NEW FUNDAMENTAL FREQUENCY ESTIMATORS FOR COCHLEAR IMPLANTS

Adrian Lee [1], Hugh McDermott [2], & W. Harvey Holmes [1]

[1] School of Electrical Engineering and Telecommunications,
The University of New South Wales
[2] Department of Otolaryngology, The University of Melbourne

ABSTRACT: Users of cochlear implants experience difficulties in listening to music. Improved processing of music signals for hearing implants requires an accurate and robust estimation of the musical pitch. This paper presents a comparative study of two different signal processing strategies for estimating the fundamental frequencies (F0) of almost-periodic input signals. Their performance with various inputs is compared, and implementation issues are discussed to ensure these strategies can be processed in real-time and are compatible with current cochlear implant technology.

1. INTRODUCTION

Fundamental frequency (F0) is a concept that is common to both speech and music signals. In speech, F0 often refers to the rate at which the vocal folds open and close (Deller, 1993), while for harmonic complex tones production in the context of music, F0 is often linked to an auditory attribute, commonly referred to as pitch. Therefore, one might postulate that a sound processing technique explicitly presenting an estimate of F0 for almost periodic sound signals should benefit implant users not only for speech perception, but also for music.

However, while advanced sound processing strategies are providing many implant users with considerable ability to understand non-tonal speech, studies have shown that implant users only appreciate the rhythmic elements of music, but not the pitch-related elements (Gfeller and Lansing, 1991 and 1992). Further studies by McDermott & McKay (1997) and Pijl (1997) revealed that, by bypassing the current sound processors used by the subjects and controlling the rate of pulsatile stimuli directly to a single electrode, musical pitch information can be conveyed. This suggests that a new sound processing strategy is warranted.

Although the technique of F0 extraction from sound signals is not a novel approach for cochlear implant strategies (Patrick and Clark, 1991), these sound processors in the past were designed specifically for speech signals. The aim of this project is, thus, to conceive new F0 estimators that are suitable for any complex harmonic tones, without explicitly assuming that the signal source is either speech or music. Such F0 estimator schemes would not only benefit the implant users in music appreciation, but could also possibly improve their perception of tonal languages, in which F0 plays a specific role for speech understanding.

2. DESIGN OF SIGNAL PROCESSING STRATEGIES

The two F0 estimators investigated have a nominal compass of 2 octaves, which, in musical terms, extends from A2 (110 Hz) to A♭4 (415.3 Hz). Their outputs are to be quantised to the twelve chroma defined by the equal tempered musical scale based on the western tonal tradition, although there is no explicit assumption that the signals are derived from musical instruments instead of from speech. Two separate signal processing strategies are examined, as follows.

2.1 Phase vocoder with sifting algorithm

This strategy is based on frequency analysis using a phase vocoder as a spectrum analyser. The outputs of the phase vocoder bandpass filters are sieved according to a template based on the harmonics of each candidate F0, quantised to a musical scale. Specifically, only those partials that lie within a quarter-tone of a harmonic are passed by the sieve. The total power collected in each template is then calculated, and the results for each candidate F0 are compared to select the best estimate of the pitch.

Since the sifting algorithm effectively performs quantisation at its input, it is important to ensure that the output of the phase vocoder is accurate in its frequency estimates within a quarter-tone bound. Using a Hamming window, it was found that in order for the sifting algorithm to produce accurate F0 estimates within the nominal range, the sampling frequency must be below 8,200 Hz when using a 128-point Fast Fourier Transform (FFT) to obtain an estimate of the underlying frequency spectrum. Therefore the F0 estimator could potentially utilise all frequencies up to 4,100 Hz for the above sifting algorithm. Note that 8,000 Hz is usually considered to be the lower bound of sampling frequency in speech processing, and moreover, there is not much pitch information available beyond 4 kHz.

However, since there is a smaller number of entries in a template based on a high F0 than that with a low F0, the algorithm must be able to counteract this bias. Furthermore, the frequency intervals in each template are spaced arithmetically but have geometrically increasing error bounds, so that those frequency intervals above the 17th harmonic are no longer disjoint, and thus are not suitable as criteria for a sifting procedure. Also, Terhardt (1979) states that the relevance of upper harmonics in the calculation of F0 decreases linearly as the harmonic number increases. This suggests that some weighting function should be incorporated into the above sifting algorithm. The formula below shows the construction of one such weighted template:

$$A2_{template} = \sum_{k=1}^{10} (1-0.1k).P(f); \qquad f \in \bigcup_{k \in \mathbb{Z}} \left[k.f_{A2}.2^{-1/24}, k.f_{A2}.2^{1/24} \right)$$

where P(f) is the power associated with its frequency, and f_{A2} denotes frequency of pitch A2 (110Hz).

