WARPED DFT AS THE BASIS FOR PSYCHOACOUSTIC MODEL

Marek Parfieniuk, Alexander Petrovsky

Bialystok Technical University, Bialystok, Poland

ABSTRACT

In this paper we consider the warped DFT as an alternative basis for psychoacoustic models. The appropriate construction of the transform is approached, aiming precise critical band power analysis. It is shown that sufficient spectral resolution can be obtained for sample block lengths several times shorter than 1024 or 2048 commonly used in the FFT based ear models. Thus temporal resolution is improved and comparable with that of more complex perceptual models utilizing filter banks. The computational complexity is estimated as acceptable for real-time processing.

1. INTRODUCTION

In the contemporary digital audio processing a crucial role is played by psychoacoustic. It is utilized both in signal coding and speech enhancement to keep inaccuracies unperceivable to the listener. The essential part of every psychoacoustically motivated systems is a model mimicking the behavior of the human auditory system analyzing sounds in nonequal critical bands (CBs). Its task is to determine the masking threshold - the power level separating strong relevant signal components from inaudible masked ones. Then the threshold controls quantization in coding or spectral subtraction / weighting in noise removal.

As the masking level depends on signal power, the first and most important step in its calculation is CB power analysis. The accuracy of this stage limits that of all subsequent ones (spreading, tonality estimation, normalization, absolute threshold checking) leading finally to the perceptual threshold. Two approaches currently compete in the CB power analysis.

The first group of the solutions is inspired by Johnston's famous article [1]. The idea consists in computing the FFT of windowed signal segment and then partitioning transform coefficients into groups corresponding to CBs of hearing. The sum of squared magnitudes of the group elements gives an estimate of according CB power. The attainment of reasonable spectral resolution in narrowest CBs requires using the FFT with rather long time window. Thus conceptual simplicity and efficiency are leveled by poor temporal resolution not sufficient to deal with finer phenomena such as pre-masking [2].

The second class was invented to eliminate this drawback, exploiting nonuniform filter bank to decompose signal. Then short term power spectral density is calculated on the frame of subband coefficients. The main shortcoming is a general complexity, especially if good CB approximation is of interest and an efficient tree-structured filter banks are not sufficient.

This work was supported by Bialystok Technical University under the grant W/WI/3/02.

No solution definitely outperforms the other, so both of them have their own applications. For example, the two coexist in quite recent standard for Perceptual Evaluation of Audio Quality - PEAQ (ITU-R Recommendation BS.1387). The efficient but simplified "Basic" version of PEAQ exploits the FFT, whereas the precise "Advanced" variant can use filter bank as well.

Our proposition is an extension of the first approach, exploiting recent advances in warped spectral analysis. Namely, we suggest that low dimensional warped DFT studied recently by Mitra [3] can successfully replace the FFT of long sample block. This is possible as warped transform allows allocating its frequency samples in accordance with CB distribution. Thus both good spectral and temporal resolutions can be reconciled in the WDFT based psychoacoustic model.

2. DFT AS TRADITIONAL BASIS FOR EAR MODELS

In the FFT based ear model, the transformation of the signal from the frequency domain to the CB (or Bark) domain reduces to the appropriate grouping of the half of the transform values in accordance with well known critical bandwidths [1]. The partitioning is determined for given transform size and sampling frequency. Two representative examples are given in Table 1.

Several evident observations can be made about these tabulated data. As the FFT has uniform spectral resolution whereas critical bandwidths strongly vary with frequency, very different numbers of bins are assigned to particular CBs. In the part A of the table, group quantities vary from 6 to 224, and from 3 to 18 in the part B. Thus wide, high frequency CBs supported with many coefficients are obviously preferred. Power estimation, as well as tonality measurement, is more accurate. In turn, lower CBs having only few data, are treated very superficially. On the contrary, all CBs are equivalent from auditory point of view, so they all should be analyzed with similar precision in perceptual domain. Even paradoxically, lower bands can be regarded as more significant as speech spectrum spans 0 .. 4 kHz and ear is more sensitive in this range. The eighteen groups corresponding to this range comprise of only ~280 of all 1024 bins of the half FFT.

