Rejoinder to M. Wagner's 2024 Letter to the Editor of the Australian Speech and Technology Association

Frantz Clermont

Speech and Language Laboratory,
School of Culture, History and Language,
Australian National University, Canberra, Australia
dr.fclermont@gmail.com

Abstract

We thank M. Wagner for his critique of our paper "Linear transformation from full-band to sub-band cepstrum" (in *Proc. 18th Australasian International Conference on Speech Science and Technology (SST2022)*). We also thank the Executive Committee of the SST2024 Conference for the invitation to respond. This rejoinder considers four central points from the critique as follows: (i) The "sub-band cepstrum" approach dating to the 1990s is recalled and contrasted with our 2022 approach; (ii) The spectral representation directly relevant to our approach is highlighted; (iii) Our mathematical formulation is further justified; (iv) Our choice of certain terms is discussed.

Index Terms: Cepstrum, full band, sub-band, band-limited, linear transformation, Fourier series.

1. On the "sub-band cepstrum"

The impetus for sub-band processing of speech arose from psychoacoustic work [1] in the 1950s and [2] in the 1990s, both postulating that the human auditory system decodes the linguistic message separately in different spectral sub-bands. It is this frequency-local mechanism which inspired the idea of increasing robustness in automatic speech recognition by de-emphasising noise-corrupted sub-bands. Its implementation promoted [3-4] also in the 1990s, involves splitting the full-band's frequency range into multiple sub-bands, extracting their own acoustic features independently, and then building statistical models to detect the noise-affected sub-bands.

Fig. 1(a) gives a schematic description of this sub-band approach applied to the cepstrum. We refer to it as the existing approach. For any frame of the speech signal, standard spectral analysis (via FFT or filter banks) is performed to obtain the log-magnitude spectrum (LMS) over the full band. Sub-bands are selected by matching lower and upper limits $[\omega_1, \omega_2]$ with the closest frequency bins of the corresponding LMS sub-regions.

The Discrete Cosine Transform (DCT) is a standard step [5: p 642] for converting either the entire LMS into the full-band cepstrum, or a sub-region of the LMS into the "*sub-band cepstrum*". This term was first coined [5: p. 642] in 1998 as far as we can ascertain, then mentioned [6: Fig. 1] in 1999 and [7: p. 242] in 2000. The DCT outputs in Fig. 1(a) are vectors of sub-band cepstral coefficients (CCs) whose size must be finite in practice. There are still no definite recommendations for determining the appropriate vector size per sub-band.

Fig. 1(b) gives a schematic description of our approach [8], which also seeks to access local spectral information using the sub-band cepstrum. The major difference is that our sub-band CCs are derived from full-band CCs via a linear transformation.

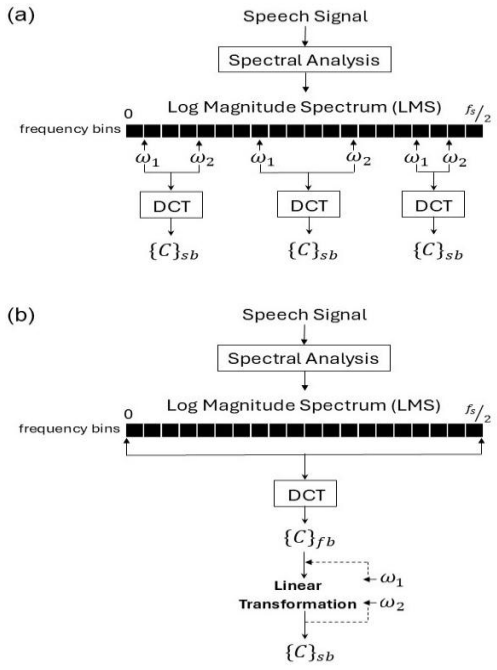

Figure 1: *Two approaches for generating sub-band cepstral coefficients (CCs): (a) the existing approach dating back 30 years; (b) our 2022 approach. Notations: {C} represents a vector of CCs; the subscripts "sb" and "fb" indicate sub-band and full band, respectively; "fs" denotes the signal's sampling frequency; the full band ranges from 0 to fs/2 (Hz) or from 0 to π (radians); ω_1 and ω_2 are parameters for the lower and upper frequency limits of a sub-band.*

It is worth noting the flexibility and efficiency advantages in our approach: (a) The DCT operation is performed once to obtain the full-band CCs; (b) These are re-used every time a new sub-band is selected; (c) The parametric formulation of the linear transformation enables the easy selection of any sub-band's frequency limits; (d) The minimum size for a vector of sub-band CCs depends on the fraction of the full-band's frequency range occupied by a sub-band's width [8: p. 137]. In Section 2, we explain our motivation for qualifying our sub-band CCs as "band-limited".

