GLOTTAL CLOSURE INSTANT ESTIMATION USING AN APPROPRIATENESS MEASURE OF THE SOURCE AND CONTINUITY CONSTRAINTS

Damien Vincent, Olivier Rosec

France Telecom
R&D Division
2, Avenue Pierre Marzin - 22307 Lannion
{damien.vincent, olivier.rosec}@francetelecom.com

Thierry Chonavel

ENST Bretagne
Signal & Communication Department
CS 83818 - 29238 Brest Cedex 3
thierry.chonavel@enst-bretagne.fr

ABSTRACT

An algorithm for GCI (Glottal Closure Instants) estimation is presented in this paper. It relies on a source filter model of speech production using a LF model for the source component. From this source filter decomposition, a ratio which measures the goodness of fit of the LF source model is introduced in the GCI estimation procedure together with fundamental frequency constraints. Then a Viterbi algorithm is applied to extract the most likely GCI sequence. Experiments performed on a real speech database show the relevance of the proposed method.

1. INTRODUCTION

Determination of glottal closure instants (GCI) is a problem that has received considerable attention for many years. Such instants are considered as important for speech analysis as their precise location is mandatory for glottal source estimation as well as for voice quality characterisation. Another direct field of application is pitch marking of speech databases for use in speech synthesis. Indeed, certain speech synthesis algorithms (eg TD-PSOLA [1]) need precise estimation of GCI for prosodic modification or speech segment concatenation. Considering the determination of GCI, several methods have been proposed using group delay functions [2] which rely on basic properties of minimum phase signals, or dynamic programming in order to get a GCI sequence in accordance with a given f_0 input sequence [3].
However existing approaches do not take into account the structure of the glottal source signal. In this paper, a measure obtained from a source-filter decomposition is used to locate potential GCI instants and thus to better constrain the GCI estimation problem. This information is combined with a previously estimated f_0 sequence by means of a dynamic programming algorithm. The paper is organised as follows. Section 2 gives an overview of the pitch determination algorithm which can be seen as an improved Yin algoritm [4] by the introduction of a f_0 tracking mechanism. Section 3 details the GCI estimation task while section 4 describes the obtained results.

2. F0 ESTIMATION PROCEDURE

2.1. The algorithm

We present an improvement of the YIN method [4] to estimate the fundamental frequency of speech, based on the introduction of continuity constraints through dynamic programming. Given a speech signal $s(n)$, the YIN algorithm estimates the fundamental frequency $f_0(n)$ using the following cumulative mean normalized difference function (refered to as CMNDF thereafter):

$$d'_n(\tau) = \begin{cases} 1 \text{ if } \tau = 0, \\ \frac{d_n(\tau)}{\sum_{k=1}^{\tau} d_n(k)} \text{ otherwise,} \end{cases}$$

where $d_n(\tau) = \sum_{k=-K}^{K} \left(s(n+k) - s(n+k-\tau)\right)^2$ is the difference function over a $2K+1$ length analysis window. The use of the CMNDF instead of a correlation or a difference function was shown to provide better performance on average. However, the global minimum of CMNDF does not always correspond to the correct fundamental frequency and so further post-processing steps are needed to make the estimation more robust. The first step alleviates the problem of subharmonic errors: among the set of minima of d' below a threshold δ (set to 0.1 in [4]), the estimated fundamental frequency is set to the highest frequency of these minima (if any). The second step enables to apply the f_0 values obtained in more reliable areas to areas of lower reliability. This processing tends to increase the continuity of the fundamental frequency but only locally.

In this paper, we propose the introduction of explicit constraints to globally improve this continuity. This tracking mechanism is introduced by means of dynamic programming. For a frame with index l, a target cost $C_{\text{targ}}(l, \tau)$ is defined based on the CMNDF and on the first post-processing step of the original Yin algorithm. To favour the f_0 values with high reliability in the target cost, if the minimum of CMNDF

is below δ then $C_{\text{targ}}(l,\tau)$ is set to infinity everywhere but in the vicinity of this minimum; otherwise, $C_{\text{targ}}(l,\tau)$ is set to the CMNDF. Moreover a transition cost $C_{\text{trans}}(f_0^{\text{prv}}, f_0^{\text{cur}})$ is also proposed to measure the continuity of two successive f_0 values. This transition cost is defined by