The significant consequence of the above facts is that only the FFT of rather large size can ensure reasonable accuracy (at least several frequency samples) in lower subbands. The computational load is not much a barrier having fast transform - the widening of time window is a real problem. For common analysis segment sizes of 1024 and 2048, corresponding time slots are 20 - 40 ms. Such a time resolution is sufficient to deal with coarser phenomena e.g. post-masking lasting even hundreds of milliseconds. But it is too big if rapid events (such as pre-masking) are modeled, having period of several milliseconds.

The only solution is to find an alternative means of transition to the Bark domain.

Table 1: Typical mappings from FFT bins to Critical Bands.

	A (FFT size = 2048, Fs = 32 kHz)			B (FFT size = 256, Fs = 8 kHz)		
CB	Bin range	No.	Freq. [Hz]	Bin range	No.	Freq. [Hz]
1	1 - 6	6	16 - 94	1 - 3	3	31 - 94
2	7 - 12	6	109 - 188	4 - 6	3	125 - 188
3	13 - 19	7	203 - 297	7 - 9	3	219 - 281
4	20 - 25	6	313 - 391	10 - 12	3	313 - 375
5	26 - 32	7	406 - 500	13 - 16	4	406 - 500
6	33 - 40	8	516 - 625	17 - 20	4	531 - 625
7	41 - 49	9	641 - 766	21 - 24	4	656 - 750
8	50 - 58	9	781 - 906	25 - 29	5	781 - 906
9	59 - 69	11	922 - 1078	30 - 34	5	938 - 1063
10	70 - 81	12	1094 - 1266	35 - 40	6	1094 - 1250
11	82 - 94	13	1281 - 1469	41 - 47	7	1281 - 1469
12	95 - 110	16	1484 - 1719	48 - 55	8	1500 - 1719
13	111 - 128	18	1734 - 2000	56 - 64	9	1750 - 2000
14	129 - 148	20	2016 - 2313	65 - 74	10	2031 - 2313
15	149 - 172	24	2328 - 2688	75 - 86	12	2344 - 2688
16	173 - 201	29	2703 - 3141	87 - 100	14	2719 - 3125
17	202 - 236	35	3156 - 3688	101 - 118	18	3156 - 3688
18	237 - 281	45	3703 - 4391	119 - 128	10	3719 - 4000
19	282 - 339	58	4406 - 5297			
20	340 - 409	70	5313 - 6391			
21	410 - 492	83	6406 - 7688			
22	493 - 608	116	7703 - 9500			
23	609 - 768	160	9516 - 12000			
24	769 - 992	224	12016 - 15500			
25	993 - 1024	32	15516 - 16000			

3. FUNDAMENTALS OF WARPED DFT

3.1. General WDFT definition

The idea of spectrum warping is not novel. It was considered in the 70s by Oppenheim *et al.* in the context of nonuniform spectral analysis [4]. The concept was to pass the analyzed signal through allpass chain to achieve an auxiliary sequence having deformed spectral contents. Then pure DFT of this sequence represents the warped spectrum of the input signal. The corresponding processing schema is shown in Fig. 1.

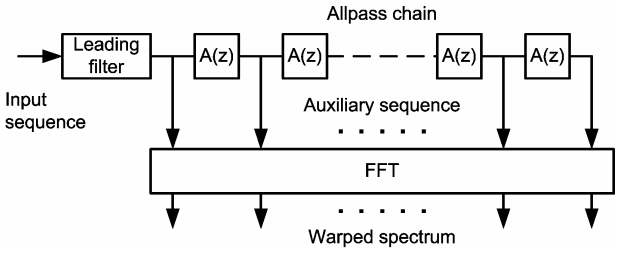

Fig. 1. Warped spectrum via allpass chain preprocessed FFT

The theory behind this is known as Laguerre expansion decomposing signal into infinite set of infinite-length Laguerre sequences. As obvious truncations must be done, a certain error is unavoidable, resulting in only approximated warped spectra obtained this way.

