A COMPARISON OF SEVERAL RECENT METHODS OF FUNDAMENTAL FREQUENCY AND VOICING DECISION ESTIMATION

[1,2]*Eric Mousset*, [1]*William A. Ainsworth*, [2]*José A.R. Fonollosa*

[1]Department of Communication and Neuroscience, Keele University, Keele, Staffordshire ST5 5BG, U.K.
[2]Universitat Politècnica de Catalunya, E.T.S.E. Telecomunicació, Barcelona 08080, SPAIN

ABSTRACT

In this paper, we are interested in the comparison of several kinds of methods for fundamental frequency estimation and GCI (Glottal Closure Instant) detection. These methods operate in various domains (time-, frequency- or joint time-frequency domains). Their performances have been compared for both fundamental frequency estimation and voicing decision tasks as well as GCI detection, where applicable. This comparison was designed to be as unbiased as possible, so as to reflect the intrinsic properties of each method. A method based on a "Born-Jordan" kernel bilinear time-frequency representation of speech signals achieves the best performance in terms of GCI detection accuracy but is not as robust to inter-speaker variability as the SIFT algorithm. An auditory model, which has been applied on the same data in a previous study, has been shown to compare favourably to other methods (such as SIFT) in adverse noisy conditions only.

1. INTRODUCTION

The work presented here should be seen as a first step towards the preparation of an experimental framework devoted to systematic evaluation and comparison of F0 (fundamental frequency) estimation methods, including GCI (Glottal Closure Instant) detection methods. A common, low ambient noise, continuous speech database has already been recorded and labeled for this purpose [8]. This study consists of a comparison between several kinds of methods for GCI detection (a SIFT-based method, a Frobenius Norm based method and two bilinear time-frequency based methods) and for F0 estimation (a modulated gaussian wavelets based algorithm and an AMPEX based algorithm).

The performance of the various methods have been evaluated for both fundamental frequency estimation and voicing decision tasks as well as GCI detection, when applicable. This evaluation is not performed for only one configuration of parameters but for a range of values of the most influencial parameters. In other words, for each tested algorithm, such values have been varied between each database processing session. A brief presentation of the methods to be compared is given in section 2. Section 3 is devoted to the comparison framework itself and presents the database for evaluation and the evaluation criteria. Results are presented and discussed in section 4.

2. PRESENTATION OF THE COMPARED METHODS

2.1. GCI estimation methods

The methods as they are presented below do not actually produce series of GCIs, but a one-dimensional temporal signal the peaks of which are expected to indicate GCIs. This resulting signal can then be used in several ways, to extract GCIs or to locate the voiced frames for example. For each of the techniques presented below, the GCI detection process can be seen as a chain of four successive processing steps:

1. Acoustic speech signal pre-emphasis (optional).
2. Transformation aiming at producing peaks at GCIs.
3. Post-processing aiming at increasing contrasts in the resulting signal (optional).
4. Peak picking operation.

These four steps are from now on referred to by their item number. In this section, we focus mostly on step 2 and ignore step 4, which will be addressed below (see section 3.2).

A SIFT-based method. This method, proposed by Plante et al. [**?**], operates in the time domain. It consists of a SIFT based filtering of the speech signal extracting the so-called *residual signal*. Steps 1 to 3 can be arbitrarily sequenced the following way:

1. The signal is first passed to a pre-emphasis module improving the accuracy of the LPC (Linear Predictive Coding) analysis (performed on 25.6 ms asynchronous windows, overlapping by 12.8 ms).
2. The filter corresponding to the vocal tract is calculated from the LPC coefficients and the *residual signal* is obtained by inverse filtering. In order to increase the residual amplitude for voiced frames, the residual signal is weighted by the energy ratio between the original and the pre-emphasised versions.
3. In practice, the *residual signal* generally contains some noise corresponding to vocal tract characteristics. To remove some of this, the signal is clamped, low-passed filtered and its envelope is calculated using a Hilbert transform.

A Frobenius norm based method. This method was proposed by Ma et al. [**?**] as an alternative to more conventional methods based on LPC. It relies on the computation of the Frobenius norm of a matrix $M_{(m,p+1)}$, the rows of which are formed with sequences of speech signal samples, using a simple rectangular sliding window of length $p+1$ samples (shifted sample by sample between two successive rows). It has been demonstrated in [**?**] that the Frobenius norm of matrix M can be also expressed in terms of its $p+1$ singular values σ_i (assuming that $m \geq p+1$ and that m has full column rank, i.e. $p+1$). The following expression C (whose computation does not require any eigenvalue decomposition) is expected to produce peaks at GCIs:

$$C = \frac{1}{p+1}\|M\|_F^2 = \frac{1}{p+1}\sum_{i=1}^{m}\sum_{j=1}^{p+1} s_{ij}^2 = \frac{1}{p+1}\sum_{i=1}^{p+1}\sigma_i^2$$

It should be mentionned that steps 1 and 3 are absent in the original version of this method.

