ROBUST F_0 AND JITTER ESTIMATION IN PATHOLOGICAL VOICES

Maurílio N. Vieira[1,2], Fergus R. McInnes, and Mervyn A. Jack
E-mail: maurilio, fmi, maj@ccir.ed.ac.uk

CCIR, Dept. of Electrical Engineering
University of Edinburgh
80 South Bridge, Edinburgh EH1 1HN, Scotland, UK

ABSTRACT

Dysphonic voices were used to compare electroglottographic (EGG) and acoustic measures of fundamental frequency (F_0) and jitter using a wavematching and an event-based technique. Continuous speech was considered in the first part of the study, where the effects of pre-filtering the acoustic signals and linearly smoothing the F_0 contours were analysed. The second part of the investigation compared jitter from sustained vowels (/i/, /a/, /u/), resulting in poor agreement for /i/ and /u/. In /a/ vowels, however, a relatively small mean normalised absolute difference (10.95%) was obtained with a method that is being proposed, which combines peak-picking and zero crossings, being able to detect a waveform pattern observed in such vowels and reject unreliable measures.

1. INTRODUCTION

Several studies based on synthetic waveforms or voices from normal speakers have systematically evaluated recording procedures or perturbation analysis algorithms [e.g., 1,2], leading to important standardisation proposals [3]. Most of the usual perturbation indices (i.e., jitter, shimmer, and measures of glottal noise) rely on the accurate determination of glottal cycle's boundaries, and were more precisely obtained (in synthetic signals) with wavematching techniques [2]. Artificially produced perturbations (additive noise, and amplitude or frequency modulation) allow better controlled experiments, but the extent to which they simulate the perturbations present in natural voices, especially from dysphonic speakers, remains uncertain. In fact, inconsistent results have been reported regarding perturbation measures from dysphonic voices based on the currently available fundamental frequency (F_0) extraction methods [4].

Similar jitter measures have been obtained, in a wide range of vowels and controlled levels of F_0 and intensity, using simultaneously recorded electroglottographic (EGG) and acoustic signals from normal speakers [5]. Electroglottography is an attractive method to obtain a clean signal related to the laryngeal phonatory function, but the comparison between microphone and EGG signals should be done with caution. EGG recordings can be affected by artifacts (e.g., position of the electrodes, laryngeal vertical movements, or metallic necklaces), as is well known, and differential effects on EGG- and acoustic-derived measures are also possible in the presence of laryngeal disorders. On the other hand, the reliability of acoustic F_0 extraction can be sensitive to such factors as the signal's bandwidth and sampling frequency, or the vowel type.

In the study described in this paper, F_0 and jitter measures based on a wavematching technique [6], and an event-based approach [7] (along with a slightly modified version of the latter) were compared with measures from EGG signals. In the first part of the study, continuous speech was used to evaluate the acoustic F_0 tracking methods, while, in the second part, sustained vowels were used to assess jitter estimation strategies.

2. F_0 ESTIMATION

Although most of the acoustic analysis in laryngeal disorders assessment is carried out using sustained vowels, continuous speech was used in this part of the study to test the ability of the algorithms to track phonatory abnormalities (e.g., pitch breaks, or modal-to-falsetto transitions), commonly present in dysphonic speakers.

Simultaneous acoustic and EGG recordings were taken from 15 adult patients (8 females, 7 males) suffering from laryngeal disorders (e.g., nodules, polyps, papillomas, or functional dysphonia). The speech material, which was part of a more elaborate protocol used in a hospital clinic, consisted of a short paragraph digitised at 22,050 samples per second, 16 bits per sample, totalling 4.5 minutes over all patients.

2.1 EGG analysis [8]

Initially, the EGG baseline drift was attenuated by bandpass filtering the signals (60-5000 Hz) using two consecutive filterings and time inversions to achieve exact zero phase shift. Assuming that up-going EGG signals indicate increasing vocal fold contact area, individual fundamental periods in the EGG signals were automatically obtained as the elapsed time between two consecutive linearly interpolated up-going zero crossings, F_0 values being limited to 50-500 Hz. The F_0 contours were visually inspected and corrected, if necessary, after a cycle-by-cycle comparison of the F_0 contours with the corresponding acoustic and EGG waveforms. The corrected files were used as reference values in the comparison presented below.