The subject became alive again in the 90s when the ideas of warped or allpass frequency transformed filter banks, wavelets and linear prediction appeared. Eventually, in recent years, Mitra *et al.* has drawn the concepts of two warped transforms - Warped Discrete Fourier Transform [3] and Warped Cosine Transform.

The Warped Discrete Fourier Transform (WDFT) of the sequence $x[n]$ of N points is defined as

$$\hat{X}(z_k) = X(\hat{z}_k) = \sum_{n=0}^{N-1} x[n]\hat{z}_k^{-n} \qquad k = 0..N-1 \qquad (1)$$

where $\hat{z}_k$ are the images of equidistant points of the unit circle in the z plane, resulting from the transformation

$$z_k^{-1} = e^{-j\frac{2\pi k}{N}} \quad \rightarrow \quad \hat{z}_k^{-1} = A(z_k) \qquad k = 0..N-1 \qquad (2)$$

with arbitrary order allpass function $A(z)$.

Bilinear mapping of the z plane into new warped $\hat{z}$ plane is done this way. The points uniformly distributed on the unit circle in the first plane still lie on the unit circle in the second plane but they become unequally spaced. This is explained in Fig. 2.

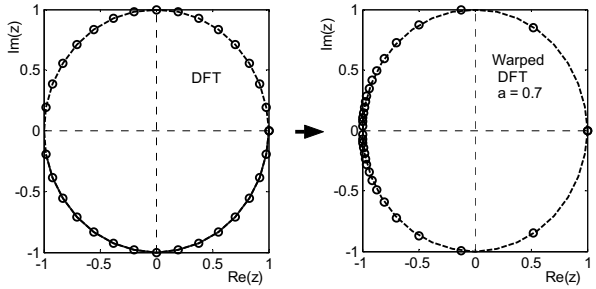

Fig. 2. Locations of DFT and WDFT frequency samples

Thus the WDFT is a generalization of the DFT, with frequency samples allocated nonuniformly but regularly over the unit circle. From the other point of view the WDFT is a special case of the most general Nonuniform Discrete Fourier Transform (NDFT) allowing sampling z transform at distinct but arbitrary selected points on the z plane.

In matrix notation (with $\hat{X}[k]$ denoting $\hat{X}(z_k)$), the WDFT can be represented as

$$\begin{bmatrix} \hat{X}[0] \\ \hat{X}[1] \\ \vdots \\ \hat{X}[N-1] \end{bmatrix} = \begin{bmatrix} 1 & A(z_0) & \cdots & A(z_0)^{N-1} \\ 1 & A(z_1) & \cdots & A(z_1)^{N-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & A(z_{N-1}) & \cdots & A(z_{N-1})^{N-1} \end{bmatrix} \begin{bmatrix} x[0] \\ x[1] \\ \vdots \\ x[N-1] \end{bmatrix} \qquad (3)$$

where the transform matrix incorporating allpass transformation is still the Vandermonde matrix as for the conventional DFT. The determinant of such a matrix is non-zero for different z_k. Thus the invertibility of the transform is guaranteed, although the problem can be ill-conditioned.

The WDFT inherits some properties of the DFT. An important one is its conjugate symmetry for real data

$$\hat{X}[N-k] = \hat{X}^*[k] \qquad (4)$$

3.2. WDFT with first-order allpass

The simplest variant of the WDFT is that based on real coefficient first order allpass [3]

$$z^{-1} \quad \rightarrow \quad A(z) = \frac{-a + z^{-1}}{1 - az^{-1}} \tag{5}$$

The coefficient can potentially be a complex number, but the corresponding mapping rotates coordinate system what is generally not desired. Stability requires $|a| < 1$. The characteristic properties of this variant of the WDFT are monotony and uniqueness of frequency mapping, what is not true for higher order allpasses [3]. Depending of the sign of a, low or high frequency range is stretched whereas the remaining part of the unit circle becomes compressed. Formally, this can be expressed as

$$\hat{\omega} = \omega + 2\arctan\left(\frac{a\sin\omega}{1 - a\cos\omega}\right) \qquad \text{for} \quad \begin{cases} z = e^{j\omega} \\ \hat{z} = e^{j\hat{\omega}} \end{cases} \tag{6}$$

4. WDFT IN CB POWER ESTIMATION

4.1. Selection of allpass function and its coefficients

The first step in employing the WDFT in psychoacoustic model is the design of the appropriate allpass transformation. The frequency samples of the z transform should be established uniformly in the perceptual domain.