Two bilinear time-frequency representation based methods. In this study we only consider one bilinear TFR (Time-Frequency Representation) based method of epoch detection. The effect of using two different kernels for this method is investigated:

- the *Born-Jordan* kernel,
- the *cone-shaped* kernel (or *cone kernel*).

The principle of TFR based methods, reported by Navarro and Esquerra, is inspired by the work of Flandrin, who has proposed a non-parametric time-frequency formulation of a general class of receivers [**?**].

Given an observation $f(t)$ of a signal and $C_{ff}(t,f;\Psi)$ its Cohen's class TFR using kernel Ψ, an adaptation of Flandrin's receiver has been proposed in [**?**] in order to make it suitable for practical issues of GCI detection (where (T) is some short integration time interval):

$$(1) \quad \Lambda(T) = \int_{-\infty}^{\infty}\int_{(T)} C_{ff}(t-1,\omega;\Psi)C_{ff}(t,\omega;\Psi)dt\frac{d\omega}{2\pi}$$

According to the four step processing scheme described above, step 1 is ignored, step 2 is achieved by evaluating expression (1) for each signal sample and step 3 includes a contrast enhancement operation followed by a compression of the dynamic range (by application of a logarithm).

2.2. F0 estimation methods

A modulated gaussian wavelets based algorithm. The algorithm proposed by Janer is based on a family of 17 gaussian wavelets, whose mother wavelet dilation parameter has been tuned so that the whole family behaves like a Bark scale filter-bank [**?**]. The first step of this algorithm consists in picking peaks in each of the 17 bands. Then, for each single band, each new peak mark is either validated or rejected according to a criterion based on the time interval between consecutive marks.

The rest of the algorithm then relies on the following twofold general assumption. For each glottal cycle, at least one of the 17 detectors will always produce a mark and such marks will always fall in a common small time interval (with respect to the current glottal cycle length). All marks which are produced during the first phase are stacked. This operation results in a time series of clusters of marks ; the time interval between two consecutive clusters being expected to provide an estimate of the corresponding local glottal cycle length. In a last step, this series of marks is processed in order to select only one mark per cluster.

F0 estimation based on an auditory model. In a previous study, F0 estimation accuracy of an auditory model was evaluated using the same speech database as in the current study [**?**]. This model contains the key elements of the AMPEX algorithm but its last step consists in using the cochlear nucleus onset units. The latter selectively enhance pitch periodicities by summing cochlear nerve activity over wide frequency bands (7 barks) and performing a sort of peak picking operation.

3. COMPARISON SCHEME

3.1. Database for evaluation

We used a database created at Keele University [**?**], which aims at providing a common general framework for the evaluation of F0 and GCI estimation methods. It includes two kinds of signals: traditional acoustic speech signals and laryngograph signals (single speaker recording). Five adult female speakers and five adult male speakers were recorded in low ambient noise conditions using a sound-proof room. Each utterance consisted of the same phonetically balanced English text. In each case, the acoustic and laryngograph signals are time-synchronised (i.e. start and end at the same instants) and share the same sampling rate value of 20,000 Hz.

For GCI detection performance evaluation, we designed our own GCI reference label files, as advised in [**?**], using a technique which consists in estimating the GCIs by looking for the minima of the first derivative of the laryngogram signal [**?**]. As far as the evaluations related to V/U (Voiced/Unvoiced) decision and F0 estimation were concerned, we used the reference files provided in the Keele database, which contain a V/U decision and a pitch estimate for each 10ms block of speech. Segments where no consistent and obvious decision could be made by visual inspection are labeled as uncertain and ignored during the evaluation.

3.2. Pre- and post-processings for GCI production

The results presented in section 4 were obtained through two series of evaluations, the first of which involves the original GCI detection methods as they are described in section 2.1 above. This means that they can differ in step 1 (preprocessing), step 3 (post-processing before peak picking) and of course in step 2, embodying each method's peculiarities.

In the second series of evaluations, step 1 was added when absent. This has been achieved by equalising the energy level of the original

acoustic speech signal over time. Step 3 has been skipped for all the tested methods. In both series of evaluations, a common module of peak picking (step 4), based on morphological filtering has been applied (see [**?**] for more detail about this module).

Parameter setting. Like any other method, the ones described above are sensitive to parameter values such as analysis window lengths, thresholds, etc. We have chosen to freeze these parameters in the evaluation sessions. Only one of them has been tuned for each method and each uttered sentence, i.e. the eventual bias in GCI estimation (see next section).

