Instantaneous Frequency Approaches for Speech and Audio Signal Analysis

J. Timoney and T. Lysaght
Department of Computer Science,
NUI Maynooth,
Maynooth,
Co. Kildare,
Ireland

Abstract
This paper investigates the use of Instantaneous Frequency Distributions for the analysis of Speech and Audio signals. In particular, methods are proposed for fundamental frequency determination of speech and audio, and method for the tracking of resonances, or formants in the case of speech, present in a signal is described. The techniques considered that are applied to the Instantaneous Frequency Distribution of the signal in order to extract a desired feature or features specifically involve the use of dynamic programming in order to produce smooth estimates.

1. Introduction

There has always been much interest in studying the amplitude and frequency modulation (AM/FM) structure of speech and audio signals as such a decomposition can indirectly describe the non-linear and time-varying phenomena that occur during their production [1][2]. A classic application of this decomposition is the well-known Phase Vocoder [3]. This type of work is also motivated by the better understanding of the signal processing function performed by the inner ear, particularly the cochlea. The cochlea is known to decompose acoustic stimuli into frequency components along the length of the basilar membrane. This phenomenon is called Tonotopic decomposition. Furthermore, it is also known that the nerve fibres emanating from a high-frequency location in the cochlea "phase-lock" to the envelope of the stimulus around that frequency, i.e. convey information about the envelope modulations in the signal. Thus, to a first-order approximation, it is often argued that the tonotopic location/place along the length of the basiliar membrane conveys the FM or frequency information about the signal, and the rate of nerve fibre activity around that location conveys the AM or envelope information [4]. Thus, the underlying FM of a signal carries information extremely important to its understanding. For speech, this has been demonstrated with the introduction of a time-frequency representation known as Instantaneous Frequency Distribution (IFD) proposed by [5]. In this work, the IFD was shown to clearly outline the formant structure of the speech signal, with a more defined resolution than the conventional spectrogram. A simple formant extraction procedure was also given based on forming a histogram of the Instantaneous Frequencies at each analysis point in time. More recently, [6], unaware of the work in [5], proposed the IF spectrogram as a means of viewing and analysing the harmonics present in voiced speech signals. The production of voiced speech involves the use of the vocal cords and the resulting sounds have a distinctive harmonic structure, a good example of this type of speech is the vowels. Based on this representation, an algorithm was proposed for the extraction of the frequency of the fundamental harmonic component, or pitch, as it varies in the signal over time [7]. This work intends to develop on the ideas presented in [7] and to modify them to produce reliable fundamental frequency detection algorithms for speech and music, an algorithm to determine if speech if voiced or unvoiced, and a resonance tracking algorithm. Such information is very useful in audio applications for tasks such as analysis and modification, and in particular the application to Timbre Morphing [8] will be highlighted in the conclusion.

2. The Instantaneous Frequency Distribution

The instantaneous frequency representation of a signal suggested by [5] is achieved through a Fast Fourier Transform (FFT)-based signal processing scheme. The running short-time Fourier transform (RSTFT) is interpreted in terms of simultaneous outputs of a bank of band-pass filters with successively offset centre frequencies, all having the given signal waveform as input. This is defined as

$$X_n\left(e^{j\omega}\right) = \sum_{m=-\infty}^{\infty} x(n+m)w(m)e^{-j\omega m} \tag{1}$$

where $w(m)$ is a real "window" sequence which determines the portion of the input signal which receives emphasis at time *n*.
(1) can also be represented in polar form as

$$X_n\left(e^{j\omega}\right) = \alpha(\omega,n)\, e^{j\theta(\omega,n)} \tag{2}$$

where $\alpha(\omega,n)$ is a real, non-negative magnitude and $\theta(\omega,n)$ is real modulo 2π. The local instantaneous frequency at any ω and *n* may be taken as

$$\begin{aligned} v(\omega,n) &= \frac{\partial}{\partial n}\theta(\omega,n) = \frac{\partial}{\partial n}\mathrm{Im}\left\{\log X_n\left(e^{j\omega}\right)\right\} \\ &= \mathrm{Im}\left\{\frac{1}{X_n\left(e^{j\omega}\right)}\frac{\partial}{\partial n}X_n\left(e^{j\omega}\right)\right\} \end{aligned} \tag{3}$$