Table 2: Mappings from WDFT bins to Critical Bands.

	A			B		
	(WDFT size = 256, Fs = 32 kHz)			(WDFT size = 256, Fs = 8 kHz)		
CB	Bin range	No.	Freq. [Hz]	Bin range	No.	Freq. [Hz]
1	1 - 4	4	22 - 86	1 - 7	7	13 - 92
2	5 - 9	5	108 - 195	8 - 15	8	105 - 198
3	10 - 13	4	217 - 283	16 - 22	7	212 - 294
4	14 - 18	5	305 - 395	23 - 29	7	308 - 394
5	19 - 23	5	417 - 509	30 - 36	7	408 - 498
6	24 - 28	5	533 - 628	37 - 44	8	514 - 627
7	29 - 33	5	653 - 752	45 - 52	8	644 - 768
8	34 - 39	6	778 - 910	53 - 59	7	786 - 904
9	40 - 45	6	937 - 1079	60 - 67	8	925 - 1080
10	46 - 51	6	1109 - 1264	68 - 74	7	1103 - 1255
11	52 - 57	6	1297 - 1469	75 - 81	7	1282 - 1458
12	58 - 63	6	1506 - 1700	82 - 88	7	1489 - 1694
13	64 - 69	6	1741 - 1964	89 - 95	7	1731 - 1972
14	70 - 75	6	2012 - 2272	96 - 102	7	2015 - 2301
15	76 - 81	6	2329 - 2642	103 - 109	7	2352 - 2688
16	82 - 87	6	2711 - 3095	110 - 116	7	2748 - 3134
17	88 - 93	6	3182 - 3671	117 - 123	7	3202 - 3629
18	94 - 98	5	3782 - 4286	124 - 128	5	3703 - 4000
19	99 - 104	6	4428 - 5269			
20	105 - 108	4	5469 - 6146			
21	109 - 113	5	6402 - 7615			
22	114 - 117	4	7973 - 9209			
23	118 - 122	5	9680 - 11886			
24	123 - 127	5	12517 - 15277			
25	128 - 128	1	16000 - 16000			

Thus far the problem of perceptual frequency warping was considered only in one but very exhaustive study [5] aiming the design of auditory filters. There was shown that first order allpass is sufficient to well approximate the perceptual Bark and ERB (Equivalent Rectanngular Bandwidth) scales. Based on the optimization results, the formula

$$a_{Bark} = 0.1957 - 1.048\left[\frac{2}{\pi}\arctan\left(0.07212\frac{f_s}{1000}\right)\right]^{\frac{1}{2}} \tag{7}$$

was pointed out as directly giving allpass coefficients appropriate for given sampling frequency. For two common frequencies 8 and 32 kHz, it gives a equal -0.4092 and -0.7056 accordingly.

Due to the same warping mechanism, these results can be directly applied to the WDFT, leading to the organization of the bins shown in Table 2. It is evident that all CB have assigned comparable numbers of the transform bins. No band is preferred. Referring to Table 1, it can be noted that the same resolution in first Barks can be achieved for the warped transforms shorter several times than the FFT. At the same size, the measures of spectral flatness and tonality can be more accurate, owing to the z-transform sampling matching the Bark scale.

4.2. Power compensation

The WDFT in its original form does not preserve signal power in corresponding parts of the unit circle before and after allpass transformation. As warping causes one frequency range to stretch but the other to narrow, it should be supported with appropriate scaling of the WDFT magnitude to retain correct power level in each CB.