2.2 Sinusoidal coding analysis

The second strategy is based on the sinusoidal coding method, first proposed by McAulay and Quatieri in 1986, and subsequently elaborated in a series of papers (e.g. McAulay and Quatieri, 1995). Sets of amplitude, frequency and phase estimates, $\{A_l, \omega_l, \phi_l\}$, are generated for each harmonic constituent of the signal by minimising the mean-square difference between the signal and the estimated waveform, i.e. minimise:

$$\varepsilon(\omega_0, \phi) = \frac{1}{N+1} \sum_{n=-N/2}^{N/2} \left| s(n) - \hat{s}(n; \omega_0, \phi) \right|^2$$

The F0 estimate is performed implicitly, assuming that harmonic constituents are perfect harmonics of the fundamental, and it is optimal in the mean-squared-error (MSE) sense.

The key to the F0 estimation lies in the pitch estimation function, which is derived from the above MSE criterion. This function, which is a maximum at the optimum choice of ω_0, is

$$\rho(\omega_0) = \sum_{k=1}^{K(\omega_0)} \bar{A}(k\omega_0) \left\{ \max_{\omega_l \in L(k\omega_0)} \left[A_l D(\omega_l - k\omega_0) - \frac{1}{2} \bar{A}(k\omega_0) \right] \right\} \qquad (1)$$

where ω_0 is the fundamental (angular) frequency, A_l is the local maximum of the underlying Discrete Fourier Transform (DFT) picked by the Spectral Envelope Estimation Vocoder (SEEVOC) algorithm (Paul, 1981), $\bar{A}(k\omega_0)$ is the piecewise constant interpolated envelope of the chosen local maxima, and $D(\cdot)$ is a distance function. This distance function has the shape of the first lobe of a sinc function over an interval of L, and is pitch adaptive, since it is always an octave wide.

The main feature of this pitch estimation criterion is that, at least for low-noise speech signals, it avoids pitch halving and pitch doubling, which is a problem with many other pitch estimators, and therefore the frequency that maximises equation (1) gives an unambiguous F0 estimate.

However, since the design of sinusoidal coding and the SEEVOC algorithm was intended mainly for speech transmission at low bit rates, the criterion in its present form fails to capture the F0 of some music signals, such as flute tones, for which the amplitude spectrum is considerably different from those encountered in speech. Therefore the following modifications were made to both the pitch estimation criterion and the SEEVOC peak-searching algorithm:

1. In order to make the sinusoidal coding analysis perform at a Signal-to-Noise-Ratio (SNR) of 0 dB, which was not considered in the original algorithm, SEEVOC peaks close to the noise range are discarded.

2. Paul (1981) assumed in his SEEVOC algorithm that the speaker's individual pitch varies reasonably slowly over only about one octave. However, it is quite different for music signals, due to the fact that a typical melody is based on a construction of pitch intervals, which implies that there may be sudden changes in pitch, unlike in speech. Therefore, an extra criterion was added to the SEEVOC peak-searching algorithm to ensure that the global maximum of the FFT spectrum is found.

3. The pitch estimation criterion was further modified so that an accurate F0 estimate can still be obtained for harmonic complex signals that are very much different from those of speech, e.g. flute sounds. The modified criterion is:

$$\rho(\omega_0) = \sum_{k=1}^{K(\omega_0)} \bar{A}(k\omega_0) \left\{ \max_{\omega_l \in L(k\omega_0)} \left[A_l D_{alt}(\omega_l - k\omega_0) - \frac{\eta k}{2} \bar{A}(k\omega_0) \right] \right\} \qquad (2)$$

   where the new parameter ηk aims to de-emphasise the contribution of the upper harmonics, similar to the weighting function employed in the aforementioned sifting algorithm.

3. RESULTS AND DISCUSSION

3.1 Test signals

Four classes of test signals were used to test the two F0 estimators:

1. Artificially generated pure sinusoids.

2. Harmonic complexes consisting of 3 artificially generated sinusoids.

3. Sampled signals of musical instruments, one representative instrument from each family in the modern orchestra of the western tonal tradition – the violin represents the string family, the flute represents the woodwind, the trumpet represents the brass, and the piano is chosen to represent a pitched percussive instrument.

4. Sampled signal of a coloratura soprano singing the vowel /æ/.

Each class was tested under different SNR conditions, ranging from 30 dB to 0 dB.