2.2 Wavematching Method

The wavematching strategy (Super Resolution, or SR) [6, 9] determines the fundamental period (T_0) at a generic discrete time "t" in a voiced interval as the duration "T" that maximises the cross-correlation coefficient between two equal-sized adjacent segments defined by the intervals [t-T, t-1], and [t, t+T-1], respectively. In the study, the range of T_0 corresponded to 50-500 Hz, and the observation instant "t" was shifted by the estimated T_0, during voiced segments, or by 5 ms, during unvoiced intervals.

A linearly smoothed F_0 contour (F'_0) was also obtained by means of a 3-point Hanning window:

$$F'_0(i) = \tfrac{1}{4} \cdot F_0(i) + \tfrac{1}{2} \cdot F_0(i-1) + \tfrac{1}{4} \cdot F_0(i-2) \qquad (1)$$

where $F_0(i)$ is an instantaneous fundamental frequency value.

Finally, the 15 acoustic recordings were lowpass filtered (Chebyshev II, cut-off at 1000 Hz, zero phase shift) and the corresponding F_0 and F'_0 contours were obtained. The cut-off frequency of the lowpass filter corresponded to at least twice the maximum estimated $F_0[1]$, while zero phase filtering was used to facilitate visual comparisons of acoustic and EGG waveforms.

2.3 Event-based Method

The event-based algorithm (KSV) was the method proposed by Schäfer-Vincent [7], where computationally expensive numerical methods are replaced by rules designed to implement a kind of "empirical correlation."

The algorithm initially identifies a "significant point" (SP), which is the maximum sample (and, similarly, the minimum) in a running 2-ms window, this length establishing the maximum detectable F_0 value (500 Hz). Six tests are then applied, combining the newly detected SP with up to 100 past SP values, in order to identify "period-twins", i.e., two consecutive and similar glottal cycles. A generic period-twin is delimited by three significant points at the instants t1, t2, and t3 (Fig. 1), having the following properties: the F_0 values corresponding to t2-t1 and t3-t2 are no less then 50 Hz (test 1); a measure of cycle-to-cycle frequency perturbation (jitter) is no more than 10% (test 2); the amplitudes of the significant points are higher than a noise threshold (test 3); a measure of cycle-to-cycle amplitude variation (shimmer) is no more than 50% (test 4); the amplitudes of the significant points are the biggest samples ("envelope") in the
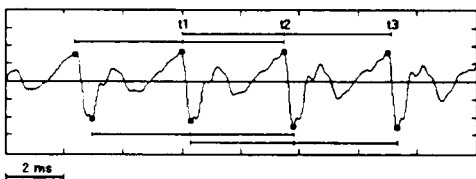

Figure 1: Significant points (•) and period-twins (├──┤)

period-twin (test 5); and, finally, the waveforms of the period-twins must be "similar" (test 6). Another relatively complex set of rules ("chaining") is then applied to concatenate the so-detected twin-periods. The reader is referred to the original paper for details.

A modified version of the algorithm (KSV mod) was also evaluated. In this version, the search for period-twins was limited to the *first* successful attempt, instead of considering *all* (up to 100) possible period-twins. Additionally, "test 6" in the original algorithm, which consumed approximately 30% of the processing time, was replaced by a simpler test. The new test 6a, shown in the pseudo-code below, compared the waveforms of the two parts of the twin-periods in a simpler way, by using linear time alignment between the intervals t2-t1 and t3-t2 to calculate an absolute difference of amplitudes (normalised to a gross estimate of the signal energy, E), which was compared to an empirically determined threshold (0.14) to accept or reject the similarity of the waveforms.

```
test6a()
{
    NormAbsDif =0;
```

$$E = (t3 - t1) \cdot \frac{\left| x_{t1} + x_{t2} + x_{t3} \right|}{3} ;$$

$$\alpha = (t3 - t2 - 1) / (t2 - t1 - 1);$$

```
    for (z=t1; z<t2; z=z+dec_factor) {
```

$$w = (\text{int}) \left(\alpha \cdot (z - t1) + t2 \right);$$

$$NormAbsDif = NormAbsDif + \left| x_z - x_w \right|;$$

```
    }
    NormAbsDif = NormAbsDif/E;
    if (NormAbsDif < 0.14/dec_factor) return TRUE;
    else return FALSE;
}
```

