The sub-band cepstrum as a tool for local spectral analysis in forensic voice comparison

Shunichi Ishihara and Frantz Clermont

Speech and Language Laboratory, Australian National University, Canberra, Australia
shunichi.ishihara@anu.edu.au, dr.fclermont@gmail.com

Abstract

This paper exploits band-limited cepstral coefficients (BLCCs) in forensic voice comparison (FVC), with the primary aim of locating speaker-sensitive spectral regions. BLCCs are sub-band cepstral coefficients (CCs) which are easily obtained by a linear transformation of full-band CCs. The transformation gives the flexibility of selecting any sub-band region without the recurrent cost of spectral analyses. Using multi-band BLCCs obtained by sliding a 600-Hz sub-band every 400 Hz across the full [0-5kHz] range, FVC experiments were attempted using citation recordings of the 5 Japanese vowels from 297 adult-male, native speakers. The FVC results give locations and ranges for the most speaker-sensitive sub-bands, and show that combining 3-4 of these yields comparable FVC performance with full-band CCs. Owing to their ability to easily extract locally-encoded speaker information from full-band CCs, it can be conjectured that BLCCs have a significant role to play in the search for meaningful interpretations of the numerical outcome of forensic analyses.

1 Introduction

In forensic voice comparison (FVC), the forensic scientist typically needs to compare a pair of speech recordings: the source-questioned and source-known samples, and to obtain the strength of evidence quantified by a likelihood ratio (LR).

For this purpose, it has become standard practice to parameterise the acoustic speech signal using low-dimensional vectors of cepstral coefficients (CCs). These are automatically extracted from any phonetic segments, and have been shown to be effective for speech and speaker classification. The effectiveness is attributable to the ability of low-ordered CCs to produce cepstrally-smoothed spectra with reduced sensitivity to "noninformation bearing variabilities" (Rabiner and Juang 1993: 169) and, thus, with increased distinctiveness. Such spectra may be obtained with full-band CCs which yield spectral representations over the full frequency range, or with sub-band CCs which give access to local regions within the full range.

Consistent with our long-term goal of interpreting the FVC outcome beyond numerical LR values, the present study focuses on sub-band CCs with the dual aim of (a) locating vowel spectral regions that are most sensitive to speaker differences, and (b) determining the extent to which such regions affect LR values compared to the full band from vowel to vowel. The motivation for this endeavour stems from an old premise (Peterson 1959: 151) that speaker information is not uniformly encoded throughout vowel spectra, i.e., there exist local regions of strong speaker and phonetic specificity. Supportive evidence has since been reported in a wide range of studies (*inter alia*: Goto et al. 2017; Hyon et al. 2012; Khodai-Joopari et al. 2004; Kitamura and Akagi 1995; Mohammadi et al. 2011; Mokhtari and Clermont 1994; Pols et al. 1973; Saito and Itakura 1982; van den Heuvel et al. 1993; Wang et al. 2016).

The presentation of our work is as follows. Sec. 2 describes and illustrates the method (Clermont 2022) adopted for obtaining sub-band CCs, hereafter referred to as band-limited CCs (BLCCs in short). The BLCC method affords flexibility and efficiency, two properties exploited in this work.

Sec. 3 recalls the basics of the LR framework. Sec. 4 concerns the multi-speaker vowel data used, the BLCC parameterisation applied to a sequence of sub-bands, the FVC procedures, and the LR-based metric for performance assessment. Sec. 5 presents full-band and sub-band FVC results for each vowel. Sec. 6 discusses the results in context of previous work, and Sec. 7 outlines potential ways forward.

Proceedings of The 21st Annual Workshop of the Australasian Language Technology Association, pages 40–50
November 29 - December 1, 2023

2 The BLCC Method

This section focuses on the method employed for obtaining BLCCs by a linear transformation of full-band CCs. The method is described in Sec. 2.1, and its mathematical formulation is outlined in Sec. 2.2. In Sec. 2.3, the numerical and spectral behaviours of BLCCs show that the practical size for a BLCC vector depends on the fraction of the full-band's frequency range occupied by the sub-band's width.

2.1 Procedural steps

The BLCC method consists of three main steps encapsulated in Fig. 1. Steps (1) and (2) describe standard procedures of spectral analysis, which are applied to short-time frames of the speech signal sampled at some frequency F_s (Hz). The final step (3) concerns the linear transformation itself.

At Step (1), the all-pole linear-prediction (LP) model of speech production is adopted for two reasons: (a) It provides a reliable characterisation of the spectral resonance patterns of non-nasalised, voiced sounds; (b) It is thus expected that speaker differences are strongly encoded in the LP cepstral representation of the vowels used for this study.

Step (1) yields a log magnitude spectral (LMS) representation based on the LP model (order M), which spans the entire frequency range $[0, (F_s/2)]$ in Hertz (or $[0, \pi]$ in radians). The dashed curve in Fig. (2a) illustrates this representation also known as the "exact" LP-based LMS. Note that the frequency scale along the horizontal axis is kept linear in our experiments, thus leaving open the possibility of finding speaker-sensitive sub-bands without pre-defined nonlinear constraints.

The purpose of the Discrete Cosine Transform (DCT) at Step (2) is to expand the exact LMS as a Fourier cosine series of the so-called cepstral coefficients C_k. These are here referred to as full-band C_k since our LMS spans the full frequency range. The average of the full-band LMS is usually assumed to be zero, hence $C_0 = 0$. In practice, the series is truncated after M terms as follows:

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

The solid curve in Fig. 2(a) depicts the cepstrally-smoothed LMS resulting from the truncated series. As noted earlier, smoothing has the beneficial effect of enhancing spectral distinctiveness.