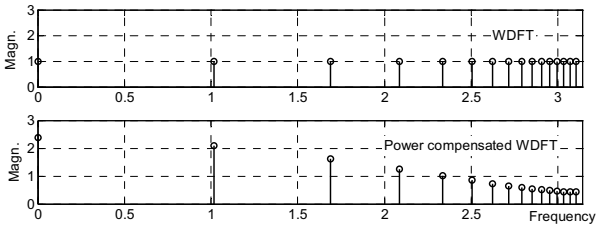

Fig. 3. Power compensation of WDFT (a=0.7) of the unit pulse

Beginning with the desired signal power equality for an arbitrary sector $\langle z_{\min}, z_{\max} \rangle$ of the unit circle and its image $\langle \hat{z}_{\min}, \hat{z}_{\max} \rangle$

$$\frac{1}{2\pi i}\int_{\hat{z}_{\min}}^{\hat{z}_{\max}} |X(\hat{z})|^2 \frac{d\hat{z}}{\hat{z}} = \frac{1}{2\pi i}\int_{z_{\min}}^{z_{\max}} |X(z)|^2 \frac{dz}{z} \tag{8}$$

we can convert the left side to be

$$\int_{\hat{z}_{\min}}^{\hat{z}_{\max}} |X(\hat{z})|^2 \frac{d\hat{z}}{\hat{z}} = \int_{z_{\min}}^{z_{\max}} \left|\left[\frac{(1-a^2)z}{(1-az)(z-a)}\right]^{0.5} X\left(\frac{z-a}{1-az}\right)\right|^2 \frac{dz}{z} \tag{9}$$

This states that the compensation factor must be the square root of the expression in the square brackets above. As only the magnitude is of interest in the power estimation and z is a root of unity, we can neglect phase and further reduce normalization factor to express the power corrected WDFT as

$$\hat{X}_{PC}(z) = \frac{\sqrt{1-a^2}}{1-az}\hat{X}(z) \tag{10}$$

This result is similar to that obtained in [6] where the power compensation of the first order allpass chain preprocessed FFT is approached by means of additional leading filter (see Fig. 1).

4.3. Computational complexity

It seems that algorithms of efficiency comparable to that of the FFT, can not be constructed for the WDFT, due to the

asymmetry of the WDFT matrix. However, the algorithm with direct complex matrix multiplication can be highly optimized.

Currently, the most advanced algorithm (although still of complexity $O(N^2)$) was proposed in [3]. It exploits the factorization of the WDFT matrix into the product of real, the DFT (implemented with the FFT) and complex diagonal matrices. This method is well suited for image processing where all data come at the same time.

In audio processing, where samples come one after another, even the direct realization can be used. If we take into consideration the WDFT symmetry for real data (4), then (3) can be rewritten as

$$\begin{bmatrix} \hat{X}[0] \\ \hat{X}[1] \\ \vdots \\ \hat{X}[N-1] \end{bmatrix} = \sum_{n=0}^{N-1} \begin{bmatrix} A(z_0)^n \\ A(z_1)^n \\ \vdots \\ A(z_{N-1})^n \end{bmatrix} x[n] \qquad (11)$$

Each term in this summation is related to only one input sample. It can be computed when the sample came and accumulated to give the result after N samples. The computational load per input sample is $O(N)$ and it is similar to that in the preprocessed FFT approach, where the chain of N allpasses must be recalculated for each input sample.

5. EXPERIMENTAL RESULTS

To check the relevance of the proposed approach, several preliminary experiments of calculating masking threshold with different bases was done for music signals. Here only one, more representative is shown. Namely, CB power analysis of castanets type signal sampled at 32 kHz (shown in Fig. 4) was performed by means of the WDFT, the allpass chain preprocessed FFT and the pure FFT. The Hann window was applied in all cases. The FFT size and partitioning were as in Table 1-A. The overlap was set at 50%. The warped solutions were of the size of 256, configured as described in Sec. 4.1 and 4.2. In this case, no overlap was used.

