Proceedings of the Institute of Acoustics

A TRANSFORM METHOD FOR GENERATING PERCEPTUALLY BIASED SPECTROGRAMS

M D Edgington & J A S Angus

Signal Processing: Voice and Hearing Research Group,
Department of Electronics, University of York, Heslington, York

ABSTRACT

The spectrogram is a widely used tool in speech research, although it represents signals in a very different way to the human auditory system. This paper describes a method of producing spectrograms which are biased to model some of the human auditory system features. The algorithm retains the computational advantages of a z-plane transform approach.

1. INTRODUCTION

The work described in this paper was motivated by the problem of objectively assessing the quality of synthetic speech generated by a copy synthesis procedure. Comparison of the Fourier transform spectrograms of the natural and synthetic speech is a traditional method of assessment. However spectrograms do not model many of the characteristics of the human peripheral auditory system; in particular, the spectrogram masks small temporal differences, and has a linear frequency axis, so a 10% variation of the first formant (F1) will tend to look less significant than a 5% variation in the fourth formant (F4). This is due to the linear frequency scale imposed by the DFT, which also leads to a fixed bandwidth in each bin. Furthermore there is an inherent trade-off between frequency and temporal resolution of a spectrogram. An alternative approach is to use filter bank auditory models, which tend to give a lower frequency resolution than transform based methods.

The auditory invariance principle states that the human auditory system decodes the pressure waveform to produce sets of invariant features characterising the speech signal. Thus for meaningful comparative measures between natural and synthetic speech, we must use a metric which at least approximates the gross features of the human auditory system. The low level processing of the human auditory system is performed by the peripheral auditory system (PAS), which is (arguably) well understood in comparison to the rest of the human hearing system. It is well known that the PAS does not have a linear frequency response, but rather a pseudo-logarithmic frequency response. (Zwicker and Zwicker [1] suggest a good model as linear frequency scaling up to 500Hz, and logarithmic scaling at higher frequencies.) The PAS response can be characterised by a set of *critical bands*, each of which respond to energy over a broad spectrum. The critical bands do not all have the same bandwidth, but it varies with the centre frequency of the band, in a similar way to the frequency scaling mentioned above. For centre frequencies above about 500 Hz, all bands have approximately the same Q-factor; i.e. the ratio between bandwidth and centre frequency is approximately constant.

The shape of the critical band response is discussed in Moore and Glasberg [2], and varies with the intensity of the presented sound energy, as well as the centre frequency.

The method described in this paper uses the Chirp-Z Transform to analyse the speech signal along a spiral path in the z-plane to produce a *perceptually biased spectrogram*. Consequently, the Q-factor of each bin is approximately constant. This leads to fine temporal resolution at high frequencies, and coarse temporal resolution at low frequencies. The biased spectrogram more closely represents the effects of the human auditory system. Since the Chirp-Z Transform can be precisely and efficiently implemented by FFT methods, the resulting algorithm has a computational complexity greater than but similar to that of an equivalent sized windowed FFT. The Chirp-Z Transform can perform evaluation of a limited sub-band of a signal, which allows the algorithm to bias the frequency scale in a psychoacoustically meaningful way. It is also possible to model an auditory filter response by appropriate windowing.

2. CHIRP-Z TRANSFORM

2.1 Definition of the Chirp-Z Transform

The Chirp-Z Transform algorithm is described in Rabiner et.al. [3], and a very brief overview is presented here. The representation of the z-transform can be sampled at a discrete and finite set of N points, z_k

$$X_k = X(z_k) = \sum_{n=0}^{N-1} x_n z_k^{-n} \qquad \text{for } k = 0,1,\ldots,N-1 \qquad (1)$$

A useful case is the set of points equally spaced around the unit circle, $z_k = e^{j\frac{2\pi}{N}k}$ which gives,

$$X_k = \sum_{n=0}^{N-1} x_n e^{-j\frac{2\pi n}{N}k} \qquad \text{for } k = 0,1,\ldots,N-1 \qquad (2)$$

i.e. the Discrete Fourier Transform (DFT).