At Step (3), BLCCs are obtained using a method

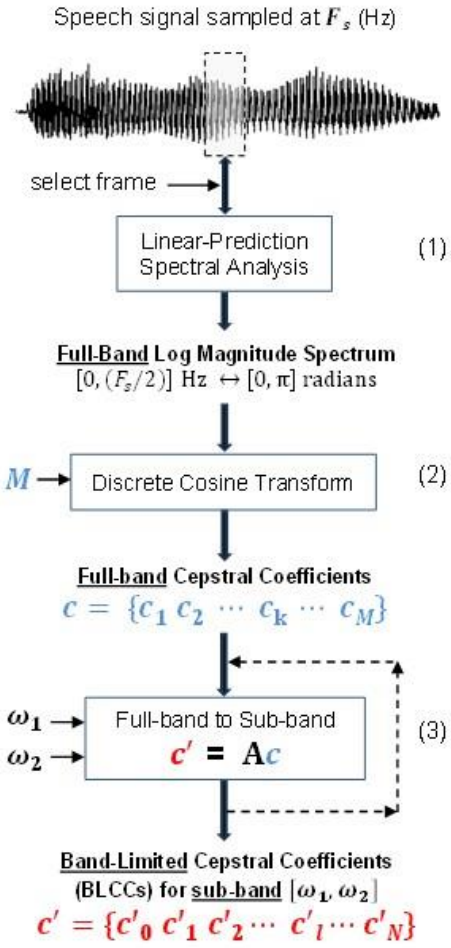

Figure 1: The BLCC method and its main steps.

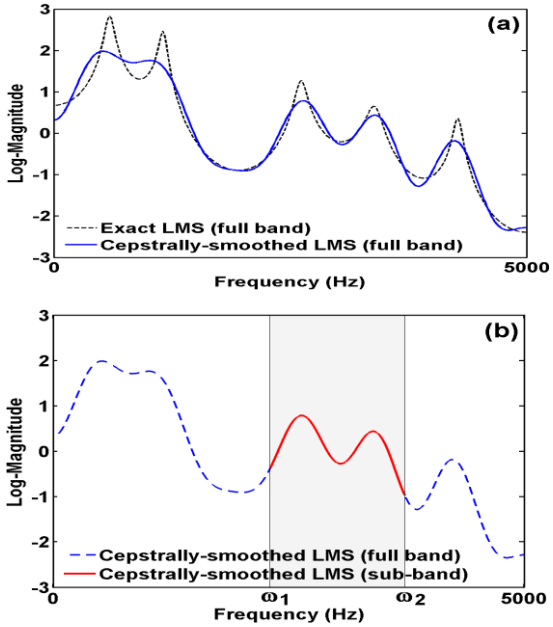

Figure 2: Spectral representations of a back vowel: **(a)** Exact LMS (full band) based on LP analysis (order M=14) at Step (1), overlaid with cepstrally-smoothed LMS based on Eq. (1) and on the C_k obtained at Step (2); **(b)** Sub-band region $[\omega_1, \omega_2]$ highlighted as an integral part of the full-band, cepstrally-smoothed LMS.

which affords the flexibility of selecting any sub-band region of the full-band spectrum without repeating the two previous steps. The central idea portrayed in Fig. 2(b) is this: Focusing on a sub-band region $[\omega_1, \omega_2]$ does not alter the fact that it forms an integral part of some full-band spectrum. It is therefore conceivable that sub-band cepstra are derivable from full-band cepstra. As shown in Clermont's (2022) study, the vector $\mathbf{c}'$ of BLCCs representing a sub-band can indeed be calculated using a linear transformation $\mathbf{A}$ of the vector $\mathbf{c}$ of full-band C_k. Sec. 2.2 outlines the transformation formulae. Key properties are illustrated in Sec. 2.3.

2.2 Linear transformation formulae

The mathematical goal is to represent a sub-band region $[\omega_1, \omega_2]$ of the full-band, cepstrally-smoothed LMS with a Fourier cosine series, such that its coefficients C'_l depend on the full-band C_k.

The band-limited analogue of Eq. (1) may be expressed as follows:

$$S(\omega(\omega')) = C'_0 + \sum_{l=1}^{N} C'_l \cos(l\omega'),\ 0 \le \omega' \le \pi \quad (2)$$

where C'_l is the l-th BLCC and N is the series' upper bound. Eq. (2) includes C'_0 because the average of $S(\omega(\omega'))$ within a sub-band may not be zero. The other $C'_{l>0}$ represent the spectral shape.

The frequency variable ω' defined below plays a key role by translating the sub-band interval $[\omega_1, \omega_2]$ to that of the full-band range $[0, \pi]$:

$$\omega' = \pi \left[\frac{(\omega-\omega_1)}{(\omega_2-\omega_1)}\right],\ \omega_1 \le \omega \le \omega_2 \quad (3)$$

From Eq. (3) it is easy to express the frequency variable ω of the full-band series as:

$$\omega(\omega') = \omega_1 + \left[\frac{(\omega_2-\omega_1)}{\pi}\right]\omega' = \omega_1 + W\omega' \quad (4)$$

where the scalar W is the ratio of the sub-band's width to the full-band's frequency range.