3.3. Evaluation scheme

Correcting the bias in the GCI estimation. A first series of tests, the results of which are not reported here, has revealed biases in the GCI estimates which are speaker dependent [**?**]. If b_{min} denotes the optimal length of time interval, which, when added to each GCI estimate (for a given method) minimises the global error, then values of b_{min} measured for each speaker and each utterance suggest that this bias is stationary at the speaker level but may vary significantly from one speaker to another (and from one method to another). Then, before starting any further performance evaluation, we decided to correct this bias for each method, each speaker and each uttered sentence. In other words, what is evaluated hereafter is the ability to produce accurate detection of some event in the glottal cycle. The problem of how this event is related to the glottis' closure has been ignored.

Error types (GCI detection). We distinguished between the four following types of error (for a given RM R_j):

- *fine error* (*f.e.*), if a given AM A_i is such that the error $E_{i,j}$ (see below) is below some threshold (set to 0.1),
- *gross error* (*g.e.*), if $E_{i,j}$ is above this threshold,
- *non-detection* (*n.d.*), when there is no AM in $V(j)$,
- *false insertion* (*f.i.*) (or *redundant insertion*), when there are more than one AM in $V(j)$. We distinguish then between three sub-cases: all these AMs fall in the gross error case, some do and other ones don't, or none of them do (referred below as *case 1*, *case 2* and *case 3*, respectively).

$E_{i,j} = |e_{i,j}|$ can be seen as the *glottal cycle synchronised and normalised error*[1]:

$$e_{i,j} = \begin{cases} (A_i - R_j)/(R_{j+1} - R_j) & \text{if } A_i > R_j \\ (A_i - R_j)/(R_j - R_{j-1}) & \text{otherwise} \end{cases} \quad (2)$$

Error types (F0 estimation). Voiced/unvoiced decision and F0 estimation were simultaneously obtained from GCI detection methods by performing an autocorrelation of their resulting signal (output of step 3). Apart from the voicing error types *VU* (Voiced-to-Unvoiced) and *UV* (Unvoiced-to-Voiced), we only looked at the fine- and gross- error types, which again are defined according to the F0 reference value of the corresponding local 10ms frame (with a threshold set to 0.1), so as to be pitch independent:

$$E_n = |(F0_A(n) - F0_R(n))/F0_R(n)| \quad (3)$$

where $F0_A(n)$ (resp. $F0_R(n)$) is the algorithm (resp. reference) F0 estimate for frame of index n.

3.4. Algorithm operating characteristics

In the case of GCI detection, the morphological filtering based peak picking process is mostly sensitive to one of its parameters, namely the size of the structuring element. As a consequence, when the database is successively processed for different values of this parameter, different scores are obtained. Performances related to some criterion (e.g. non-detection error) improve whilst others are get worse (e.g. false insertion percentage). The same phenomenon occurs with the autocorrelation when one is varying the *voicing decision threshold*. Performance results discussed in the next section were obtained by varying these parameters between each database processing session.

4. COMPARISON RESULTS AND DISCUSSION

4.1. Evaluation results

Performance results take the form of two-dimensional cross-plots, The legend associated with the last figure also applies to the previous ones. The results obtained with the "origina" methods, i.e. as they are presented in section 2.1 are plotted with solid lines whereas results obtained with "homogenised" pre- and post-processings (see section 3.2) are plotted with dashed lines[2].

4.2. Discussion

The graphs presented in this paper do not show the high degree of inter-speaker variability existing in the results. Hence, the distance between curves that one can visually observe should not be interpreted as a statistically significant difference. As far as GCI detection is concerned (figure 1), the best results were obtained by the bilinear TFR based method when associated with the *Born-Jordan* time-frequency kernel. However, this method seems particularly sensitive to the choice of its associated kernel. The SIFT-based method turned out to be the most robust to speaker characteristics as far as the bias in GCI estimation was concerned (see section 3.3).

The modulated gaussian wavelets based method realises the best performance relative to the voicing decision ability tests. Considering F0 estimation (figures 2 and 3), the best results were obtained by the SIFT-based method in its "homogenised form", i.e. without any application of a Hilbert transform before autocorrelation. This changing in score indicates that the corresponding resulting signal is less contrasted than the one resulting from the Born-Jordan kernel method. The auditory model has been shown in previous studies to compare favourably to the SIFT-based method in adverse noisy

[1] The normalisation included in this definition frees the performance results from any dependency on F0.

[2] Once displayed from the proceedings CD-ROM, graphs should appear in color.

conditions [**?**] while achieving poorer performances when applied to the database used in this study.