Due to the high sample rate (22050 Hz) and the close resemblance between the two "halves" of the candidates to period-twin that reach this test, the decimation factor (dec_factor) can assume a value between 2 and 8, without noticeable deterioration in performance (the results presented in the paper corresponding to dec_factor = 2). Regarding speed, the pruned search for twin-periods and the simplified "test 6a" reduced the processing time by approximately 30%.

Fundamental frequency contours, as well as the respective smoothed time series (Eq. 1), were obtained with the original and modified versions of the algorithm running on both the unfiltered and lowpass filtered acoustic waveforms (processed as described in the end of section 2.2).

2.4 Results

The comparison of the acoustic and EGG F_0 contours was carried out in terms of precision and voicing-detection errors, and is summarised in Table 1, where the values (percentages) are

	H	L	IFI	FSD	U2V	V2U
UNFILT.						
SR	1.51	2.84	1.72	1.71	1.10	9.49
KSV	0.47	0.91	2.05	1.95	0.76	9.14
KSV mod	0.28	0.80	1.98	1.88	0.58	9.69
Lin. Sm.						
SR	2.58	3.52	2.24	1.97		
KSV	0.53	1.01	1.59	1.58		
KSVmod	0.36	1.06	1.55	1.55		
FILT.						
SR	1.83	3.01	1.82	1.77	1.77	6.05
KSV	0.34	0.81	1.79	1.74	0.80	8.64
KSVmod	0.24	0.76	1.75	1.70	0.63	8.98
Lin. Sm.						
SR	2.73	3.30	2.37	2.04		
KSV	0.37	0.89	1.49	1.53		
KSVmod	0.35	0.95	1.47	1.51		

Table 1: H (high errors, i.e., above +10%) and L (low errors, i.e., below -10%) are shown proportionally to the number of values in the acoustic F_0 contours being assessed; IFI is the mean absolute fine error (i.e., within ±10%) and FSD is the respective standard deviation; unvoiced (or silent)-to-voiced errors (U2V) and voiced-to-unvoiced (or silent) errors (V2U) are shown proportionally to the utterance duration.

averages over the 15 patients, no sex difference being observed. The missing entries for the linear smoothed contours (Lin. Sm.) are identical to the respective values for the non-smoothed contours in both the unfiltered (unfilt.) and filtered (filt.) cases.

Note that (a) high (H) and low (L) errors were more frequent in the SR method than in the event-based methods; the relative immunity to H and L errors (caused, mainly, by F_0 doublings and halvings, respectively, which are common in pathological voices), seems to be a strong feature of the KSV algorithm; (b) smoothing increased H and L errors in all methods; (c) apart from voiced-to-unvoiced (or silent) errors, filtering worsened all indices for the SR method, with a further degradation after the linear smoothing operation; (d) filtering and linear smoothing of F_0 contours reduced the fine errors (IFI) in the event-based method.

If a choice has to be made, a compromise between speed and performance seems to be the "KSV mod" algorithm (which is approximately 8 times faster than the SR algorithm) running on unfiltered signals, with linear smoothing of the F_0 contours. It is important to emphasise, though, that smoothing can mask true perturbations and an accurate F_0 estimation does not necessarily represent an accurate jitter extraction [2], as will be shown next.

3. JITTER ESTIMATION

Acoustic and EGG signals were digitised, as described before, during the maximum sustained phonation of 3 English vowels (/i/, /a/, /u/ as in sheep, father, and food, respectively).