To avoid explicit differentiation of $v(\omega,n)$, (1) can first be rewritten as

$$X_n\left(e^{j\omega}\right) = e^{j\omega n}\sum_{r=-\infty}^{\infty} x(r)w(r-n)e^{-j\omega r} \tag{4}$$

The derivative of (4) is given by

$$\frac{\partial}{\partial n}X_n\left(e^{j\omega}\right) = j\omega X_n\left(e^{j\omega}\right) + e^{j\omega n}\sum_{r=-\infty}^{\infty} x(r)\left[\frac{\partial}{\partial n}w(r-n)\right]e^{-j\omega r} \tag{5}$$

Consequently, the Instantaneous Frequency Distribution (IFD) is

$$v(\omega,n) = \omega + \mathrm{Im}\left\{\frac{\sum_{m=-\infty}^{\infty} x(n+m)\,w'(m)e^{-j\omega m}}{\sum_{m=-\infty}^{\infty} x(n+m)\,w(m)e^{-j\omega m}}\right\} \tag{6}$$

where

$$w'(n) = -\frac{\partial}{\partial n}w(n)$$

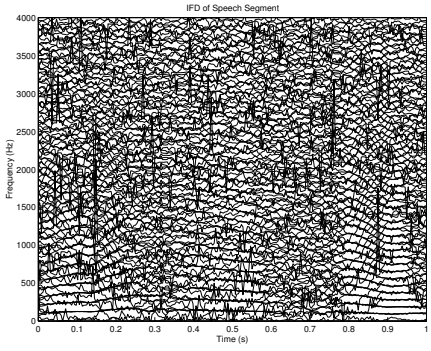

Figure 1. IFD for Speech Segment

Therefore, to find the instantaneous frequency distribution two RSTFTs are computed, one with $w'(n)$ in place of $w(n)$. The imaginary part of their ratio is taken at each ω and *n*, and finally, ω is

added. Difficulties can arise in calculating (6) when the denominator becomes very small but can be overcome by smoothing of the instantaneous frequency estimate [5]. Figure 1 is a plot of the IFD for a segment of speech. The areas of voiced speech can be discerned as those where the harmonic structure of the instantaneous frequencies are obvious while unvoiced speech appears as random fluctuations in the instantaneous frequency values. To extract the pitch of the speech at the voiced section, it is necessary to create a method that will determine the areas of the greatest concentration in the distribution and ignore the others.

3. Pitch Extraction

To determine the pitch of the signal, it is first proposed to recalculate the IFD so that it is nonlinearly spaced in frequency, to prevent interference from nearby spurious components in the estimation procedure. By using a Log-spaced frequency scale, the harmonics will appear at frequencies of $\log(P), \log(P)+\log(2), \log(P)+\log(3)\ldots$, where P is the fundamental period. Moreover, the nonlinear sampling will ensure that more instantaneous frequencies will be found in the lower frequency portion of the signal where the pitch value is most likely to lie. The log-spectrum at a time slice *n* is found by doing the following [10]:

First, create a vector of log sampled frequencies,

$$\omega l_i = 2\pi e^{(\ln f_b + id\ln f)} T_s \qquad i = 0,\ldots,N-1 \tag{7}$$

where f_b is the start frequency for the frequency sampling, T_s is the sampling period, *N* is the length of the frequency vector, and the sampling increment is given by

$$d\ln f = (\ln f_e - \ln f_b)/(N-1)$$

where f_e is the end frequency.

A log-spaced DFT can then be calculated by computing equation (8) directly

$$Xl_n(e^{j\omega l}) = \sum_{m=0}^{N-1} x(n+m)w(m)e^{-j\omega lm} \tag{8}$$

Similarly, (8) can be computed with the differentiated window function and then using (6) to create a log-spaced IFD. To extract the pitch frequency by applying the Dynamic programming (DP) procedure, it is necessary to create a likelihood function to which the DP algorithm can be applied. It is proposed here to use an evaluation function of the form

$$E(n, f_p) = \sum_{\omega l = f_b}^{f_e} Xl_n(\omega l) S(f_p, \omega l) H(f_p) \tag{9}$$

where

$Xl(\omega l)$ is the enhanced magnitude of the log-spectrum. This is created using the procedures outlined in [9] and [7]. First, the magnitude is flattened using a μ -law compression to remove the influence of the formants [9]. Next, as in [7], the spectral peaks are integrated at the points where the instantaneous frequencies are within the range of their frequency bin size, and are then weighted by the local second-order spectral moment. This ensures that the harmonics become well emphasised in the magnitude spectrum.