5. CONCLUSION

We have presented a comparison of methods for F0 estimation and GCI detection. Although none of them was originally designed with the intention to be used for voicing decision, we also evaluated their performance in this task. The Born-Jordan kernel time-frequency representation based method achieves the best global results and suggests that joint time-frequency analysis is a promising technique for GCI detection. Nevertheless, none of the methods investigated is significantly better than the others, neither globally nor if the evaluation criteria are considered individually.

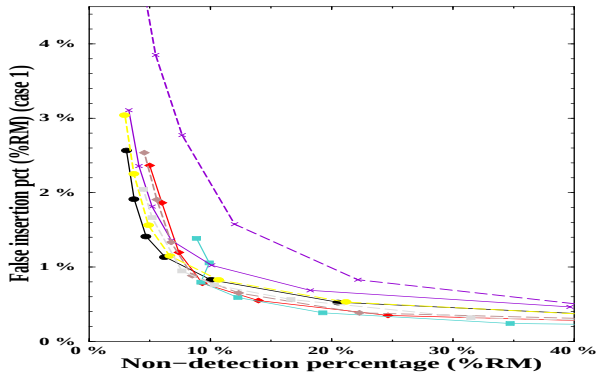

Figure 1: f.i. pct. (case 1) vs n.d. pct. (GCI detection).

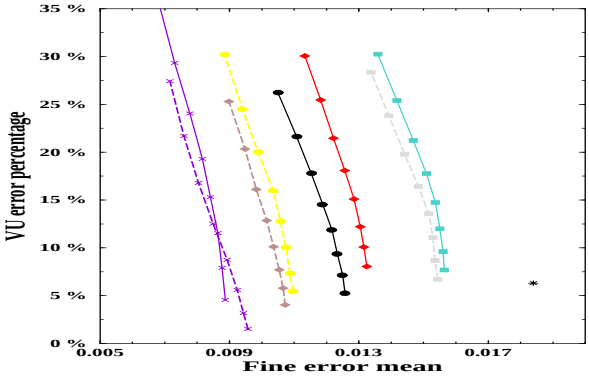

Figure 2: VU error pct. vs fine error mean.

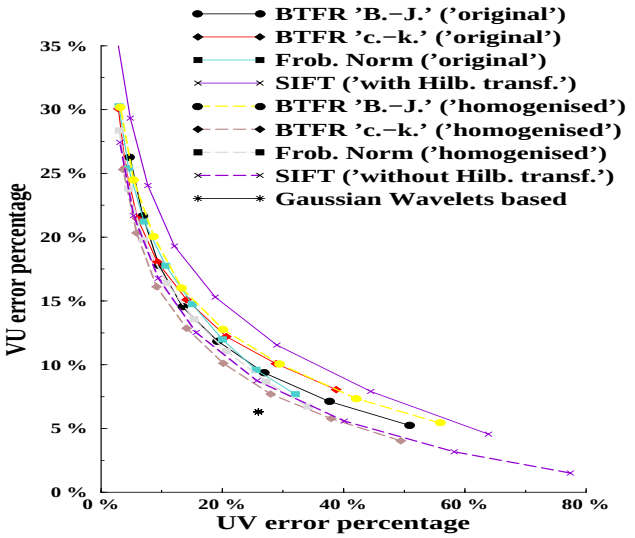

Figure 3: VU error pct. vs UV error pct.

6. ACKNOWLEDGMENTS

This work was supported by the HCM contract ERBCHRX-CT93-0098. We would like to thank Ignasi Esquerra, Leonard Janer, Brian Karlsen, Juan-Luis Navarro and Gethin Williams for their friendly and great help. We are also very grateful to Fabrice Plante for several enlightening discussions.

7. REFERENCES

1. Y. Kamp C. X. Ma and L.F. Willems. A frobenius norm approach to glottal closure detection from the speech signal. *IEEE Trans. on Speech and Audio Processing*, 2(2):258–265, 1994.
2. G. F. Meyer F. Plante and W. A. Ainsworth. Pitch detection: Auditory model versus inverse filtering. *Proc. I.O.A.*, 16(5):81–88, 1994.
3. G. F. Meyer F. Plante and W. A. Ainsworth. A pitch extraction reference database. *Proc. EUROSPEECH'95*, pages 837–840, 1995.
4. L. Janer. A modulated gaussian wavelet transform based speech analyser pitch determination algorithm. *Proc. EUROSPEECH'95*, pages 401–404, 1995.
5. A. K. Krishnamurthy and D. G. Childers. Two-channel speech analysis. *IEEE Trans. ASSP*, 34(4):730–743, 1986.
6. E. Mousset. *Comparison of glottal closure instant detection methods*. Tech. Rep. TR95-18/ISSN: 1353-7776, Computer Science Dept., University of Keele, 1995.
7. J.-L. Navarro and I. Esquerra. A time-frequency approach to epoch detection. *Proc. EUROSPEECH'95*, pages 405–407, 1995.