$$C_{\text{trans}}(f_0^{\text{prv}}, f_0^{\text{cur}}) = \begin{cases} 0 & \text{if } c \leq c_1 \\ \frac{c-c_1}{c_2-c_1} & \text{if } c_1 < c < c_2, \\ 1 & \text{if } c \geq c_2 \end{cases}$$

where $c = \frac{|f_0^{\text{cur}} - f_0^{\text{prv}}|}{0.5(f_0^{\text{cur}} + f_0^{\text{prv}})}$ is the normalized local f_0 variation. In the experiment performed in this paper, $c1$ and $c2$ are respectively set to 0.15 and 0.25, which means that no penalty is applied if the variation is less than 15%, while the maximum penalty is attributed when this variation is more than 25%. As the analysis is done every 20ms, the algorithm allows to double ($1.15^{\frac{100}{20}} \approx 2$) the pitch every 100ms without penalty: this delay corresponds to a physical limit of most human voices to change the fundamental frequency by one octave.

The CMNDF is a normalized function and is therefore independent of the signal amplitude. Thus it gives as much importance to a strict voiced part as to a weak spurious sinusoidal component added to a silence zone. In order to make the estimation more robust, high energy parts of the signal should be given a higher weight in the pitch estimation process. This is done by adding a factor $\alpha_E(l)$ to the concatenation cost $C_{\text{trans}}(f_0^{\text{prv}}, f_0^{\text{cur}})$ and to the target cost. From the previous considerations, we propose the following expression for $\alpha_E(l)$:

$$\alpha_E(l) = \beta^{\frac{1}{3} 10 \log_{10}\left(\frac{E(t_l)}{E_{\text{mean}}(t_l)}\right)} \quad ,$$

where $E(t_l)$ and $E_{\text{mean}}(t_l)$ are respectively the signal powers estimated using an analysis window respectively of length 25ms and 500ms centered on t_l. Every 3dB, the factor $\alpha_E(l)$ is multiplied by β: this constant has been set to 1.15 in the experiments.

2.2. F0 estimation performance

The evaluation was carried out on the *Arctic* database [5] which provides both the speech signals and the EGG signals (ElectroGlottoGraphic). The EGG derivative signal (DEGG) allows to easily extract the GCIs as these correpond to clear peaks in the DEGG signal. The instantaneous reference fundamental period is therefore defined as the difference between two successive GCIs. However, the intantaneous f_0 is not suitable for an irregularly voiced signal since it cannot be linked to the perceived pitch. We prefer to use a mean fundamental period which is a good tradeoff between getting a correct pitch value for voiced pitch and getting a meaningful value for irregularly voiced signals: denoting the reference glottal closure instants by $t_c(l)$, the mean fundamental period is defined by $T_0(l) = \frac{t_c(l+2) - t_c(l-2)}{4}$.

To get the reference GCIs over the entire database, an automatic peak picking algorithm is required. The algorithm is iterative. For each iteration, a GCI is picked according to its confidence level: the confidence is defined by the amplitude of the DEGG peak. Then, once a GCI has been selected, an exclusion zone is defined so as to prevent picking another GCI in this zone during the following iterations. The length of the exclusion zone is defined using an a priori pitch period. It should be noted that the peak picking algorithm is robust regarding errors on this a priori pitch period.

Finally, the estimation results are given in terms of a gross error rate: the percentage of $\hat{f}_0$ which deviates from f_0 by more than 20%. Two kinds of results are provided: the first evaluation is done over all reference fundamental frequencies, the second one ignores the irregularly voiced parts of the signal from the results. A GCI is defined as irregular if the left instantaneous pitch period differs from the right pitch period by more than 20% or if the left or right pitch period differs from the mean pitch period $T_0(l)$ by more than 20%. The results are given in table 1 for both YIN method and the present algorithm. Tests using both configurations show a significant gross error rate decrease by adding the proposed tracking mechanism to the YIN method.

Algorithm	Error rate (1)	Error rate (2)
YIN	3.19%	1.78%
Proposed (YIN+Viterbi)	1.68%	0.66%

Table 1. Gross error rates on the Arctic database using the YIN method and the proposed algorithm. Gross error rate (1) includes irregularly voiced parts while error rate (2) ignores them.