3.2 Results of F0 estimations

The performance of the two F0 estimators in estimating artificial signals is summarised in Table 1, while their performance in estimating sampled signals with quasi-stationary steady-states is summarised in Table 2. The performance of the estimators in estimating highly non-stationary sounds such as the piano and the coloratura sung vowel is presented in Table 3.

From Table 1, it is evident that both strategies perform well as both chroma and pitch estimators. From Table 2, we can conclude that, even though both estimators are good chroma estimators, the strategy that is based on the phase vocoder analysis and sifting algorithm is sometimes affected by pitch halving and/or doubling. Table 3 shows that both estimators perform well above a segmental SNR (time-segmental SNR) level of 2 dB.

Figure 1 shows that both approaches are able to track the vibrato of the coloratura, which roughly equates to 5 beats per second. However, it is important to point out that pitch quantisation is applied at the input of the sifting algorithm, and cannot be reduced further at the output. This is contrary to the quantisation method of the sinusoidal coding analysis method, which is applied at the output. Thus, the sinusoidal coding analysis method can be more easily modified to incorporate an adaptive frequency resolution at the output.

Table 1. Summary of the two F0 estimators' performance as pitch estimators at SNR = 0dB for artificial samples, ie. correctly estimating both chroma and height. The accuracy of the estimators in providing the correct pitch estimate given for each time frame of analysis over the whole duration of the signal is expressed as a percentage. Note that the sinusoidal coding analysis in its present implementation cannot distinguish one particular pitch, since it lies between two FFT analysis bins.

CLASSES / ACCURACY (%)	PHASE VOCODER + SIFTING ALGORITHM	SINUSOIDAL CODING ANALYSIS
Single Tone, (F0)	98%	97%
Harmonic Tones (F0, 2F0, 3F0)	100%	99%

Table 2. Summary of the two F0 estimators' performance at SNR = 0dB for sampled signals with quasi-stationary steady-state waveforms. Their performance as pitch estimators, ie., correctly estimating both chroma and height, and as chroma estimators, ie., correctly estimating chroma only, are evaluated. Note: If chroma estimation varies from chroma + height estimation, it implies that pitch halving or doubling has occurred.

INSTRUMENTS		FLUTE		VIOLIN		TRUMPET	
Pitch / Accuracy (%)		G3	C4	G3	C4	C3	G3
Sifting Algorithm	Chroma + Height	92%	94%	56%	92%	40%	98%
	Chroma only	92%	94%	96%	92%	92%	98%
Sinusoidal Coding Analysis	Chroma + Height	93%	95%	93%	88%	93%	98%
	Chroma only	93%	95%	96%	88%	98%	98%

Table 3. Summary of the two F0 estimators' performance at segmental SNR = 2dB for piano G4, which has a short steady-state waveform and a long decay tail. The segmental SNR below which the algorithm fails to operate is also provided.

INSTRUMENTS		PIANO	
PITCH = G4		Accuracy	Segmental SNR Cutoff
Sifting Algorithm	Chroma + Height	96%	0.8dB
	Chroma only	96%	
Sinusoidal Coding Analysis	Chroma + Height	100%	1.0dB
	Chroma only	100%	

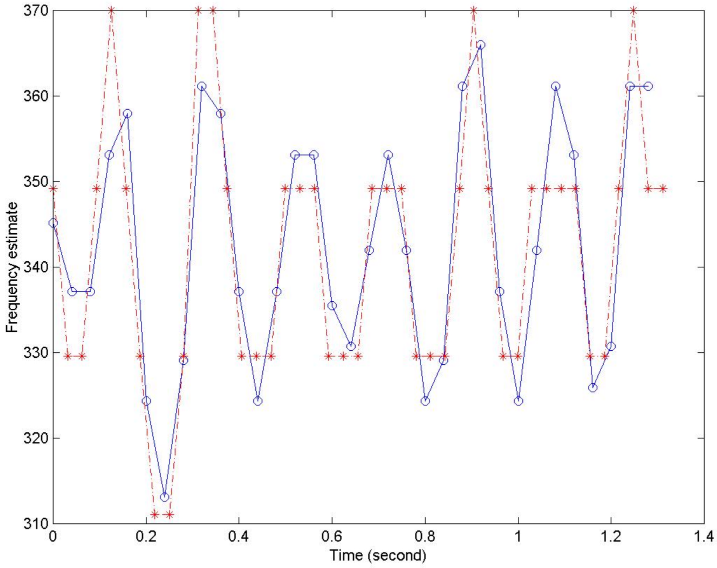

Figure 1. Graphical summary of the two F0 estimators when tracking vibrato of the coloratura (sung vowel /æ/ at pitch F4). [*] denotes the output of the sifting algorithm, and [^{o}] denotes the output of the sinusoidal coding analysis. Note the different quantisation levels in the two outputs.