The notation $\omega(\omega')$ is a reminder that ω is itself a (band-dependent) function of ω', thus making it possible to substitute ω in Eq. (1) for Eq. (4) and to use standard formulae for the coefficients of the BLCC series in Eq. (2). These operations lead to:

$$C'_l = \sum_{k=1}^{M} a_{lk}\, C_k,\ l = 0,1,\dots,N \quad (5)$$

and to the matrix form $\mathbf{c}' = \mathbf{A}\mathbf{c}$ laid out below:

$$\begin{bmatrix} C'_0 \\ C'_1 \\ \vdots \\ C'_l \\ \vdots \\ C'_N \end{bmatrix} = \begin{bmatrix} a_{0,1} & \cdots & a_{0,k} & \cdots & a_{0,M} \\ a_{1,1} & \cdots & a_{1,k} & \cdots & a_{1,M} \\ \vdots & & \vdots & & \vdots \\ a_{l,1} & \cdots & a_{l,k} & \cdots & a_{l,M} \\ \vdots & & \vdots & & \vdots \\ a_{N,1} & \cdots & a_{N,k} & \cdots & a_{N,M} \end{bmatrix} \begin{bmatrix} C_1 \\ \vdots \\ C_k \\ \vdots \\ C_M \end{bmatrix} \quad (6)$$

The band-dependent weights a_{lk} are given in Eq. (7a) for $l = 0$ and in Eqs (7b)-(7c) for $l > 0$.

$$a_{lk,\, l=0} = \beta_k[\sin(k\omega_2) - \sin(k\omega_1)] \quad (7a)$$

$$a_{lk,\, l\neq kW} = \gamma_{lk}[(-1)^{l+1}\sin(k\omega_2) + \sin(k\omega_1)] \quad (7b)$$

$$a_{lk,\, l=kW} = \cos(k\omega_1) \quad (7c)$$

where:

$$\beta_k = \frac{1}{k(\omega_2-\omega_1)} \text{ and } \gamma_{lk} = \frac{2(kW)}{\pi[l^2-(kW)^2]} \quad (7d)$$

The implementation of Eqs (6) and (7) raises the question of how large N needs to be in practice. The empirical solution suggested in Clermont's study is to fix N at $M \times W$ (MW in short) rounded to the nearest integer, where W is the ratio defined above and M the size of the vector of full-band C_k.

2.3 Numerical illustrations of key properties

What do BLCCs look like, and how effective are they at preserving spectral resolution in a sub-band region for $N = MW$?

Fig. 3(a) gives a glimpse of BLCC series for two sub-bands selected from the same back vowel illustrated in Fig. (2). The full-band $C_{k=\{1\cdots M=14\}}$ were obtained by DCT of the full-band LP-based LMS ranging from 0 to 5 kHz. Eqs (6) and (7) were then used to calculate BLCCs for these sub-bands: [0.1-0.814]-kHz and [2.3-3.728]-kHz, the latter being twice as large as the former.

The coefficient C'_0 in Fig. 3(a) is visibly much larger in the [0.1,0.814]-kHz range, thus indicating a prominent region in the lower part of the spectrum. The next C'_l exhibit a consistent trend for both sub-bands: A major drop in magnitude is noticeable after MW, followed by a clear decay towards zero.

Is the proposed truncation after MW detrimental to the spectral resolution in a sub-band region? To gain insights into this question, it is instructive to observe cepstrally-smoothed spectra representing the full band and the two sub-bands. The latter are overlaid in Figs 3(b)-(d) for $N = 0, 1, MW$, respectively. The $N = 0$ cases in Fig. 3(b) correspond to using only C'_0. While the spectral fits are expectedly very poor, these coefficients alone give a good indication of the respective levels of the prominences in the two sub-bands. Recruiting the next BLCC with $N = 1$ improves the approximation by capturing the overall slopes in Fig. 3(c). Finally, the spectral fits become very tight in Fig. 3(d) with $N = MW$.

In sum, the numerical evidence described above indicates that BLCCs after MW tend to contribute relatively little to the spectral representation of a sub-band. This is supported by the consistent decay towards zero seen in Fig. 3(a).