$S(f_p, \omega l)$ is an evaluation function computed across the range of values, given by f_p, at which the DP algorithm will search for the fundamental frequency. It is given by

$$S(f_p, \omega l) = \begin{cases} e^{v(\omega l)-f_p} & if\ v(\omega l) - f_p < 0 \\ e^{-(v(\omega l)-f_p)} & if\ v(\omega l) - f_p \geq 0 \end{cases} \tag{10}$$

This function will produce a peak of one when the DP frequency and the instantaneous frequency are the same, and should be sufficiently great when they are close in value. Otherwise, it should be very small.

Lastly, the function $H(f_p)$ is the value of the histogram, evaluated at a grid size corresponding to the DP search. This further weights the evaluation function towards the desired fundamental frequency.

The DP algorithm aims to find the values of f_p that will maximise the value of (9) over the entire signal length. The DP process basically finds the cumulative maximum of the evaluation function over time, and then extracts it via a backtracking search. Starting with the initial value

$$V(f_p, n_0) = E(f_p, n_0) \quad f_{pb} \le f_p \le f_{pe} \tag{11}$$

and carrying out the iteration

$$V(f_p, n) = E(f_p, n) + \max_{|i| \le i_0} V(f_p + i, n-1) \quad f_{pb} \le f_p \le f_{pe},\ n_0 < n \le n_1 \tag{12}$$

where n_0 and n_1 are the beginning and end time instants, *i* is the grid size, that is, the resolution of the search, and $\lfloor f_{pb}, f_{pe} \rfloor$ is the range over which the search takes place. The optimal fundamental frequency contour is then found by backtracking the local maximum of this global evaluation function.

$$f_p(n_1) = \max_{f_{pb} \le f_p \le f_{pe}} V(f_p, n_1)$$
$$f_p(n) = \max_{|i| \le i_0} V(f_p(n+1) + i, n) \quad n = n_1 - 1, n_2 - 1, \ldots \tag{13}$$

In order to produce a high resolution estimate, the grid size of the DP-search for the fundamental frequency was chosen as 1Hz. Figure 2 shows an example of this procedure carried out for both a male and a female speaker. The contour is shown by the black line, and it is clear from the plots that the match is good, with that for the male speaker being slightly better.

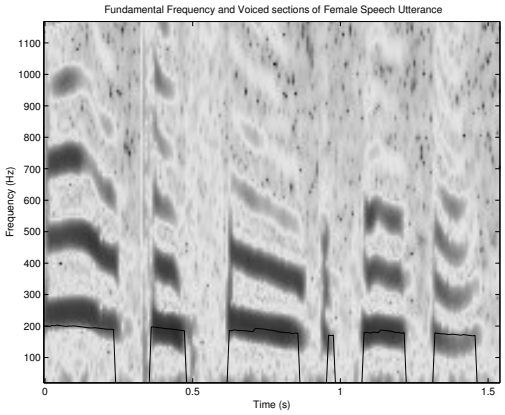

Figure 2 Female Speech Example

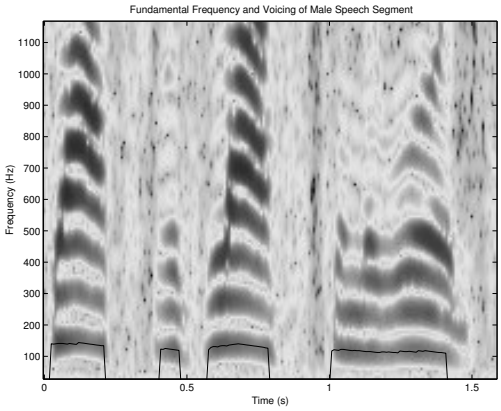

Figure 3 Male Speech Example

Voicing detection was achieved by measuring the variance of the IFD at each time instant and the amplitude of the fundamental harmonic. These quantities were median filtered, the mean subtracted, the sign value taken, and then combined to get the estimate. These regions appear on the plots as points of no contour. As can be seen voicing detection is good but not entirely accurate, the errors appearing at the edges of the voiced portions.