Now consider a more general contour, Figure 1,

$$z_k = AW^{-k} \qquad \text{for } k = 0,1,\ldots,M-1 \qquad (3)$$

where M is an arbitrary integer, and both A and W are arbitrary complex numbers of the form $A = A_0 e^{j2\pi\theta_0}$ and, $W = W_0 e^{j2\pi\phi_0}$. Substituting Equation (3) into Equation (1), the sampled z-transform for this contour becomes,

TRANSFORM METHOD FOR BIASED SPECTROGRAMS

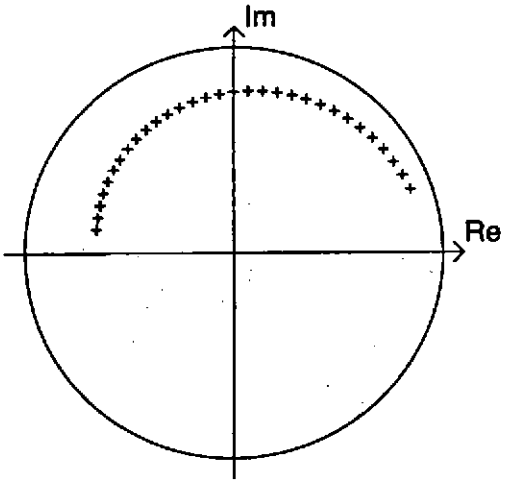

Figure 1: The Chirp-Z Transform contour in the z-plane

$$X_k = \sum_{n=0}^{N-1} x_n A^{-n} W^{nk}, \qquad \text{for } k = 0,1,\ldots,M-1 \qquad (4)$$

Equation (4) is the definition of the Chirp-Z Transform (CZT). Note that for the special case of $A = 1$, $M = N$ and $W = e^{-j\frac{2\pi}{N}}$, Equation (4) is the same as the DFT, Equation (2); thus the CZT is the more general transform.

2.2 Implemention of the Chirp-Z Transform

The CZT can be implemented efficiently if it is noted that

$$nk = \frac{n^2 + k^2 - (k-n)^2}{2}$$

from [3]. Thus Equation (4) can be written as,

$$X_k = \sum_{n=0}^{N-1} x_n A^{-n} W^{\frac{n^2}{2}} W^{\frac{k^2}{2}} W^{-\frac{(k-n)^2}{2}} \qquad (5)$$

This can be calculated in three stages,

windowing $y_n = x_n A^{-n} W^{\frac{n^2}{2}}$, for $n = 0,1,\ldots,N-1$

convolving y_n with v_m, where $v_m = W^{-\frac{m^2}{2}}$ to give g_n,

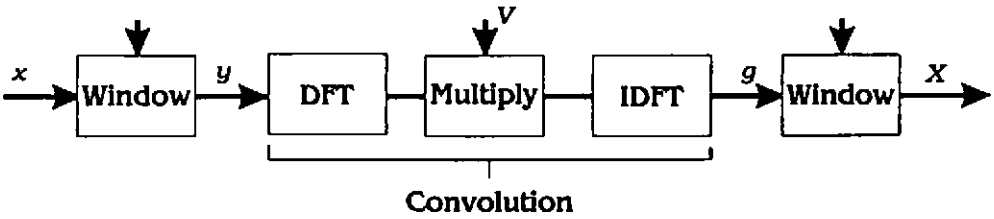

Figure 2: The Chirp-Z Transform implementation

$$g_n = \sum_{n=0}^{N-1} y_n v_{k-n} \qquad \text{for } k = 0,1,\ldots,M-1 \qquad (6)$$

windowing again, $X_k = g_k W^{\frac{k^2}{2}}$, for $k = 0,1,\ldots,M-1$

The convolution stage, Equation (6), can be performed by using the fast convolution DFT method,

$$g \otimes h = \mathcal{F}^{-1}\{\mathcal{F}(g)\mathcal{F}(h)\}$$

where $\otimes$ represents the convolution function, $\mathcal{F}$ the DFT function, and $\mathcal{F}^{-1}$ the Inverse DFT (IDFT) function.

If the analysis frame size is fixed, part of the convolution can be pre-calculated since the Fourier transform of v, $\mathcal{F}(v)$ or V, is independant of the data sequence x_n. Thus the procedure reduces to a windowing operation followed by a DFT, a further windowing operation, an IDFT, and a final windowing stage. This procedure is shown schematically in Figure 2.

An FFT has a length of N' points, where $N' = 2^i$, i is a positive integer and $N' \geq N$. The DFT used in the fast convolution calculation has a length of at least $N+M-1$ points, which will generally be no greater than $2N'$. Therefore, allowing for the IDFT, the new algorithm requires no more than four times (normally about twice) the computational effort of an FFT, with significant advantages in the flexibility of specifying the analysis contour.

3. ALGORITHM

3.1 Modelling the Auditory Filter Shape

Moore [4] and Moore and Glasberg [2] suggest that the shape of the auditory filter can be approximated to a rounded exponential function, of the form $(1+r)\exp(-r)$, where r is a function of normalised frequency. For normal speech levels, the filter response is symmetric on a linear frequency scale, while for higher intensities the response becomes very asymmetric,

with a much shallower low frequency slope, and a steeper high frequency slope. Moore and Glasberg [2] give a set of equations to predict the auditory filter shape for particular sound pressure levels (SPL), although if we assume that speech is at moderate SPL (50 - 60 dB), the filter shape can be considered constant, and fairly symmetrical. There are addiditonal practical reasons for making this assumption, mentioned below. The auditory filter shape can be simulated by applying an appropriate window to the speech signal before the CZT is performed. This window can be combined with the initial window in the CZT, so only one operation is necessary. The design of this window must also take into account the effect of the spectral zeros present in the CZT. The design of this windowing function is currently in progress.

