IMPROVING PITCH ESTIMATION WITH SHORT DURATION SPEECH SAMPLES

W A Ainsworth, C R Day, G F Meyer

Department of Communication and Neuroscience,
Centre for Human & Machine Perception Research,
Keele University, Keele, Staffs ST5 5BG, UK
{bill,charles,georg}@cs.keele.ac.uk

ABSTRACT

Hermes' Sub Harmonic Summation (SHS) pitch determination algorithm is an effective technique for extracting the percept of pitch from human speech [1]. Effective determination of the pitch in a passage of speech is believed to be fundamental for higher level speech processing applications such as speech or speaker recognition.

Of particular interest is the need to extract pitch from speech in less than ideal conditions eg. in the presence of noise or using very short analysis windows.

In an attempt to deliver accurate pitch estimates from relatively short analysis windows this paper describes an evaluation of two forms of the SHS procedure: in one case, FFT-SHS, the procedure uses the conventional Fast Fourier Transform (FFT) in its spectral analysis step; in the second case, RAFT-SHS, the ReAssigned Fourier Transform (RAFT) technique [2] is used instead of the FFT.

1. INTRODUCTION

The RAFT technique has the potential to deliver higher resolution spectrograms than the FFT for a given signal analysis window [3]. The reassignment of energy with respect to both time and frequency, such that the reassignments model the time-frequency fluctuations of the sampled signal, allows the RAFT to deliver the improved resolution.

An evaluation of the RAFT-SHS and FFT-SHS techniques has been carried out using data from a *pitch extraction* database held in an open ftp site at Keele University in the UK [4]. An important feature of the Keele database is the availability of reliable ground-truth pitch-estimates for the voiced speech of 5 male and 5 female speakers. The ground-truth pitch estimates were derived using an autocorrelation technique, applied to laryngograph traces which were made when each speaker's utterance was recorded.

Using analysis windows of 50, 25 and 12 ms the two forms of SHS have been tested to see how robustly their pitch estimates compare with the laryngograph ground-truth as the amount of information present in the sampled signal was systematically reduced.

2. COMPARING RAFT & FFT PROPERTIES

The cause of poor spectral or temporal resolution in conventional spectrographic analyses, such as the FFT, arises from the use of time-frequency windows in which all of the energy is assigned to the centre of a frequency/time window [5]. Kodera et al [6] showed that the restriction of assigning energy to the centre of the analysis window could be overcome to produce spectrograms which more accurately modeled the input signal. Kodera's insight was to reassign the energy to points in the analysis window which corresponded to the *centre of gravity* of the signal's energy within the analysis window. These points of reassignment can be precisely calculated in both the temporal and spectral domains: equations 1 and 2, where t and f are the time and frequency points calculated by the conventional FFT.

$$t_r = t - \frac{1}{2\pi}\frac{\partial\phi}{\partial f} \tag{1}$$

$$f_r = f + \frac{1}{2\pi}\frac{\partial\phi}{\partial t} \tag{2}$$

Application of equations 1 and 2 means that the point of assignment is moved in both time and frequency according to the the partial derivative of the phase (ϕ). For spectral analyses, only the point of reassignment in the frequency domain needs to be calculated. Auger et al [2] devised a computationally efficient means for calculating the frequency displacement, without using the partial derivative of the phase, equation 3:

$$f_r = f + \frac{1}{2\pi} Im\left[\frac{FFT_{dh}(f) \times FFT_h(f)}{|FFT_h(f)|^2}\right], \tag{3}$$

where f is the frequency point determined by the conventional FFT, FFT_h is the Fast Fourier Transform using the time-window h and FFT_{dh} is the FFT calculated using the derivative of the window with respect to time. (Auger et al also gave a similar, computationally efficient, expression for the points of reassignment in time where FFT_{dh} is replaced by FFT_{th} - the FFT using the window multiplied by the time t.)

When using the reassigned method to transform into the frequency domain the frequency resolution of the resulting spectrum is unlimited [7], but the separation between frequency points is no longer uniform. In order to obtain a useful reassigned spectrum it is necessary to re-sample the reassigned spectrum at regular intervals. By a process of summation within each interval a spectrum with a semantics which is similar to that of the FFT can be derived. Unlike the semantic interpretation of an FFT spectrum however, those frequency points to which the RAFT has not assigned any energy cannot be strictly interpreted as having zero energy. The energy at any of the unassigned frequency points is *undefined*, only those points to which the RAFT reassigns zero energy can be interpreted as zero valued.