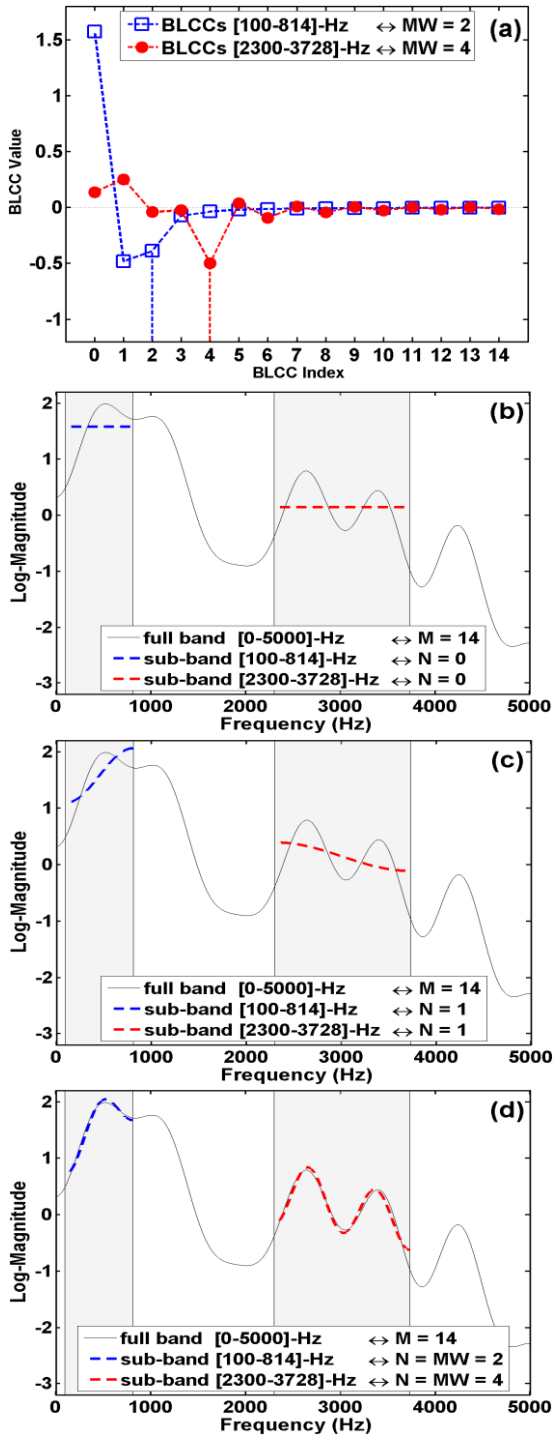

Figure 3: **(a)** BLCC series for two selected sub-bands. Cepstrally-smoothed spectra (full band based on Eq. (1), and sub-bands based on Eq. (2)) are superimposed for the following upper bounds: **(b)** $N = 0$, **(c)** $N = 1$, and **(d)** $N = MW$.

3 Likelihood Ratio Framework

The LR framework provides the theoretical foundation upon which voice evidence is analysed for source-inference purposes. In FVC, the task of the expert is to estimate the strength of voice evidence using the LR expressed as follows:

$$\mathrm{LR} = \frac{p(E = (X, Y)|H_p)}{p(E = (X, Y)|H_d)} \qquad (8)$$

The LR is the ratio of two conditional probabilities: the numerator is the probability (p) of the evidence (E) given the prosecution (same-speaker) hypothesis (H_p), while the denominator is the probability given the defense (different-speaker) hypothesis (H_d).

The evidence (E) typically consists of the source-questioned sample (X) and the source-known sample (Y). In theory, the belief of the trier-of-fact regarding the hypotheses, which was developed by the previously presented evidence, is to be updated by the LR; the assessment of the newly presented evidence. In other words, the belief of the decision maker regarding the suspect being guilty or not changes as a new piece of evidence is presented to them in the form of a LR.

The further away from LR=1, the more strongly the LR supports either of the competing hypotheses.

4 Experimental Procedures

4.1 Speech material and parametrisation

The speech materials were taken from a Japanese dataset of 297 speakers (between 20 and 60 years old) as described in Osanai et al. (1995). The citation recordings (landline telephone calls) of the 5 vowels (2 non-contemporaneous sessions × 2 tokens) were used for the FVC experiments.

The sampling frequency is 10 kHz because the high-end of the telephone bandpass is around 4.5-kHz in Japan, i.e., the available full-band extends from 0 to 5 kHz. Full-band CCs were extracted by linear-prediction (LP) analysis (order 14) of each vowel's central frame.

Using the sub-band transformation explained in Sec. 2, BLCCs were obtained from the full-band CCs by scanning the full range with a 600-Hz sub-band shifted every 400 Hz. This process yielded 12 vectors of BLCCs corresponding to the 12 sub-bands listed in Table 1.

1	[0, 0.6]	2	[0.4, 1.0]	3	[0.8, 1.4]
4	[1.2, 1.8]	5	[1.6, 2.2]	6	[2.0, 2.6]
7	[2.4, 3.0]	8	[2.8, 3.4]	9	[3.2, 3.8]
10	[3.6, 4.2]	11	[4.0, 4.6]	12	[4.4, 5.0]

Table 1: Limits $[\omega_1, \omega_2]$ in kHz of the 12 sub-bands.

Following the definition given in Sec. 2.2, the upper bound for the BLCC series representing a 600-Hz sub-band may be fixed at $MW = 14 \times \frac{600}{5000} = 1.68$ and then rounded up to 2 for practical use. Per sub-band, the total number of BLCCs is 3 including the 0th-order one. A FVC system incorporating BLCCs was then employed to calculate LRs for each of the 12 sub-bands.

4.2 Data partitioning and LR calculation

The 297 speakers were randomly divided into three mutually-exclusive batches (99 speakers each). These were used as the test, background, and calibration databases in a cross-validation manner, resulting in six-fold cross-validation experiments. The results of the six experiments were averaged for comparison.

The LR calculation is a two-stage process consisting of a feature-to-score stage and a score-to-LR stage. A statistical model commonly used in linguistic-phonetic FVC is the Multivariate Kernel Density (MVKD) model for the feature-to-score stage (Aitken and Lucy 2004). The output of the MVKD model is a score, and the score is converted to a LR value at the score-to-LR stage. The MVKD returns a score for a pair of recordings under comparison by assessing their similarity and typicality. The necessary statistical information for typicality is obtained from the background database. The score-to-LR conversion, also called "calibration", is performed via logistic regression (Morrison 2013). The logistic regression weights are determined using the calibration database.

4.3 Performance assessment

The log-LR-cost (C_{llr}) is a standard metric for assessing LR-based inference systems in forensic science. Eq. (9) is the formula for C_{llr}, where N_{SS} and N_{DS} are the numbers of the same-speaker (SS) and different-speaker (DS) LRs, respectively. The SS LRs are indexed by i and the DS LRs by j.

$$C_{llr} = \frac{1}{2}\left(\frac{1}{N_{SS}}\sum_{i}^{N_{SS}} log_2\left(1+\frac{1}{LR_{SS_i}}\right) + \frac{1}{N_{DS}}\sum_{j}^{N_{DS}} log_2\left(1+LR_{DS_j}\right)\right) \quad (9)$$

The first $log_2(\cdot)$ is the cost function for the SS LRs and the second one is for the DS LRs. The C_{llr} is the grand average between the mean cost of the SS LRs and that of the DS LRs. The lower the C_{llr}, the better in performance.