3.2 Frequency Scaling

It is well known that human listeners do not percieve the frequency of a tone as a linear variable; a tone increasing in frequency from 100 Hz to 200 Hz will invoke a very different perception to that of a tone increasing from 1100 Hz to 1200 Hz.

Zwicker and Zwicker [1] suggest the critical band rate, Z measured in Barks, as a phsycoacoustic measure of frequency, where one Bark corresponds to one critical band width. Zwicker and Zwicker give the following approximations to map from f, (frequency in Hz) to Z.

$$Z \approx \begin{cases} 0.01f & \text{if } f < 500 \text{ Hz} \\ 9 + 4\frac{\log f}{1000 \log 2} & \text{for } f > 500 \text{ Hz} \end{cases} \tag{7}$$

The effect of mapping the frequency scale to the Bark scale is mainly noticable in the frequency resolution at different frequencies. For a typical speech bandwidth of 5KHz, these equations predict that for $f = 5$KHz, $Z \approx 18$ Bark, and for $f = 1$KHz, $Z \approx 9$ Bark. Thus half of the resolution is concentrated in the lowest 20% of the frequency range.

Unfortunately, the CZT uses an analysis contour which is linear in frequency, so Equation (7) cannot be used directly. However, the CZT allows analysis of arbitrary frequency ranges, so by splitting the desired frequency range into smaller bands, each of which uses a different frequency resolution, it is possible to closely follow the mapping of Equation (7), by selecting the output of the CZT bin closest in frequency.

For example, the analysis of a 5KHz bandwidth signal at a resolution of 0.2 Bark, requires 92 points. This range can be split into three sub-bands of about 31 points, 0 to 640 Hz, 676 Hz to 1936 Hz, and 2000 Hz to 4960 Hz. The sub-bands can be analysed with a linear frequency scale of 20 Hz, 30 Hz and 80 Hz respectively. After the CZT calculation, the data is formed into the final 92 points, by choosing the bin with a centre frequency closest to that required. Figure 3 shows the curves produced by an ideal implementation of Equation (7), and the actual implementation using the three sub-band method. The centre frequency error never exceeds 1.7% for any bin, and has an average value of 0.3%. 114 points are calculated

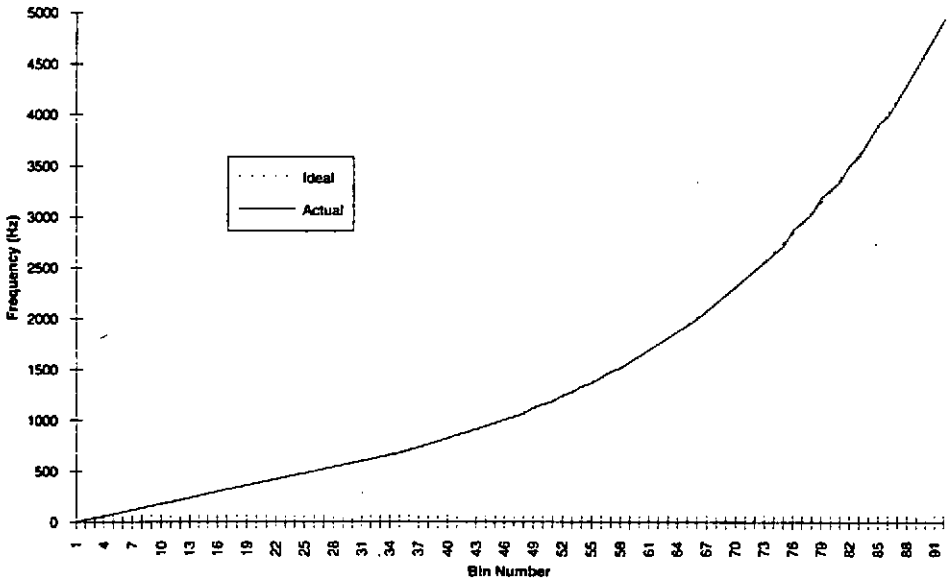

Figure 3: Comparison of 'Ideal' Bark scaling with that produced from a three sub-band linear approximation

in total, so less than 20% are discarded.