4. Application to Musical signals

To apply this procedure to fundamental frequency detection of musical signals, some modifications were necessary. As the log-sampling of the frequency spectrum produces widely spaced samples at

higher frequencies, it was decided that for Music signals, whose pitches are generally greater than that of speech, to revert to the linear sampled spectrum when calculating the IFD. It was also found that the bandwidth of the harmonics of instrument sounds are generally broader than those of speech, and produce spectral resonances rather than the narrow spectral lines typical of speech harmonics. Therefore, based on the work in [5] concerning speech, it was decided to use a very short time window when calculating the IFD and, for smoothness, to compute a short-time averaged version given by

$$\bar{v}(\omega, m) = \omega + \mathrm{Im}\left\{ \frac{\sum_{l=n_0}^{n_0+N} X_l'\left(e^{j\omega}\right) X_l^*\left(e^{j\omega}\right)}{\sum_{l=n_0}^{n_0+N} X_l\left(e^{j\omega}\right) X_l^*\left(e^{j\omega}\right)} \right\} \qquad (14)$$

where $X_n'\left(e^{j\omega}\right)$ is the time derivative of $X_n\left(e^{j\omega}\right)$ as given by (5) and $X_n^*\left(e^{j\omega}\right)$ is the complex conjugate of $X_n\left(e^{j\omega}\right)$.

The example below shows this function applied to the waveform of a sequence of notes that were played on a clarinet. The harmonics of the notes and their stepwise movement can be clearly seen.

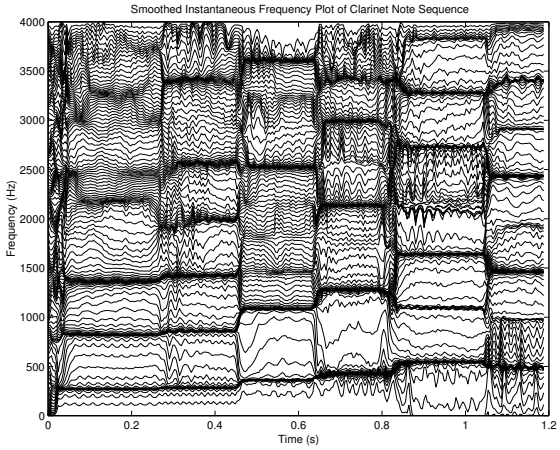

Figure 4. Smoothed IFD of Clarinet Sequence

Again, the fundamental frequency estimation algorithm was applied, this time using a linear spectrum. It was found that a larger grid size for the DP search produced better results than a smaller one. This was thought to be because the spectral transitions in the signal were sharper than the smooth contours in typical speech.

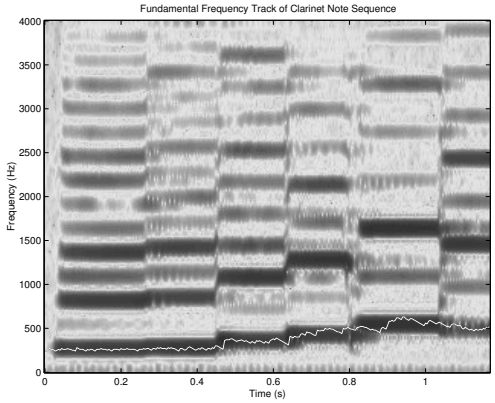

Figure 5 Fundamental Frequency of Clarinet Note Sequence

Figure 5 shows the resulting plot where the contour is given by the white line for clarity. The contour closely follows the fundamental spectral peak as it changes over time, the only inaccuracies appearing at the frequency transition points.