2. On our mathematical formulation

Section 2.3 of the critique argues that the mathematical basis of our approach is incorrect. We beg to differ and, in response, we recall our key equations, clarify their respective roles, provide a verification of their validity, and discuss the choice of certain terms. Note that the recalled equations carry their original numbers and, therefore, they appear below neither in sequential order nor necessarily in the same order as they do in [8].

2.1. Spectral representations and terminology

The LMS, denoted for example as $log|X(\omega)|^2$, is our starting point as shown in Fig. 1a) from the critique and in Fig. 2(a) below. Our focus however is the cepstral representation of the LMS formulated in Eq (1), where $S(\omega) \stackrel{\text{def}}{=} log|X(\omega)|^2$ is an *even* function of ω for real signals such as speech. That is, all necessary spectral information is already contained within the *positive* frequencies. The LMS can thus be expanded [9: Ch. 7; 10: p. 429; 11: p. 163] as a Fourier-cosine series of CCs, hereafter called the full-band C_k across $[0, \pi]$.

The cepstrally-smoothed spectrum (or envelope) based on Eq. (1) is overlaid on the full-band LMS in Fig. 2(a). Its shape is determined by the $C_{k>0}$, and its smoothness results from truncating the cosine series after M terms. The zeroth-order coefficient $C_{k=0}$ over the full band is usually excluded.

$$S(\omega) \cong \sum_{k=1}^{M} C_k \cos(k\omega), \quad 0 \le \omega \le \pi \tag{1}$$

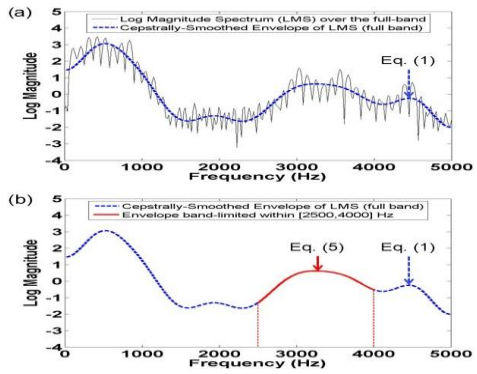

Figure 2: *(a) Overlaid: LMS obtained using FFT (black solid line), and cepstrally-smoothed envelope over the full band [0, 5000] Hz based on Eq. (1) and* $M = 14$ *(dashed blue line); (b) Overlaid: same cepstrally-smoothed envelope as in (a), and a "band-limited" (red solid line) section of the full-band envelope.*

Fig. 2(b) gives a visual illustration to clarify our approach. The marked sub-band interval is seen to contain a *band-limited section* of the *full-band envelope*, from which follows our assumption of a relation between sub-band and full-band CCs. This contrasts with the existing approach which generates sub-band CCs by applying the DCT to every selected sub-region of the LMS without direct reference to the full-band domain.

We have thus chosen the term *band-limited* CCs (or BLCCs) to reflect our shift of perspective from the existing approach, and proposed Eq. (5) to represent $S(\omega)$ over a sub-band interval $[\omega_1, \omega_2]$ in a manner analogous to Eq. (1).

$$S(\omega(\omega')) \cong C'_0 + \sum_{l=1}^{N} C'_l \cos(l\omega'), \quad 0 \le \omega' \le \pi \tag{5}$$

Our sub-band CCs are the coefficients C'_l of the series in Eq. (5), where $C'_{l=0}$ accounts for a possibly non-zero, average level of a sub-band. The series' upper bound N is set at about $M \times W$ (or MW in short), where M is the vector size for the full-band C_k and W the ratio of the sub-band's width to the full frequency range. Sections 4.2 and 4.3 of [8] give empirical evidence that MW-truncated series can only approximate the shape of the spectral envelope within sub-bands. This is why we have stated that our sub-band cepstrum is "estimated" from the full-band cepstrum, albeit with sufficient accuracy for practical purposes.

2.2. Change of variables $\omega \to \omega'$

Our mathematical problem involves two frequency intervals: $[\omega_1, \omega_2]$ for the sub-band and $[0, \pi]$ for the full band, the former being interior to the latter. To obtain our Fourier-cosine series representation on $[\omega_1, \omega_2]$, a change of variables $\omega \to \omega'$ is introduced and a one-to-one mapping is established between the ω-axis and the ω'-axis. This is a standard procedure [12: pp. 82-85; 13: pp. 286-287] for handling intervals other than $[0, \pi]$.

$$\omega' = \pi \left[\frac{(\omega - \omega_1)}{(\omega_2 - \omega_1)}\right] \ni \omega = [\omega_1, \omega_2] \to \omega' = [0, \pi] \tag{2}$$

$$\omega = \omega_1 + \left[\frac{(\omega_2 - \omega_1)}{\pi}\right] \omega' \ni \omega' = [0, \pi] \to \omega = [\omega_1, \omega_2] \tag{3}$$