3.1 EGG Jitter

After the attenuation of EGG baseline fluctuations (section 2.1), F_0 contours were obtained from events (interpolated zero crossings) and wavematching. Due to poor EGG signals, recordings from 5 /i/ vowels, 3 /u/ vowels, and 3 /a/ vowels were excluded from the remaining analysis. Jitter time series, PF1(i) ("first order perturbation factor" [2]), and "restricted" perturbation factors, $PF1_{10}$, were calculated from the non-excluded F_0 contours, as

$$PF1(i) = \frac{|F_0(i+1) - F_0(i)|}{\frac{1}{2} \cdot [F_0(i+1) + F_0(i)]} \times 100, \quad (2a)$$

$$PF1_{10} = \frac{1}{N_{10}} \sum_{i=1}^{N_{10}} PF1_{10}(i), \quad (2b)$$

where $PF1_{10}(i)$ is an instantaneous jitter estimate not higher than 10%, and N_{10} is the total number of such values in each file. The 10% restriction was introduced to limit EGG and acoustic jitter to the same range since, contrary to EGG analysis, instantaneous jitter above 10% was difficult to extract from acoustic signals due to excessive voicing detection errors.

3.2 Acoustic Jitter

The recordings were lowpass filtered (cut-off at 1000 Hz, zero phase shift filtering), and jitter was obtained from the filtered and unfiltered signals as described below.

Wavematching. The Super Resolution algorithm was used to obtain jitter (Eq. 2), but the linear smoothing was not applied, based on the results of the previous section.

Peaks. Jitter measures (Eq. 2) were obtained from the negative and positive period-twins detected by the KSV mod algorithm. Jitter from parabolically interpolated peaks were also calculated, but only the 4th decimal place (or less) of the jitter values was affected, probably because (a) the higher level of perturbation in the dysphonic voices masked the effects of a less precise peak determination, or (b) there was a smoothing due to the high number of cycles analysed (i.e., from ~300 to ~7,800).

The chaining strategy of the KSV mod algorithm concatenates twin-periods of both polarities, so that jitter based on smoothed F_0 could not be extracted from positive and negative peaks separately; jitter deteriorated with smoothing and was calculated only for /a/ vowels.

Zero Crossings. (PZN, NZP). Jitter from down-going (PZN) and up-going (NZP) zero crossings were calculated as:

$$PF1_{10} (PZN) = \tfrac{1}{2}[PF1_{10}(\bar{P}) + PF1_{10}(\bar{N})] \quad (3a)$$

$$PF1_{10} (NZP) = \tfrac{1}{2}[PF1_{10}(\bar{N}) + PF1_{10}(\bar{P})] \quad (3b)$$

where $\bar{P}$ and $\breve{P}$ represent F₀ contours based on zero crossings after or before positive peaks, respectively, $\bar{N}$ and $\breve{N}$ being similarly defined for negative peaks. To increase the speed and to avoid noisy situations, the search for zero crossings was limited to 1.25 ms.

Max. Ratio (/a/ vowels). A "ratio" parameter, designed to assess the accuracy in the identification of the zero crossings, was calculated as:

$$\text{ratio (PZN)} = \min[PF1_{10}(\bar{P}) / PF1_{10}(\bar{N}), \quad PF1_{10}(\bar{N}) / PF1_{10}(\breve{P})]$$

(4a)

$$\text{ratio (NZP)} = \min[PF1_{10}(\breve{N}) / PF1_{10}(\breve{P}), \quad PF1_{10}(\breve{P}) / PF1_{10}(\breve{N})]$$

(4b)

Several cases with high ratios (>80%) in both directions of zero crossings but discrepant jitter estimates occurred in /i/ and /u/ vowels. In /a/, however, high ratios resulted in only one zero crossing direction, and *the jitter value associated to the maximum ratio, when ratio was higher than 80%, was taken as the measure for the vowel (no measure being provided when ratio was less than 80%)*. This peculiarity in /a/ seems to be related to the higher frequencies and smaller bandwidths of the first formant, as compared to the other two vowels.