5. Application to Formant Tracking

The algorithm was then applied to the tracking of formants from speech signals. Two modifications were made to the search procedure. Firstly, the histogram-based component of the evaluation function was found to produce better peak emphasis if it was evaluated at the spectral frequencies rather than the frequencies of the DP search, i.e. $H(\omega)$ was used. Furthermore, a hierarchical approach was taken to finding the formants. The procedure began with a search for the first formant, then the second and finally the third. The search frequency space used for the DP algorithm was based on typical values for the location of these formants in the spectrum. The plots below shows the IFD for the Japanese word Sayonara on the left, and the tracked formants, using black lines, superimposed over the spectrogram on the right. There is a good match. However, it was found that the grid size of the DP search was an important factor in determining the accuracy of the final values and that great care had to be taken in selecting it.

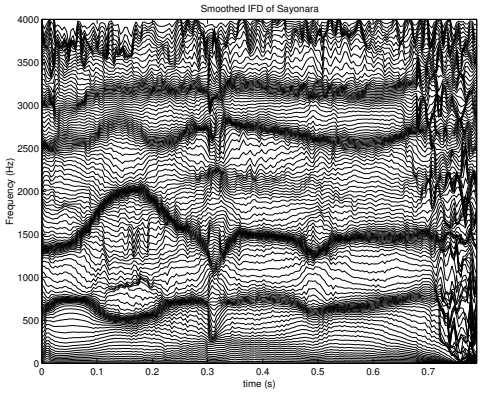

Figure 6 IFD of 'Sayonara'

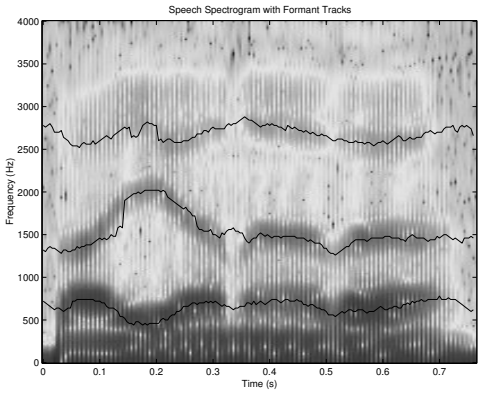

Figure 7 Spectrogram and Formant Tracks

6. Conclusion

This paper has provided a methodology to extract information from the IFD, an instantaneous frequency time-frequency representation. Examples of its performance were given with speech and audio signals. In the future it is hoped to apply this extraction technique to Timbre Morphing, which is a way of bringing two sounds together to make one new one. A vital component of this is the accurate estimation of the most perceptually relevant frequencies in the signals. Further tests of this technique may be required to determine its robustness but so far the results have been promising.

7. References

[1] P. Maragos, J.F. Kasier and T.F. Quatieri, "Energy separation in signal modulations with application to speech analysis," *IEEE Trans. Acoust. Speech Sig. Proc.*, vol. ASSP-41, no. 10, Oct. 1993, pp. 3024-3051.

[2] R. Sussman, "Analysis and re-synthesis of musical instrument sounds using energy separation," *ICASSP 1996*, pp. 997-1000.

[3] J.L. Flanagan, *Speech analysis, synthesis and perception*, Berlin: Springer-Verlag, 2nd ed., 1972.

[4] R. Kumaresan and A. Rao, "Model-based approach to envelope and positive instantaneous frequency estimation of signals with speech applications," *Journl. Acoust. Soc. Amer.*, vol. 105, no. 3, March 1999, pp. 1912-1924.

[5] D. Friedman, "Instantaneous frequency distribution vs. time: an interpretation of the phase structure of speech," *ICASSP 1985*, pp. 1121-1124.

[6] Toshihiko Abe et al, "The IF spectrogram: a new spectral representation," *Proc. ASVA 97*, pp. 423-430.

[7] Toshihiko Abe et al, "Robust pitch estimation with harmonics enhancement in noisy environments based on instantaneous frequency," *Proc. ICSLP 96*, pp. 1277-1280.

[8] T. Lysaght, D. Vernon and J. Timoney, " Subgraph Isomorphism applied to feature correspondence in Timbre Morphing," ISSC 2000, UCD, Dublin, June 2000.

[9] C. Wang and S. Seneff, "a study of tones and tempo in continuous mandarin digit strings and their application in telephone quality recognition," *Proc. ICSLP 98*, pp. 635-638.