5 Experiment Descriptions and Results

Two FVC experiments were run separately per vowel, and the results are jointly charted in Fig. 4.

In Experiment 1, speaker information locally-encoded in the spectrum was investigated vowel-by-vowel by conducting the experiments with the multi-band BLCCs (see Table 1 for the specific locations of the sub-bands).

The C_{llr} values obtained for the 12 sub-bands are displayed as a red curve at the bottom plot of each panel included in Fig. 4. Each C_{llr} value (Y-axis) is given against the central frequency (X-axis) of the sub-band. The horizontal dashed line (in blue) indicates the overall mean of the 12 C_{llr} values for the vowel. Expected formant-frequency ranges (F1, F2 and F3) taken from Kinoshita et al. (2022) are also marked for each vowel.

In Experiment 2, the sub-band LRs obtained from Experiment 1 were fused from two to all sub-bands as per the following list (r) = {2,3,…,12}. All possible combinations of r sub-bands $\binom{12}{r}$ were also included.

In the top plot of each panel, the best (lowest) C_{llr} value is given for each r together with the C_{llr} value of the single best sub-band (r=1). The C_{llr} for the full-band CCs is indicated by the horizontal dotted line. The three best sub-bands (fused) are highlighted in blue in the bottom plot, and the three worst sub-bands (fused) are highlighted in pink.

5.1 Results: Experiment 1

The red curve included in the bottom plot of each panel (Fig. 4) stays consistently below C_{llr}=1, implying that every spectral region specified by the sub-bands carries some useful speaker information for FVC. However, the fluctuations in the C_{llr} curves indicate that speaker-specific information is not evenly distributed throughout the entire frequency range, and the distributional patterns are distinctive for each vowel. It is worth noting that the C_{llr} value consistently increases for the rightmost sub-band [4.4, 5.0] kHz, meaning that this spectral region contains relatively less speaker-specific information. This may be due to the upper