3.3 Intensity Scaling

To model the PAS more accurately, there is a requirement to consider the non-linear effects of intesity variation, as described in Moore, Chapter 2 [4]. The implications of the different models of loudness perception are beyond the scope of this paper, and are not implemented in this algorithm. Furthermore, compensation for intensity scaling would necessitate the recording of sound pressure level (SPL) with all acoustic recordings. The SPL would also need to be calibrated at every processing stage.

4. RESULTS

Figure 4 shows the frequency response of several bins of the perceptually biased spectrogram, using a linear frequency scale. The bin responses were calculated by passing a Blackman windowed sinewave through the algorithm. This was repeated for sinewaves in the range 0 to 4900 Hz at 100 Hz intervals. It can be seen that the bandwidths are greater than those predicted, due to the main lobe spreading effect introduced by the window.

Clearly there is a roll-off in the peak response of the bins. A correction factor can be applied

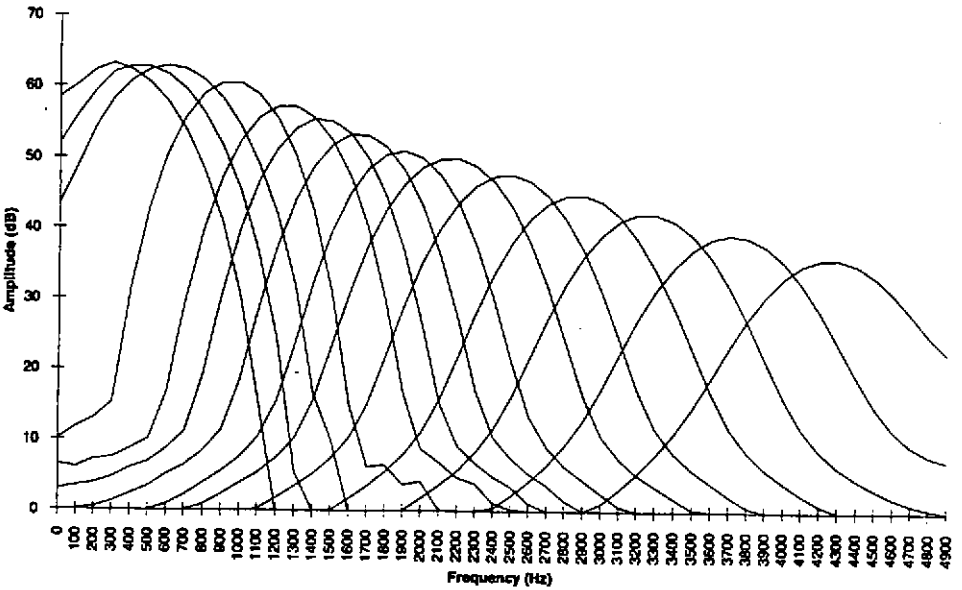

Figure 4: Frequency response of several perceptually biased transform bins

to equalise all bins' peak response, however this means that their energy response to a white noise source will not be equalised, since the width of the response increases with increasing bin centre frequency.

The frequency response of each bin is clearly affected by the locations of spectral zeros in the CZT. By splitting the frequency range into sub-bands, and evaluating each with different resolutions, a problem occurs at the joins between the sub-bands, since the CZT zeros in the lower and upper sub-bands are distributed differently, influencing the detail of the bin response in different ways. By overlapping the sub-bands, this effect is reduced.

5. CONCLUSION

The described algorithm is not intended to provide an accurate model of human auditory perception, but to provide an easy way of producing displays which retain a spectrographic quality, but are biased towards the characteristics of the PAS. This is achieved by using a Chirp-Z Transform to analysis the signal at points on a set of spiral segments in the z-plane.

Some of these points are discarded to produce appropriate frequency scaling. The algorithm is decomposed into a sequence of filtering and FFT operations, thus enabling an efficient and straightforward implementation.

6. ACKNOWLEDGEMENTS

This work was supported by the Science and Engineering Research Council and British Telecom Laboratories as part of CASE award.

REFERENCES

[1] Zwicker, E., Zwicker, U.T. "Audio Engineering and Psychoacoustics: Matching Signals to the Final Receiver, the Human Auditory System." *Journal of the Audio Engineering Society.* **39**, pp115–126. (1991)

[2] Moore, B.C.J., Glasberg, B.R. "Formulae Describing Frequency Selectivity as a Function of Frequency and Level, and their use in Calculating Excitation Patterns." *Hearing Research.* **28**, pp209–225. (1987)

[3] Rabiner, L.R., Schafer, R.W., Rader, C.M. "The Chirp-Z Transform Algorithm and its Application." *Bell Systems Technical Journal.* **48**, pp1249–1292. (1969)

[4] Moore, B.C.J. *An Introduction to the Psychology of Hearing.* Academic Press, London. (1989)