Since the computationally efficient implementation of the RAFT relies upon a ratio of FFTs (equation 3) the input and output vectors for a RAFT implementation should have a length which corresponds to an appropriate power of 2.

3. EXPERIMENTAL PROCEDURE

The utterances and ground-truth data available in the Keele *Pitch Extraction Database* all used a sampling frequency of 20 kHz. The ground-truth data was evaluated for overlapping samples of 512 points (25 ms), taken at 200 point (10 ms) intervals. Using an auto-correlation technique on all of the frames of speech for a single speaker allows a voiced/unvoiced decision to be made and an estimate of the pitch for each voiced frame to be derived. Table 1 shows the number of voiced frames and the mean pitch for each of the 5 male and 5 female speakers.

	Voiced Frames	Mean Pitch		Voiced Frames	Mean Pitch
M1	1818	100 Hz	F1	1531	192 Hz
M2	1382	134 Hz	F2	1902	226 Hz
M3	1461	134 Hz	F3	1510	190 Hz
M4	1624	93 Hz	F4	1803	230 Hz
M5	2071	107 Hz	F5	1858	228 Hz

Table 1: Pitch Extraction Database summary.

For each speaker, two forms of the SHS procedure, FFT-SHS and RAFT-SHS, were used to deliver a pitch estimate for each frame of voiced speech. FFT-SHS and RAFT-SHS are both slightly modified versions of Hermes' original algorithm specification. The modifications were made to allow the algorithm to be run for speech samples of varying duration (50, 25 and 12 ms) rather than the 40 ms samples used by Hermes.

3.1. The RAFT-SHS & FFT-SHS Procedures

The SHS procedure can be broken down into three stages:

Pre-processing where the raw signal sample is adjusted ready for spectral analysis (importantly Hermes' pre-processing ends with the production of a 256 point zero-padded vector, used as input to the spectral analysis, only the first 40% (100 points) of this input vector contain any form of speech signal);

Spectral Analysis tries to identify the spectral components present in the sample - Hermes uses an FFT to carry out this task;

Post-processing where the output of the spectral analysis is honed and subjected to a summation procedure which delivers a pitch estimate.

A generic form of the SHS procedure can be specified to allow flexibility in the size of speech sample to be analysed. This generic form is the basis of the FFT-SHS procedure used for the experiments described in this study:

Pre-processing

1. The speech signal is sampled to give a sample of n points such that

$$n = \frac{V \times 40 \times A}{100}, \tag{4}$$

where V is the size of the vector that will be the input to the spectral analysis and A is a *compression factor* which is applied to the signal sample in the succeeding step.

2. The signal sample of n points is compressed by a simple averaging of succeeding sequences of A points. (Hermes specified $A = 4$ in conjunction with speech signals sampled with $F_s = 10$ kHz, in the experiments reported here $F_s = 20$ kHz requiring that $A = 8$).

3. A *Hanning Window* is applied to the compressed signal sample.

4. The windowed and compressed signal sample is then zero-padded (a process intended to "*increase the resolution of the spectrum*") so that the resulting vector has a length which is a power of 2 and has a composition which is 40% compressed signal and 60% zero-padding. (Hermes' vector contained 100 points of signal and 156 points of zero-padding.)

Spectral Analysis

1. An FFT spectral analysis is carried out on the zero-padded vector.

Post-processing

1. A spectral-peak enhancement is performed. By taking each peak in the spectrum and setting all off-peak spectral energies to zero Hermes removes spectral noise without impacting the magnitude or frequency of the energy peaks present. The *off-peak* energies are those which do not lie within 2 FFT points of a relative energy maximum. For the experiments described in this paper only the 20 largest peaks are retained by the peak enhancement procedure.

2. A Hanning filter is applied to the peak-enhanced spectrum.

3. The spectrum is converted from a linear frequency abscissa to a $\log_2(f)$ abscissa.

4. A cubic spline interpolation generating 48 equidistant points per octave (approximately 528 points for FFT-SHS over the range of frequencies retained for these experiments (ie. 0 - 2048 Hz), is applied to the log-scale spectrum to enhance the effectiveness of the subsequent harmonic summations.