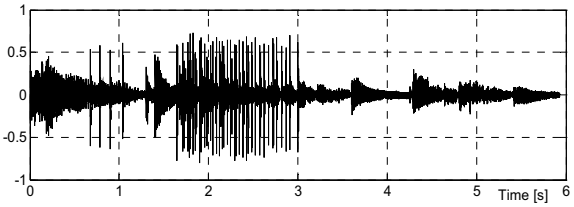

Fig. 4. Analyzed audio signal - castanets

The resulting power spectrograms are shown in Fig. 5. The warped solutions of the transform size eight times shorter than the FFT give very clear analysis results. In lower CBs (2 - 10) power estimation seems more accurate, though there are fewer coefficients in corresponding transform subsets. Power estimation for the preprocessed FFT is very similar to that due to the WDFT, though image is smeared. One could expect that it can serve as sufficiently good approximate of the exact WDFT in some cases.

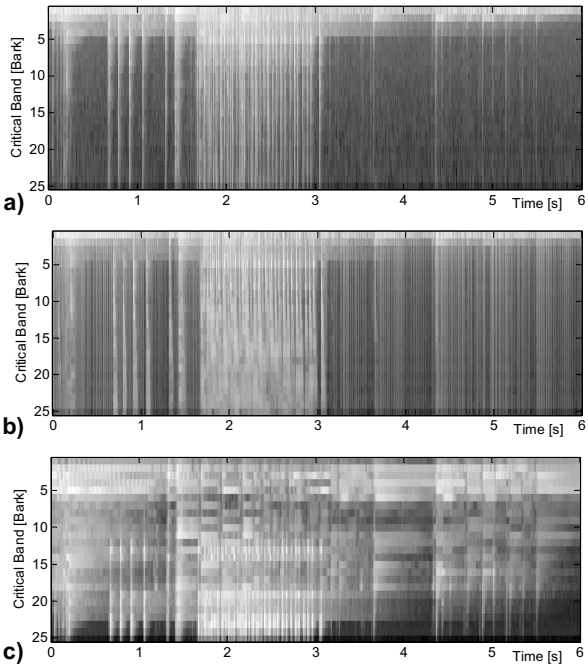

Fig. 5. Bark power spectrogram obtained using: a) WDFT, b) allpass chain preprocessed FFT, c) FFT

6. CONCLUSIONS

The presented facts indicate that the WDFT can really serve as the basis for psychoacoustic models. Preliminary experimental results seem promising. The warped transform, with frequency samples allocated in accordance with perceptual scale, outperforms the FFT, however at the cost of increased complexity. For full evaluation of the new approach, it should be compared with those based on filter banks, and applied in practical psychoacoustic system of coding or enhancement. The works have been started and results are expected soon.

7. REFERENCES

[1] J.D. Johnston, "Transform coding of audio signals using perceptual noise criteria," *IEEE J. Selected Areas in Comm.*, vol. 6, pp. 314-323, Feb. 1988.

[2] T. Thiede, E. Kabot, "A New Perceptual Quality Measure for Bit Rate Reduced Audio," *Proc. 100th AES Convention*, Copenhagen, Preprint 4280, 1996.

[3] A. Makur, S.K. Mitra, "Warped Discrete-Fourier Transform: Theory and Applications," *IEEE Trans. Circuits Systems I*, vol. 48, pp. 1086-1093, Sept. 2001.

[4] A.V. Oppenheim, D.H. Johnson, K. Steiglitz, "Computation of spectra with unequal resolution using the FFT," *Proc. IEEE*, vol. 59, pp. 299-301, Feb. 1971.

[5] J.O. Smith III, J.S. Abel, "Bark and ERB Bilinear Transforms," *IEEE Trans. Speech, Audio Processing*, vol. 7, pp. 697-708, June 1999.

[6] T. von Schroeter, "Frequency Warping with Arbitrary Allpass Maps," *IEEE Signal Processing Letters*, vol. 6, pp. 116-118, May 1999.