3.3 Discussion of implementation issues

In the above F0 estimator simulations, the phase vocoder method employs a 128-point FFT for the appropriate frequency resolution, while the sinusoidal coding analysis algorithm employs a 1024-point FFT. Depending on the architecture of the Digital Signal Processing (DSP) chip and on how often the pitch estimation is made, the added complexity of 1024-point FFT could translate to as much as a twenty-four-fold increase in computation instruction cycles. A rough estimate of the computational cost reveals that the sifting algorithm requires in the order of 10^3 instruction cycles, while for the sinusoidal coding analysis the requirement is in the order of 10^4. For a DSP chip operating at around 15 MHz, this translates to 0.1 ms and 1 ms delay, respectively.

In the process of adapting sinusoidal coding for music signals, one modification is that the SEEVOC algorithm is called upon recursively, which could present problems in a real-time implementation. Some kind of escape mechanism must be incorporated. This could also potentially cause unacceptable delay of the signal to the implant users. In our simulation, the SEEVOC algorithm was called upon no more than five times recursively, which perhaps provides a worst case approximation.

However, the more critical point to consider is the width of the analysis frame. In the sinusoidal coding analysis, the analysis frame is constrained to be at least 20 ms. While it is conceivable to run the sinusoidal coding analysis as an F0 estimator, another algorithm, like that of a conventional implant sound-processor, needs to operate in parallel to give the implant users auditory feedback of their own voice with a perceptually acceptable delay. A delay of 20 ms is probably not acceptable in such applications. The phase vocoder, in this present version, provides spectral information for the user every 7.7 ms, and pitch information every 4 such consecutive frames of analysis using the sifting algorithm. This is a more practical scheme to implement considering the current DSP chip technology, and it is easily integrable with existing cochlear implant processors.

4. CONCLUSION

In this paper, we present two new F0 estimators that can be employed with current cochlear implant sound-processing strategies. The sinusoidal coding method provides frequency estimates whose resolution can be made adaptive at the output. It also substantially reduces the possibility of pitch halving or doubling. The sifting method, however, is shown to be more suitable for real-time implementation in current cochlear implant strategies, especially due to the fact that the phase vocoder can provide auditory feedback to users with an acceptable delay. Furthermore, if implant users typically cannot distinguish tones that are less than a semitone apart, it might be an acceptable trade-off of finer frequency resolution and occasional pitch doubling or halving for a computationally simpler sifting algorithm. Nonetheless, it is hoped that the F0 estimates generated from either of the above mentioned strategies can improve not only pitch perception for the cochlear implant users, but also provide them extra cues for tonal language understanding.

5. ACKNOWLEDGEMENTS

The first two authors would like to acknowledge the support provided by the Garnett Passe and Rodney Williams Memorial. The first author would also like to thank Department of Otolaryngology, The University of Melbourne, for their financial support.

6. REFERENCES

Deller, J., Proakis, J., and Hansen, J. (1993). *Discrete-Time Processing of Speech Signals*, Macmillan Publishing Company, New York.

Gfeller, K. and Lansing, C. (1991). “Melodic, rhythmic, and timbral perception of adult cochlear implant users”, *Journal of Speech and Hearing Research* 34, pp. 916-920.

Gfeller, K. and Lansing, C. (1992). “Musical perception of cochlear implant users as measured by the primary measures of music audition: An item analysis”, *Journal of Music Therapy* 29, No. 1, pp. 18-39.

McAulay, R. and Quatieri, T. (1995). “Sinusoidal coding”, in Kleijn, W., Paliwal, K. (eds), *Speech Coding and Synthesis*, Elsevier Science B.V., Amsterdam, pp. 123-173.

McDermott, H. and McKay, C. (1997). “Musical pitch perception with electrical stimulation of the cochlea”, *Journal of the Acoustical Society of America* 101, No. 3, pp. 1622-1631.

Patrick, J. F. and Clark, G. M. (1991). “The Nucleus 22-channel cochlear implant system”, *Ear and Hearing* 12, No. 4, Suppl., pp. 3S-9S.

Paul, D. (1981). “The spectral envelope estimation vocoder”, *IEEE Transactions on Acoustics, Speech and Signal Processing* ASSP-29, No. 4, pp. 786-794.

Pijl, S. (1997). “Labeling of musical interval size by cochlear implant patients and normally hearing subjects”, *Ear and Hearing* 18, No. 5, pp. 364-373.

Terhardt, E. (1979). “Calculating virtual pitch”, *Hearing Research* 1, pp 155-182.