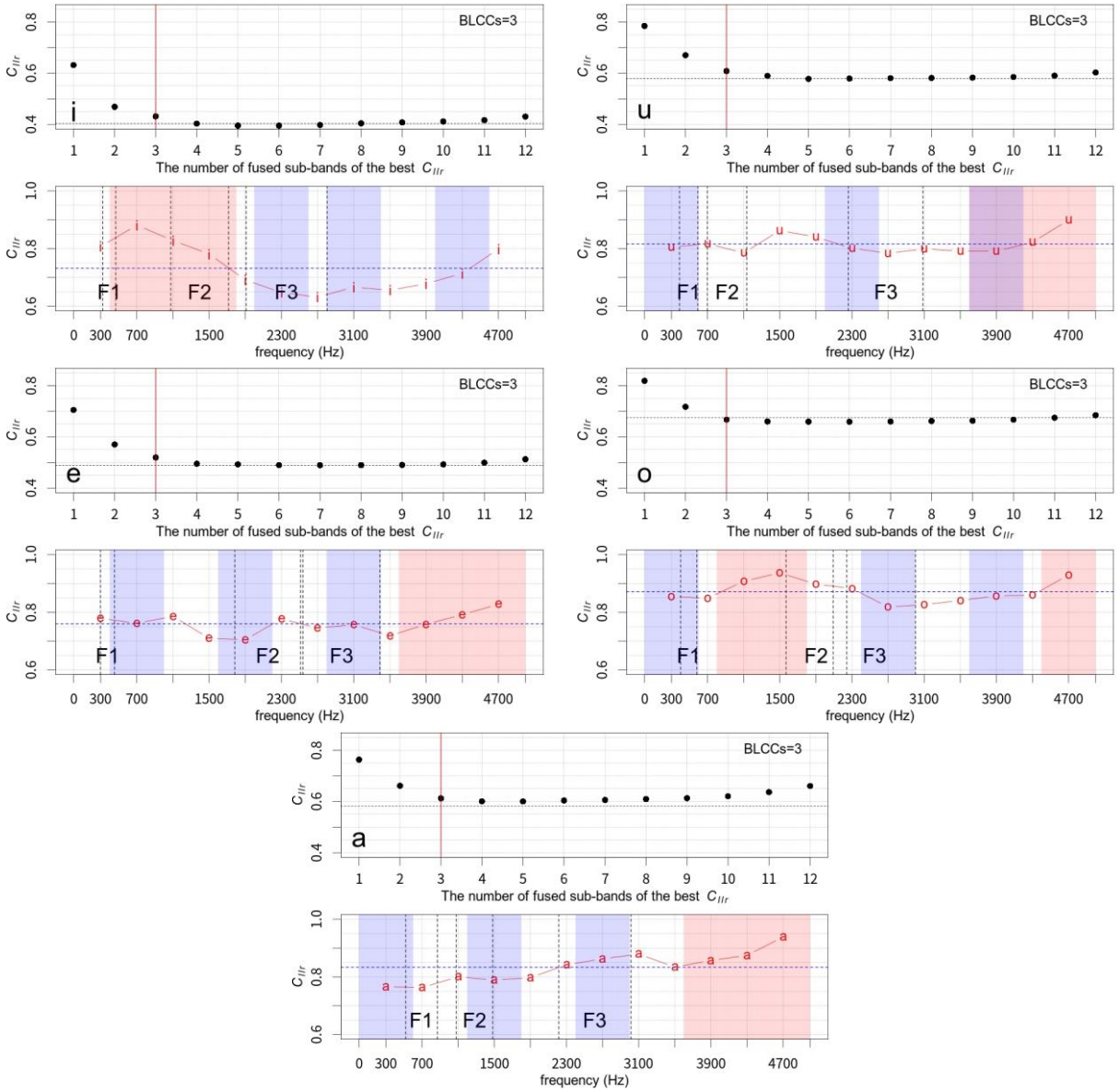

Figure 4: The results for each of the five Japanese vowels are grouped in a separate panel. The top plot in each panel contains the best C_{llr} values for the fused r (=1 to 12) sub-bands. The horizontal dotted line indicates the C_{llr} value for the full-band CCs. The vertical red solid line indicates r=3 for which sub-band performance becomes very close to the full-band result. The bottom plot in each panel gives the profile of C_{llr} values (in red) for the 12 sub-bands. The horizontal dashed line (in blue) indicates the C_{llr} value averaged over the 12 sub-bands. The sub-band regions highlighted in blue are the three best-performing sub-bands (fused), and the sub-bands highlighted in pink are the three worst-performing sub-bands (fused).

limit of the Japanese telephone band-pass located near 4.5 kHz (Rose et al. 2003).

The distributional patterns of speaker-specific information are particularly contrastive between /i/ and /a/. The information is more strongly encoded in the mid- and high-frequency regions of /i/, roughly between 1.9 and 4.3 kHz (covering F3 and beyond). By contrast, for /a/, it is the low-frequency region up to about 1.9 kHz (spanning F1 and F2) which carries the bulk of speaker-specific information. These findings agree with the observations reported in Osanai et al. (2018). Their study based on sub-band cepstral distances points to roughly 2.0 kHz as the frequency below which speaker verification accuracy was relatively higher for /a/, and roughly 1.9 kHz above which speaker verification performed relatively better for /i/.

For the other vowels (/u, e, o/), the ups and downs of the C_{llr} curves are overall less dynamic than those for /i/ and /a/. Yet, some alternations in C_{llr} are still evident. For instance, the C_{llr} values are marginally lower in the range between approximately 2.7 and 3.9 kHz (spanning F3 and

beyond) for /u/ and /o/ and around 1.5-1.9 kHz (spanning F2) for /e/.

Kinoshita (2001) found that F2 of /i/ and F2 and F3 of /e/ are strong acoustic features for Japanese FVC. With more specific details based on F-ratios, Khodai-Joopari et al. (2004) reported that the spectral regions of 1.7-2.7 kHz (spanning F3) and 3.7-4.5 kHz (extending beyond F3) of /i/, and the spectral regions of 1.4-2.4 kHz (spanning F2), 2.6-3.7 kHz (spanning F3) and 3.8-4.5 kHz (extending over F3) of /e/, are potentially useful for speaker classification based on Japanese vowels.

The results obtained from Experiment 1 mostly agree with the findings from the two studies referenced above, in that the spectral or formant regions identified as promising returned categorically low C_{llr} values. For example, the lowest C_{llr} value (=0.63171) of all vowels lies within the frequency range pointed out by Khodai-Joopari et al. (2004) for /i/. Likewise, the lowest C_{llr} value (=0.70510) for /e/, also the lowest amongst vowels /e, a, o, u/, also occurs in the F2 range pointed by Kinoshita (2001).

For the back vowels (/u, o, a/), the frequency range spanning and/or extending beyond F3 is reportedly a good candidate for Japanese speaker classification (Khodai-Joopari et al. 2004). Some studies also report the usefulness of F3 of back vowels as a speaker discrimination feature in English (Mokhtari and Clermont 1996; Sambur 1975). As noted above, the importance of the frequency region spanning F3 and beyond holds true in our results for /u/ and /o/. This point will be revisited in describing results from Experiment 2.

5.2 Results: Experiment 2

Turning our attention to the top plots in each panel, it can be observed that regardless of the vowels, performance is improved by fusing multiple sub-band LRs. The performance is substantially enhanced when 3 or 4 sub-bands are fused in contrast to using only the best single sub-band. As a matter of fact, the fusion of 3 or 4 optimal sub-bands brings the system to nearly the same performance level as that obtained with full-band CCs or even marginally better.

Note that 3 or 4 sub-bands are here represented with 9 or 12 BLCCs in total, respectively. Thus, only a few BLCCs are necessary to achieve nearly the same performance as that obtained with the 14 full-band CCs. This a notable advantage of BLCCs in terms of computational efficiency.

The performance stays basically unchanged even when more sub-bands are included for fusion, except for a slight deterioration in performance towards the higher numbers of fused sub-bands.

Together with the results from Experiment 1, the above observations based on Experiment 2 would seem to indicate that locally-encoded speaker information is not necessarily unique as per its spectral region. In other words, pieces of speaker information may be redundantly encoded across different spectral regions. Otherwise, the continuous decline in C_{llr} (an incessant gain in performance) should have been observed as more sub-bands are totalled for fusion.

The bottom plots in each panel clearly show that the three best-performing sub-bands span different spectral regions depending on the vowel. For /i/, they are in the mid- and high-frequency ranges above 2 kHz, which generally correspond to the spectral regions with strong speaker information. On the other hand, for /e/ and /a/, the three best sub-bands are dispersed in the low- and mid-frequency ranges below 3.0-3.4 kHz. For /u/ and /o/, the three best sub-bands are most widely separated in the range approximately between 0 and 4.2 kHz.