3. GCI ESTIMATION PROCESS

3.1. The algorithm

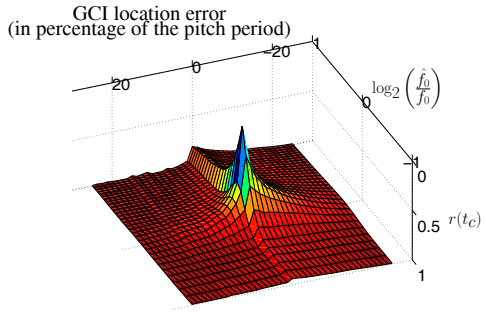

Fig. 1. Ratio $r(t_c)$ with respect to the fundamental frequency and the difference with the true GCI.

The GCI estimation process proposed in this paper takes into account the information obtained from a source-filter in-

version procedure described in [6] which is based on the Liljencrants-Fant model (LF) [7]. For that purpose, given a candidate GCI t_c, a measure is defined by $r(t_c) = \min_u \frac{E_u(t_c)}{E_0(t_c)}$ which is the ratio of the prediction error using the best LF glottal source to the LPC prediction error. To illustrate the usefulness of this measure, figure 1 depicts the variation of this measure for a synthetic signal generated with the LF model with a constant fundamental frequency of 100 Hz. Interestingly, for a given f_0 value, $r(t_c)$ is minimal at the true t_c instant, while the minimum over f_0 is obtained at 100 Hz. Thus, this measure seems to be a good indicator for the location of the GCI.

In order to track the variation of the fundamental frequency, continuity constraints need to be introduced in the GCI estimation algorithm so that the distance between two consecutive GCIs is close to the fundamental period. This leads to the introduction of a composite cost function comprising: i) a target cost $C_{\text{LF}}(l, t_c)$ given by the $r(t_c)$ measure and ii) a transition cost which favours GCI sequences in accordance with the f_0 information. On one hand, the desirable behaviour of the concatenation cost $C_{\text{concat}}(l, t_c^{\text{cur}}, t_c^{\text{prv}})$ is to penalize the delay $\Delta t_c(l) = t_c^{\text{cur}} - t_c^{\text{prv}}$ between two consecutive GCIs too far from the pitch period. On the other hand, the penalty must be weighted according to the confidence in the estimated pitch period so that for instance, in irregularly voiced parts of the signal, the concatenation cost is not too constraining. The concatenation cost $C^1_{\text{concat}}(l, t_c^{\text{cur}}, t_c^{\text{prv}}) = g\left(\frac{f_s}{t_c^{\text{cur}} - t_c^{\text{prv}}}\right)$ is thus represented in figure 2 where the minimal and maximal frequencies are defined by:

$$\begin{array}{l} \frac{f_0(t_c^{\text{cur}})}{f_0^{\min}} = \frac{f_0^{\max}}{f_0(t_c^{\text{cur}})} \\ \frac{f_0^{\max}}{f_0(t_c^{\text{cur}})} = \gamma \frac{d'_{t_c^{\text{cur}}}(T_0(t_c^{\text{cur}})) + \delta}{1+\delta} \\ d''_n(\tau) = \min(d'_n(\tau); 1) \end{array} .$$

The constant δ allows a small difference between the pitch period $\Delta t_c(l)$ and the estimated pitch period even if the CMNDF is close to zero. To make the concatenation cost more robust to pitch estimation errors, the cost is modulated to favour pitch periods $\Delta t_c(l)$ corresponding to low CMNDF values, which gives the following final concatenation cost:

$$\begin{array}{l} C_{\text{concat}}(l, t_c^{\text{cur}}, t_c^{\text{prv}}) = C^1_{\text{concat}}(l, t_c^{\text{cur}}, t_c^{\text{prv}})\ C^2_{\text{concat}}(l, t_c^{\text{cur}}, t_c^{\text{prv}}) \\ C^2_{\text{concat}}(l, t_c^{\text{cur}}, t_c^{\text{prv}}) = \min\left(d'_{t_c^{\text{cur}}}(\Delta t_c(l)) - \min_\tau d'_{t_c^{\text{cur}}}(\tau); 1\right) \end{array}$$