Note that ω' stretches the sub-band interval $[\omega_1, \omega_2]$ to the full range $[0, \pi]$ and, by writing Eq. (2) as Eq. (3), ω is recast as the sub-band dependent function $\omega(\omega')$. Our sub-band C'_l are then derived via the (well-established) inverse Fourier-cosine formulae in Eqs (6) and (7), where $S\big(\omega(\omega')\big)$ is replaced by the full-band expansion in Eq. (1) and ω by Eq. (3):

$$C'_{l=0} = \frac{1}{\pi} \int_0^{\pi} S\big(\omega(\omega')\big)\, d\omega' \tag{6}$$

$$C'_{l>0} = \frac{2}{\pi} \int_0^{\pi} S\big(\omega(\omega')\big) \cos(l\omega')\, d\omega' \tag{7}$$

2.3. Mathematical verification

Evaluating Eqs (6)-(7) yields Eq. (10) where sub-band and full-band CCs are related to one another through a weighted linear sum or a linear transformation in matrix form. The weight vectors a_{lk} in Eqs (11a)-(11c) are trigonometric functions of ω_1 and ω_2:

$$C'_l = \sum_{k=1}^{M} a_{lk} \cdot C_k, \quad l = 0, 1, \ldots, N \tag{10}$$

$$a_{lk,\, l>0,\, l \neq kW} = \beta_{lk}[(-1)^{l+1} \sin(k\omega_2) + \sin(k\omega_1)] \tag{11a}$$

$$a_{lk,\, l>0,\, l=kW} = \cos(k\omega_1) \tag{11b}$$

$$a_{lk,\, l=0} = \gamma_k[\sin(k\omega_2) - \sin(k\omega_1)] \tag{11c}$$

where $\beta_{lk} = \frac{2(kW)}{\pi[l^2 - (kW)^2]}$, $\gamma_k = \frac{1}{k(\omega_2 - \omega_1)}$, $W = \left[\frac{(\omega_2 - \omega_1)}{\pi}\right]$

It can easily be verified that $C'_l = C_k$ for $[\omega_1, \omega_2] = [0, \pi]$.

3. Summary

We agree that the context of our SST2022 paper [8] would have been more complete with relevant literature on the sub-band cepstrum. We have therefore made a diligent attempt to rectify this in Section 1 above. We have also paid attention to central points raised against our mathematical formulation. We believe that the formulation of our sub-band approach continues to hold in the light of additional information offered in this response.

4. Acknowledgements

We thank the reviewers for their comments and suggestions.

5. References

[1] H. Fletcher, *Speech and Hearing in Communication*, (New York: Krieger), 1953.

[2] J.B. Allen, "How do humans process and recognize speech?" *IEEE Trans. on Speech & Audio Processing*, 1994, pp. 567-577.

[3] H. Hermansky, S. Tibrewala and M. Patel, "Towards ASR on partially corrupted speech," in *Proc. 4th Int. Conf. on Spoken Language Processing (ICSLP 96)*, 1996, pp. 1579-1582.

[4] S. Tibrewala and H. Hermansky, "Sub-band based recognition of noisy speech", in *Proc. Int. Conf. on Acoustics, Speech & Signal Processing*, 1997, pp. 1255-1258.

[5] S. Okawa, E. Bocchieri and A. Potamianos, "Multi-band speech recognition in noisy environments", in *Proc. Int. Conf. Acoustics, Speech & Signal Processing*, 1998, pp. 641-644.

[6] K. Yoshida, K. Takagi and K. Ozeki, "Speaker identification using subband HMMS", in *Proc. 6th European Conf. Speech Communication & Tech. (Eurospeech'99)*, 1999, pp. 1019-1022.

[7] P.M. McCourt, S.V. Vaseghi and B. Doherty, "Multi-resolution sub-band features and models for HMM-based phonetic modelling", *Computer Speech & Language*, 2000, pp. 241-259.

[8] F. Clermont, "Linear transformation from full-band to sub-band cepstrum," in, *Proc.18th Australasian Int. Conf. on Speech Science & Technology*, 2022, pp. 136-140.

[9] L.R. Rabiner and R.W. Schafer, *Digital Processing of Speech Signals*, (New Jersey: Prentice-Hall), 1978.

[10] K. Shikano and F. Itakura, "Spectrum distance measures for speech recognition, in S. Furui and M.M. Sondhi (Eds), *Advances in Speech Signal Processing*, (New York: Marcel Dekker), 1992, pp. 419-451.

[11] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*, (New Jersey: Prentice-Hall), 1993.

[12] J.W. Brown and R.V. Churchill, *Fourier Series and Boundary Value Problems (5th ed.)*, (New York: McGraw-Hill), 1993.

[13] J. Makhoul, "Spectral linear prediction: Properties and applications", *IEEE Trans. on Acoustics, Speech & Signal Processing*, 1975, pp. 283-296.