It is noticeable that the 3 best sub-bands are not only spaced apart from each other, but they also tend to fall in the speaker-sensitive spectral regions. This leads us to conjecture that those sub-bands are likely to contain more locally-distinctive speaker information. In support of the conjecture, it can be observed that the 3 worst sub-bands (coloured in pink) are in immediately neighbouring positions. For /u, e, a/, they are the 3 contiguous sub-bands appearing in the high-end of the spectrum and, for /i/, the 3 sub-bands flock together towards the low-frequency end, where C_{llr} values are worse. It can therefore be surmised that those sub-bands did not perform well after fusion because they are largely redundant in speaker information in addition to being less sensitive to speaker individuality, as demonstrated in Experiment 1.

Following on from Experiment 1, the importance of the F3 region for FVC is also evident for the back vowels from the bottom plots given in Fig. 4, in that one of the 3 sub-bands falls in the F3 region. The sub-band spanning F3 does not seemingly contain strong speaker information for vowel /a/; the C_{llr} values of the region are higher than the average C_{llr}. Nevertheless, the speaker information encoded in the F3 region is judged to be complementary with the sub-bands spanning F1

and F2 for /a/. For /u/ and /o/, one sub-band appears in the frequency range beyond F3, in agreement with Khodai-Joopari et al. (2004).

As can be seen from the bottom plots for the back vowels, the first sub-band [0, 0.6] kHz turned out to be a good one when fused with the other 2 sub-bands. Judging from the commonly shared empirical knowledge that more speaker-specific information is encoded in higher spectral regions (Hayakawa and Itakura 1994; Kitamura and Akagi 1995), this result is counter-intuitive. However, Khodai-Joopari et al. (2004) also sighted a peak of speaker F-ratio below F1 region for /o/ and /a/, and suggested their glottal-source characteristics as a possible cause for the peak.

6 Discussion

The FVC results presented in Sec. 5 confirm the existence of speaker-sensitive spectral regions, which principally agree with previous acoustic and articulatory studies of vowels. As such, it can demonstrably be argued that BLCC is a useful analytical tool equipped with flexibility and precision in selecting any sub-band of interest.

The formant frequencies (F1, F2 and F3) are common phonetic features in linguistic-phonetic FVC (Rose et al., 2003; Morrison 2008, Rose, 2017). The analytical potential of the multi-band BLCCs, however, unavoidably led us to notice that the regions corresponding to formant frequencies do not always contain strong speaker information. For example, the C_{llr} values for sub-bands spanning the F1-F2 region of /i/, the F2 region /o/ and the F3 region of /a/ are relatively high compared to the other regions. This suggests that sub-band selection based strictly on formant ranges is an unnecessarily constraining and even sub-optimal solution.

A case in point is Kinoshita et al's (2022) results based on sub-band cepstral distances and on prior knowledge of fixed F1, F2 and F3 sub-bands. A set of FVC experiments was done with the sub-bands that were selected according to the fixed F1, F2 and F3 ranges provided in Kinoshita et al. (2022) for the same experiments performed in the current study. The resultant C_{llr} values are shown in Table 2, together with the C_{llr} values with the 3 optimal sub-bands (fused) selected empirically (see Fig. 4), i.e., without prior acoustic-phonetic knowledge. The C_{llr} values for the full-band CCs are also listed.

The results from the 2 rightmost columns of Table 2 indicate that BLCCs can achieve nearly full-band performance with 3 optimal sub-bands and, thus, with fewer cepstral features. This finding illustrates the power of BLCCs in locating such sub-bands without any prior knowledge.

Vowels	Kinoshita et al (2022)	This Study	
	3 sub-bands (with prior knowledge)	3 sub-bands (without prior knowledge)	full band
/i/	0.52191	0.43142	0.40342
/u/	0.68992	0.60858	0.57934
/e/	0.54173	0.51947	0.48843
/o/	0.73500	0.66732	0.67500
/a/	0.65428	0.61239	0.58130
Ave.	0.62856	0.56783	0.54549

Table 2: Middle columns: C_{llr} values for 3 fused sub-bands selected using two approaches. Kinoshita et al's (2022) approach with prior knowledge, i.e., based on their formant ranges; and this study's approach without prior knowledge, i.e., guided by empirical selection. Rightmost column: full-band C_{llr} values from this study are included for reference.

It is relevant to point out that while our FVC experiments and Kinoshita et al's (2022) involve phonologically the same vowels and about the same number of speakers, their vowel tokens were produced in various consonantal contexts, whereas ours were produced without any such contexts. Thus, the exact formant ranges could be different for the vowels included in these two studies.

Notwithstanding this discrepancy for now, the trend of C_{llr} values in the 2 middle columns of Table 2 is consistent and encouraging: Our approach (without prior knowledge) outperforms the one employed by Kinoshita et al. (2022) (with prior knowledge). Further investigations with BLCCs applied to Kinoshita et al's vowel data and to other datasets will be necessary to confirm the apparent superiority of our sub-band approach in FVC.

The results obtained in this study are based only on male speech samples. While this is practically justified because males tend to commit crimes more often than females, further experimentation is desirable with a wider variety of speakers. However, the analytical power of BLCCs should remain unaffected by gender or age. It is the locations and ranges of speaker-sensitive spectral regions that could differ with these factors.

While retaining intrinsic properties of the cepstrum (e.g., ease of extraction, immunity to insignificant spectral details), the analytical power of BLCCs allows the forensic scientist to flexibly shift the focus of scrutiny and interpretation according to the selected sub-band region(s). This

is an invaluable contribution that BLCCs can bring to the task of communicating the FVC outcome to the trier-of-fact in a more approachable way.