3.2. Implementation considerations

The first GCI is constrained to be in an interval $[t_1^s, t_2^s]$ by applying minor modifications to the previous target cost: the target cost is not modified on $[t_1^s, t_2^s]$ but is set to $+\infty$ outside. Constraining the first GCI is needed to avoid skipping several GCIs at the begining of the signal. The length L of the treillis, which corresponds to the number of GCIs, is not known a priori. Thus, we have to deal with the end of the algorithm: a new GCI is added to the treillis until the location of the last GCI of the optimal path lies in the interval $[t_1^e, t_2^e]$.

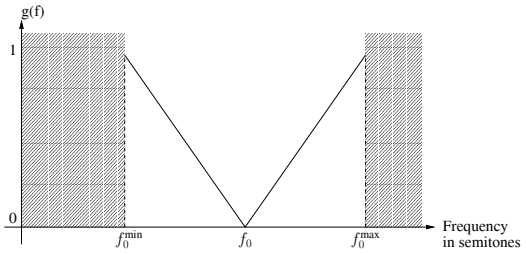

Fig. 2. Pitch deviation penalty function used in the concatenation cost.

The state space of the Viterbi algorithm is composed of $N = t_2^e - t_1^s$ samples. The complexity of each iteration of the algorithm is $O(N(T_0^{max} - T_0^{min}))$ which is much lower than the theoretical complexity $O(N^2)$, as many transitions are just forbidden. The complexity could be further reduced by working on each voiced part of the speech signal separately in order to decrease the number N of samples in the state space. The overall complexity is given by the estimation of $r(t_c)$ which is much more demanding than applying the present Viterbi algorithm.

4. RESULTS

The reference GCIs are selected as explained in section 2.2. In the same way, the test was done on the *Arctic* database [5] using two configurations: either all reference GCIs were included in the results or the irregularly voiced parts were discarded from the results. For a speech signal, the association between the reference GCIs and the estimated GCIs is done using an iterative algorithm: for each iteration, find the closest pair of reference GCI and estimated GCI; if the time delay between the reference and the estimate is below 5ms, add them to the list of associated GCIs, otherwise add the reference GCI to the missing set and the estimate to the false alarm set, then go to the next iteration. If there are more estimated GCIs than reference GCIs, the remaining GCIs are included in the false alarm set; if there are more reference GCIs, the remaining GCIs are included in the missing set. Let $t_{\text{ref}}(k)$ and $t_{\text{est}}(k)$ denote respectively the reference GCI and the associated estimated GCI; and $\Delta t(k) = t_{\text{est}}(k) - t_{\text{ref}}(k)$. The quality of the estimation is given by four measures: the standard deviation of the distribution Δt using only values below 2.5 ms, the gross error rate GER $= \frac{\text{card}\{k/\Delta t(k) > 2.5 \text{ ms}\}}{\text{Number of reference GCIs}}$, the missing rate MR $= \frac{\text{card}\{\text{Missing set}\}}{\text{Number of reference GCIs}}$ and the false alarm rate FAR $= \frac{\text{card}\{\text{False alarm set}\}}{\text{Number of reference GCIs}}$.

The results of the proposed algorithm have been compared to the method described in [3]. The algorithm [3] is not strictly speaking a GCI estimation algorithm but just a

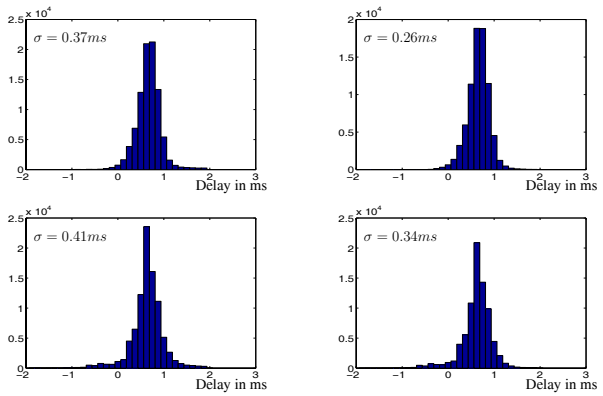

Fig. 3. Distribution of the difference between the estimated GCIs and the reference GCIs given by the DEGG signal. Top: proposed algorithm. Bottom: algorithm [3]. Left: using all reference GCIs. Right: discarding too irregular GCIs.