5. The interpolated spectrum is multiplied by a raised arc-tangent function to simulate some characteristics of the human auditory system.

6. Finally, a harmonic summation procedure was carried out to produce a summed spectrum in which the energy at each point is the sum of the energies found at the 15 succeeding harmonic frequencies present in the spectrum. At the conclusion of the summation procedure the estimate for pitch is the frequency giving the maximum energy in the summed-spectrum.

The RAFT-SHS procedure differs from the FFT-SHS procedure as follows:

1. The zero-padding applied to the compressed sample of speech was completely omitted. In the original SHS algorithm Hermes chose to carry out the zero-padding to extract *higher resolution* from the FFT. Since the RAFT at re-sampling rate (RS) of 4 times is already delivering finer spectral resolution, this step was considered unnecessary. The RAFT-SHS procedure supplied only the compressed speech signal in RAFT's input vector.

2. In the peak-enhancement step of RAFT-SHS the number of peaks retained is $20 \times RS$.

3. For the cubic-spline interpolation the approximate number of interpolated points is $528 \times RS$.

4. The harmonic summation operation which is the final step of the SHS procedure was amended. The amendment was intended to reflect the higher spectral resolution obtained from the RAFT's re-sampling. Accordingly, instead of generating a sum-spectrum in which the energy found only at precise harmonic intervals is added together, the RAFT-SHS's sum-spectrum was generated by also including the energy from each harmonic's four nearest neighbours (2 lower points and 2 higher points).

3.2. Results

For each frame of voiced speech a pitch estimate was deemed correct if it deviated from the frame's ground-truth pitch-estimate by no more than ± 20 Hz. The percentage of incorrect pitch estimates was then calculated with respect to the total number of voiced frames for each speaker. The results are summarised in Table 2 below.

	SHS Error Rates (%)					
	50 ms		25 ms		12 ms	
Speaker	RAFT	FFT	RAFT	FFT	RAFT	FFT
M1	7.8	6.7	13.4	49.2	62.9	92.0
M2	16.7	19.6	13.3	36.1	49.7	96.1
M3	5.8	8.0	5.3	26.0	43.6	98.0
M4	3.4	7.2	12.1	70.5	59.7	95.2
M5	7.6	6.2	11.1	26.1	65.1	87.8
Mean	8.3	9.5	11.0	41.6	56.2	93.8
SDev	5.0	5.6	3.3	18.7	9.1	4.0
F1	13.3	17.1	10.0	17.2	10.3	94.3
F2	10.7	12.0	7.0	11.8	7.0	83.0
F3	12.0	15.7	8.7	18.2	9.6	92.5
F4	14.3	16.5	10.5	16.0	10.4	70.1
F5	8.5	11.5	5.9	9.9	4.8	81.7
Mean	11.7	14.6	8.4	14.6	8.4	84.3
SDev	2.2	2.6	1.9	3.6	2.4	9.7

Table 2: FFT-SHS & RAFT-SHS pitch estimation errors.

The presentation of the results in Table 2 has been structured to reflect the differing performance rates of both forms of SHS depending upon whether the speaker was male or female. To illustrate the diverging performance of both techniques for male or female speakers more clearly, the results are also displayed in Figure 1 below.

4. DISCUSSION

The results presented in Table 2 and Figure 1 show that when the speech-samples used for pitch determination are of long duration (ie. 50 ms) there is no difference in performance between FFT-SHS and RAFT-SHS regardless of whether the speaker is male or female.

