlike MMEs based on the cost function for each position, $F(p_z)$, given by (15). This is done by Algorithm 1, which simply checks if $p_{z+1} - p_z \leq D$ and eliminates the pulse which has smaller $F(\cdot)$.

**Algorithm 1 Short-term pulse elimination.**

**Require:** $D$: minimum distance

1: $z \leftarrow 0$
2: while $z < N - 1$ do
3: if $p_{z+1} - p_z < D$ then
4: $Z \leftarrow Z - 1$
5: if $F(p_z) < F(p_{z+1})$ then
6: Remove $p_z$
7: else
8: Remove $p_{z+1}$
9: $z \leftarrow z + 1$
10: end if
11: else
12: $z \leftarrow z + 1$
13: end if
14: end while

### 3.3. Position refinement

The position refinement step takes as input some initial locations and optimizes them so as to maximize the likelihood of speech given the assumed model. Effectively, the same procedure employed in the pulse detection stage is used here with two differences: (1) the universe of samples $\mathcal{P}$ is set around the neighborhood of each initial pitch mark; (2) the procedure is carried on a pulse location basis rather than frame by frame.

Therefore, for a given input set of positions $\{p_0, \ldots, p_{Z-1}\}$, where $Z$ is the number of pulses, the updated position is given by
\[
F(p_z) = \frac{(g_\mathcal{P}^n R^{-1}s)^2}{g_\mathcal{P}^n g_{\mathcal{P}_n}},
\] (17)
\[
\hat{p}_z = \arg\max_{p_z \in \mathcal{P}} F(p_z),
\] (18)

where $\mathcal{P}$ is a range of samples in the neighborhood of the initial $p_z$. Among others, one possibility, for instance, is to set $\mathcal{P} = \left\{\frac{p_z + p_{z-1}}{2}, \ldots, \frac{p_z + p_{z+2}}{2}\right\}$. After refining the positions, the corresponding amplitudes are calculated by
\[
a_z = \frac{g_\mathcal{P}^n R^{-1}s}{g_\mathcal{P}^n g_{\mathcal{P}_n}}.
\] (19)

### 3.4. $F_0$-based post processing

In the last step of MME detection procedure, $F_0$ information is used to: (1) remove remaining pulses in the unvoiced region; (2) detect half pitch problems. The first procedure is done by simply removing pulses $p_z$ so that pitch at position $p_z$, $F(p_z)$, is zero. The local pitch is calculated through the mean across $2L + 1$ frames, where $L$ is even and the center frame is the one in which $p_z$ belongs to.

For the detection of half pitch problem, $F_0(z)$ is also employed together with $F_0$ information. This task is performed by Algorithm 2. Basically what is being done in this part is: (1) check whether the distance $p_{z+1} - p_z$ is consistent with the corresponding local pitch at position $p_z$, $F(p_z)$; (2) if not, remove either $p_z$ or $p_{z+1}$ depending whether $F_0(p_{z+1}) > F_0(p_z)$ or $F_0(p_{z+1}) \leq F_0(p_z)$.

**Algorithm 2 Detect half pitch problems.**

**Require:** $K$: number of samples in 1 ms

**Require:** $\{P(p_0), \ldots, P(p_{Z-1})\}$: pitch at every pulse position

1: while $z < Z - 1$ do
2: if $p_{z+1} - p_z \leq 0.625P(p_z)$ and $F_0(p_{z+1}) \leq F_0(p_z)$ then
3: Remove $p_{z+1}$
4: $z \leftarrow z + 1$
5: else if $p_{z+1} - p_z \leq 0.5P(p_z)$ and $F_0(p_{z+1}) > F_0(p_z)$ then
6: Remove $p_z$
7: $z \leftarrow z + 1$
8: else
9: $z \leftarrow z + 1$
10: end if
11: end while

Figure 3 shows a speech segment and corresponding MME output by each one of the steps shown in Figure 2, where it can be clearly seen the effect of each one the procedures described in this section. The fact of starting with the short-frame MME detection prevents from missing MME, what according to experiments with complex cepstrum analysis and synthesis, is the most crucial problem.
Figure 3: Speech waveform and corresponding MMEs, as the results of each step shown in 2. Top: the outcome of the short-term MME detection; second row: after applying Algorithm 1; third row: outcome of the pulse refinement; fourth row: another run of Algorithm 1; fifth row: after removing MME from the unvoiced regions; sixth row: final MME, after half pitch problem detection.