pitch marking algorithm, however this is not a drawback for the evaluation since the global bias of the difference distribution over the database is removed from the results. To make a fair evaluation, the same f0 estimate is used for both algorithms. The proposed algorithm achieves better performance: the estimation errors exhibit a smaller standard deviation (see figure 3) while table 2 shows a decrease for both the false alarm and missing rates. Figure 4 gives an example of the estimation process on a signal which exhibits both strongly periodic GCIs and some irregularly spaced GCIs in the middle. As a result, the proposed algorithm is able to estimate GCIs properly, even the irregularly spaced ones.

Test	Algorithm	GER	FAR	MR
Config1	Proposed algorithm	0.73%	1.04%	0.39%
	Algorithm [3]	0.97%	2.89%	0.86%
Config2	Proposed algorithm	0.15%	0.09%	0.16%
	Algorithm [3]	0.11%	2.30%	0.15%

Table 2. Gross error rate, false alarm and missing rates for both methods using either test configuration 1 (using all reference GCIs) or configuration 2 (excluding too irregular GCIs).

5. CONCLUSION

The proposed method enables to estimate the GCIs with good accuracy. This accuracy is not obtained at the expense of a false alarm or missing rate increase; these two rates remain low in the experiments. Concerning the applications, correct GCI estimation is often a required condition especially when estimating vocal quality parameters, as these parameters greatly depend on the location of GCIs. Some resynthesis experiments using the LF source model will then be performed in order to quantify the improvements resulting

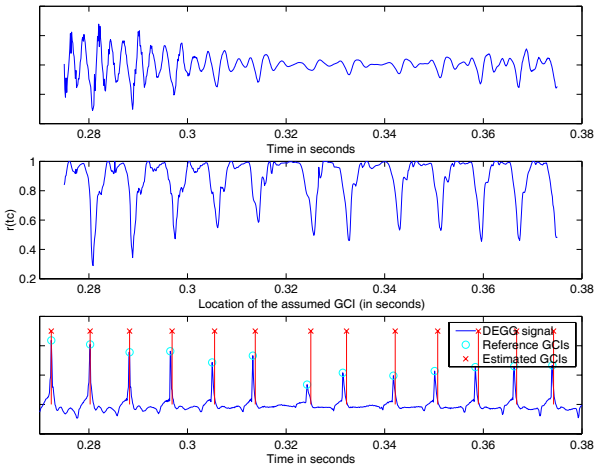

Fig. 4. Result of the estimation on the speech signal corresponding to the middle of the words '*It was*'. From top to bottom: (a) the speech signal, (b) the ratio $r(t_c)$, (c) the DEGG signal and the estimated GCIs.

from the proposed GCI selection algorithm. Another interesting application would be to pitch-mark the FTR&D speech database to check the effect of the proposed algorithm on speech synthesis quality: we expect better concatenations as pitch-marks are more consistent.

6. REFERENCES

[1] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Communication*, vol. 9, pp. 453–467, 1990.

[2] R. Smits and B. Yegnanarayana, "Determination of instants of significant excitation in speech using group delay function," *IEEE Trans. on Speech and Audio Processing*, vol. 3, no. 5, pp. 325–333, 1995.

[3] Yves Laprie and Vincent Colotte, "Automatic pitch marking for speech transformations via TD-PSOLA," *EUSIPCO*, pp. 1133–1136, 1998.

[4] A. de Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *J. Acoust. Soc. Am.*, vol. 111, no. 4, pp. 1917–1930, 2002.

[5] "Arctic speech database," http://festvox.org/cmu_arctic/.

[6] D. Vincent, O. Rosec, and T. Chonavel, "Estimation of LF glottal source parameters based on ARX model," *Interspeech*, pp. 333–336, 2005.

[7] G. Fant, J. Liljencrants, and Q. Lin, "A four-parameter model of glottal flow," *STL-QPSR*, vol. 4, pp. 1–13, 1985.