When the duration of the speech-samples is reduced to 25 ms differences in performance begin to emerge. For FFT-SHS using male speech the performance is much worse

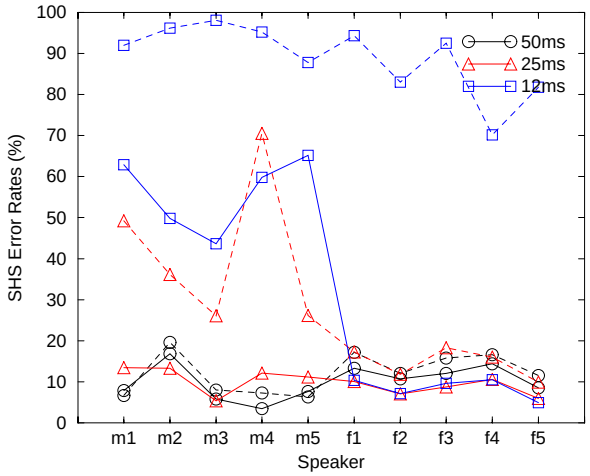

Figure 1: FFT-SHS (dashed lines) and RAFT-SHS (solid lines) results using the Keele Pitch Extraction Database.

than for female speech and the performance of FFT-SHS for both sexes is significantly worse than RAFT-SHS (average FFT-SHS error rate for male speakers is 41.6% and 14.6% for female speakers). Furthermore RAFT-SHS is much less sensitive to the speaker's gender (average RAFT-SHS error rate for male speakers is 11% and 8.4% for female speakers).

Finally, when the speech-sample duration is halved once more, to 12 ms, the difference in performance between the two forms of SHS becomes more pronounced and RAFT-SHS's performance becomes very sensitive to the gender of the speaker. (Clearly the average 12 ms performance of RAFT-SHS for male speakers (56% error) is unacceptably high for any worthwhile form of pitch estimation, but it is still far better than the equivalent performance of FFT-SHS (93.8% error).)

The relatively robust performance of the RAFT-SHS technique as the analysis window is halved is an indication that the RAFT technique is delivering important benefits not obtained with the FFT.

The RAFT-SHS procedure involves small deviations from the SHS procedure outlined by Hermes. The fine-tuning of the RAFT-SHS technique is justified since the deviations from Hermes' procedure merely eliminate some FFT fine-tuning (fine-tuning which was of course retained in the FFT-SHS procedure outlined above) and involve an undersampling of the high-resolution sum-spectra delivered by RAFT-SHS.

5. FURTHER WORK

For both male and female speakers at 50 and 25 ms the RAFT's good performance is robust. The cause of the RAFT's wide male/female performance difference at 12 ms is currently being investigated. It may be the case that since the average male speaker's pitch in the pitch extraction database is around 100 Hz and the average female speaker's pitch is around 200 Hz (see Table 1 above), the 12 ms window may be approaching the duration of the male speakers' average glottal periods - whilst the 12 ms window would be comfortably exceeding the female speakers' glottal periods. If this turns out to be the case the chances of restoring the RAFT-SHS 12 ms male-speaker performance to levels similar to those obtained for the female speakers at 12 ms would appear to be slim.

Examining this 12 ms discrepancy and finding methods to overcome it are the immediate topics for further work.

6. ACKNOWLEDGEMENT

This work is supported by the UK Government's Engineering & Physical Sciences Research Council (EPSRC), grant number GR/K77754.

7. REFERENCES

[1] Dik J. Hermes. Measurement of pitch by subharmonic summation. *Journal of the Acoustical Society of America*, 83(1):257–264, 1988.

[2] F. Auger and P. Flandrin. Improving the readability of Time-Frequency and Time Scale representations by the reassignment method. *IEEE Transactions on Signal Processing*, 43:1068–1089, 1995.

[3] Fabrice Plante, Georg F. Meyer, and William A. Ainsworth. Improvement of Speech Spectrogram Accuracy by the Method of Reassignment. *IEEE Transactions on Speech and Audio Processing*, 6(3):282–287, 1998.

[4] Fabrice Plante, Georg F. Meyer, and William A. Ainsworth. A Pitch Extraction Reference Database. In *EUROSPEECH'95*, volume 1, pages 837–840. European Conference on Speech Communication and Technology, 1995.

[5] J. W. Cooley and J. W. Tukey. An algorithm for the machine calculation of complex Fourier series. *Mathematics of Computation*, 19(90), 1965.

[6] K. Kodera, R. Gendrin, and C. deVilledary. Analysis of time-varying signals with small BT values. *IEEE Transactions on ASSP*, 34:64–76, 1978.

[7] Georg F. Meyer, Fabrice Plante, and F. Berthommier. Segregation of concurrent speech with the reassigned spectrum. In *International Conference on Acoustics Speech and Signal Processing*, pages 1203–1207, Munich, 1997.