3.3 Results

The comparison between EGG and acoustic jitter was carried out in terms of normalised absolute differences, Δ:

$$\Delta(\cdot) = \frac{|PF1_{10}(EGG) - PF1_{10}(\cdot)|}{PF1_{10}(EGG)} \times 100\%$$

(5)

whose average was calculated for each vowel and acoustic jitter extraction method (·). Acoustic and EGG measures based on the same strategy (i.e., events or wavematching) had a better agreement than crossed comparisons (i.e., wavematching methods *versus* event-based methods), apart from wavematching on /i/ vowels (Table 2). This table summarises the conditions that resulted in the smallest differences between the acoustic and EGG jitter measures.

	/i/	/a/	/u/
Neg. Peaks	90.87[f]	43.03[f]	116.70[f]
Pos. Peaks	95.79[f]	36.51[f]	136.86[f]
Lin. Sm. Peaks		62.21[f]	
PZN	47.51[f]	26.20[u]	64.30[f]
NZP	66.83[f]	73.71[u]	87.29[f]
SR	*36.44[f]	27.99[f]	31.45[u]
Max. Ratio		10.95[u]	

Table 2: EGG and acoustic jitter differences. The superscripts "f" and "u" refer to filtered and unfiltered acoustic signals, respectively. Acoustic jitter from the SR method in the vowel marked (*) was compared with EGG jitter from zero crossings.

Poor agreement resulted from measures based on peaks and zero crossings (PNZ and NZP). In general, the agreement was better for the wavematching method, but the differences were still considerably high. There was a tendency towards smaller differences in /a/ vowels and the "max. ratio" method explored this feature, combining up-going (NZP) and down-going (PZN) zero crossings to provide relatively small differences.

4. CONCLUSION

Acoustic jitter extraction from /a/ vowels using the "max. ratio" method resulted in a relatively low mean normalised absolute difference (10.95%), considering that two different signals were being compared. Zero crossings from /i/ and /u/ vowels, detected as described in the paper, are not reliable acoustic correlates of the glottal closure, as detected in EGG signals. The results in /a/ vowels are encouraging and the integrated analysis of EGG, acoustic, and videostroboscopic recordings can lead to a better understanding of the laryngeal behaviour, especially in the presence of disorders.

4. REFERENCES

1. Titze, I.R., Horii, Y., and Scherer, R.C., "Some technical considerations in voice perturbation measurements." *J. Speech Hear. Res.* 30: 252-260, 1987.

2. Titze, I.R., Liang, H. "Comparison of F₀ extraction methods for high-precision voice perturbation measurements. *J. Speech Hear. Res.* 36: 1120-1133, 1993.

3. Titze, I.R. "Workshop on Acoustic Voice Analysis: Summary Statement." NCVS, Iowa, 1995.

4. Bielamowicz, S., Kreiman, J., Gerratt, B.R., Dauer, M.S., and Berke, G.S. "Comparison of voice analysis systems for perturbation measurements." *J. Speech Hear. Res.* 39: 126-134, 1996.

5. Orlikoff, R.F. "Vocal Stability and vocal tract configurations: An acoustic and electroglottographic investigation." *J. Voice* 9: 173-181, 1995.

6. Medan, Y., Yair, E., and Chazan, D. "Super resolution pitch determination of speech signals." *IEEE Trans. Signal Proc.* 39: 40-48, 1991.

7. Schäfer-Vincent, K. "Pitch period detection and chaining: Method and evaluation." *Phonetica* 40: 177-202, 1983.

8. Vieira, M.N, McInnes, F. R., and Jack, M.A. "Analysis of the effects of electroglottographic baseline fluctuations on the F₀ estimation in pathological voices." *J. Acoustic. Soc. America.* In press.

9. Bagshaw, P.C., Hiller, S.M, and Jack, M.A. "Enhanced pitch tracking and the processing of F₀ contours for computer aided intonation teaching" *Proc. Eurospeech'93*: 1003-1006, 1993.