4. Experiments

4.1. Speech material

To evaluate the proposed algorithm, 200 utterances from four different speakers, two male and two female, 50 sentences from each speaker, were randomly selected from a multi-speaker corpora. Speakers F1 (American English female), F2 (UK English female), and M1 (US English male) are proprietary databases that were recorded at studio with high quality for the purpose of designing TTS systems. Speaker M2 (UK English) corresponds to the clean part of the data released to the Hurricane Challenge [15]. The sampling frequency was 16 kHz.

4.2. Evaluation condition

Although measures like identification accuracy, mean squared error, miss rate and false alarm rate are generally used in the assessment of MME detection methods with an assumed ground truth [9, 12, 13], here we evaluated the performance of the proposed algorithm in terms of practical application of the final MME: speech analysis and reconstruction using complex cepstrum-based speech representation [16]. The choice for complex cepstrum-based analysis and synthesis was made based on the fact that this type of parameter is highly sensitive to the locations of the analysis instants. In addition, the complex cepstrum not only represents amplitude but also phase information of speech. If the provided MMEs are accurate, speech analysis and synthesis based on the complex cepstrum should be able to reconstruct close-to-natural speech. In this work we utilized the complex cepstrum analysis method proposed by [16] because factors such as choice of analysis window and phase unwrapping are greatly relaxed by this approach. In addition, because the goal of complex cepstrum-based analysis and synthesis is to reconstruct the original time-domain speech waveform, the segmented signal-to-noise ratio of voiced regions, (SNR-seg) was chosen as measure of performance.

The MME detected by the proposed algorithm were compared with the MME extracted by two other well-known algorithms: (1) the residual-based normalized cross-correlation method (ESPS) [4]; and (2) the Dynamic Programming Projected Phase-Slope Algorithm (DYPSA) [7]. Although ESPS has been proposed in the end 1980s, the method has still been used in some recent papers for evaluation [12]. ESPS is part of Festvox and DYPSA is released as a tool in voicebox, a speech processing toolkit for MATLAB.

The original sentences were used to detect MME using the proposed method, ESPS and DYPSA approaches. For the proposed method, minimum-phase cepstra were calculated at every 1 ms with cepstrum order $C = 24$. During MME detection, minimum-phase impulse response order was set to $P = 256$. The error covariance matrix, $R$, was implemented as a diagonal matrix, where each term was given by $R(n) = j^2(n)$, $n = 0, \ldots, P$, where $\{j(0), \ldots, j(P)\}$ is a right-half Blackman window so that $j(0) = 1$ and $j(P) = 0$. In fact, choice of $R$, which can be seen as a weighting filter, and $P$ and $C$ are the only variables to be set in the proposed algorithm. The MME from the three approaches were fed into the MSE complex cepstrum analysis. In all the cases, speech was reconstructed by converting the frame-based complex cepstra into corresponding non-causal impulse response, while the MME were used to create the excitation signal. Speech was then synthesized by convolving the filter impulse response with the excitation signal.

4.3. Results

Table 1 shows SNR-seg taken as the average from the speech material used in the test. It can be seen that the proposed method achieves better results for F1 and M2, while DYPSA is better for F2 and M1. However, overall the quality of the samples was similar. The advantage of the proposed method is its simplicity and consistency across different speakers.

<table>
<thead>
<tr>
<th></th>
<th>F1</th>
<th>F2</th>
<th>M1</th>
<th>M2</th>
</tr>
</thead>
<tbody>
<tr>
<td>ESPS</td>
<td>15.90</td>
<td>18.29</td>
<td>15.45</td>
<td>14.31</td>
</tr>
<tr>
<td>DYPSA</td>
<td>16.36</td>
<td>18.35</td>
<td>17.30</td>
<td>15.40</td>
</tr>
<tr>
<td>Proposed</td>
<td>16.70</td>
<td>17.85</td>
<td>16.50</td>
<td>15.53</td>
</tr>
</tbody>
</table>

5. Conclusions

In this paper we have proposed an approach for the detection of moments of maximum excitation in speech. The algorithm assumes a model in which speech is estimated by the convolution of an excitation signal composed of pulses located at the MME, convolved with a time-varying minimum-phase impulse response. The algorithm then obtain the locations and amplitudes of the excitation pulse through maximum likelihood estimation. Experiments in a complex cepstrum analysis-synthesis framework have shown that the proposed method is in average better than ESPS, and similar to DYPSA in terms of SNR-seg of the reconstructed speech. The advantage of the proposed method is that is simple, consistent, and fully automatic. In the future we intend to perform a quantitative evaluation of the proposed method.
6. References