7 Future Work

The BLCCs exploited here are based on LP modelling of the speech signal and extracted on a linear frequency scale. However, there may be further insights to be gained by applying the same linear transformation to CCs from filter-bank outputs, combined with a nonlinear mapping of the frequency axis such as the often-used Mel scale. It is interesting to note that, except for /i/, our best-performing sub-bands include the lower-spectral regions that are precisely emphasised with Mel-Frequency CCs (MFCCs). A deeper investigation of MFCCs with differing sub-band widths and overlaps is therefore possible using our flexible approach to selecting local spectral regions.

From a forensic point of view, it is coherent to extend the application of BLCCs to non-vowel sounds (Rose 2022), whose speaker-sensitive spectral properties have received relatively less attention. From a linguistic point of view, it is conceivable that BLCCs could also be used as an ancillary or alternative parameter in the areas of acoustic-phonetics (e.g., efficient encoding of contrastive features as in Iskarous (2018)) and socio-phonetics (e.g., exploration of accent-specific sub-bands) (Arslan and Hansen 1997). In connection with these applications, it would be useful to study correlations between BLCCs and formant frequencies via the linear regression models developed by Broad and Clermont (1989) and Clermont (2013), and recently explored by Hughes et al. (2020) in the FVC context.

Finally, it is hoped that the sub-band approach embedded in BLCCs will bring new perspectives in other areas of speech science and technology, such as speech classification (Mokhtari and Clermont 1994), spoofing detection (Chettri et al. 2020; Soni et al. 2016), language identification (Salesky et al. 2021), and speech emotion recognition. Any pieces of information related to speaker variability, speech emotion, or synthesised speech, which are found to be notably encoded in specific sub-bands, would be advantageous for building robust classification systems, or for training deep-learning models. These technological pursuits are likely to benefit from the flexibility and efficiency afforded by the BLCC approach to sub-band spectral analysis.

Acknowledgments

The authors thank the four reviewers for their valuable comments.

References

Aitken, C. G. G. and Lucy, D. (2004) Evaluation of trace evidence in the form of multivariate data. *Journal of the Royal Statistical Society, Series C (Applied Statistics) 53*(1): 109-122. https://dx.doi.org/10.1046/j.0035-9254.2003.05271.x

Arslan, L. M. and Hansen, J. H. L. (1997) A study of temporal features and frequency characteristics in American English foreign accent. *The Journal of the Acoustical Society of America 102*(1): 28-40. https://doi.org/10.1121/1.419608

Broad, D. J. and Clermont, F. (1989) Formant estimation by linear transformation of the LPC cepstrum. *The Journal of the Acoustical Society of America 86*(5): 2013-2017. https://doi.org/10.1121/1.398581

Chettri, B., Kinnunen, T. and Benetos, E. (2020) Subband modeling for spoofing detection in automatic speaker verification. *Proceedings of Odyssey 2020 The Speaker and Language Recognition Workshop*: 341-348.

Clermont, F. (2013) Cepstrum-to-formant mapping of spoken vowels. *Paper presented at the Conference of the International Association in IAFPA 2013 – 22nd Annual Conference of the International Association for Forensic Phonetics and Acoustics*, July 21-24, Tempa, Florida, 2013.

Clermont, F. (2022) Linear transformation from full-band to sub-band cepstrum. In R. Billington (ed.), *Proceedings of the 18th Australasian International Conference on Speech Science and Technology*: 136-140.

Goto, R., Misawa, K. and Okada, Y. (2017) Analysis of individual characteristics in vowel spectral envelopes. *Proceedings of the International MultiConference of Engineers and Computer Scientists*: 113-116.

Hayakawa, S. and Itakura, F. (1994) Text-dependent speaker recognition using the information in the higher frequency band. *Proceedings of the 1994 IEEE International Conference on Acoustics, Speech and Signal Processing*: I/137-I/140.

Hughes, V., Clermont, F. and Harrison, P. (2020) Correlating cepstra with formant frequencies: Implications for phonetically-informed forensic voice comparison. *Proceedings of Interspeech 2020*: 1858-1862.

Hyon, S., Wang, H., Wei, J. and Dang, J. (2012) An investigation of dependencies between frequency components and speaker characteristics based on phoneme mean F-ratio contribution. *Proceedings of the 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference*: 1-4.

Iskarous, k. (2018) The encoding of vowel features in mel-frequency cepstral coefficients. *Il parlato nel contesto naturale [Speech in the Natural Context]*: 9-18. https://dx.doi.org/10.17469/O2104AISV00000

Khodai-Joopari, M., Clermont, F. and Barlow, M. (2004) Speaker variability on a continuum of spectral sub-bands from 297-speakers' non-contemporaneous cepstra of Japanese vowels. *Proceedings of the 10th Australian International Conference on Speech Science and Technology*: 504-509.

Kinoshita, Y. (2001). *Testing Realistic Forensic Speaker Identification in Japanese: A Likelihood Ratio Based Approach Using Formants.* Unpublished PhD thesis, the Australian National University, Canberra.

Kinoshita, Y., Osanai, T. and Clermont, F. (2022) Sub-band cepstral distance as an alternative to formants: Quantitative evidence from a forensic comparison experiment. *Journal of Phonetics 94*: 101177. https://doi.org/10.1016/j.wocn.2022.101177

Kitamura, T. and Akagi, M. (1995) Speaker individualities in speech spectral envelopes. *Journal of the Acoustical Society of Japan (E) 16*(5): 283-289. https://doi.org/10.1250/ast.16.283

Mohammadi, S. H., Sameti, H., Tavanaei, A. and Soltani-Farani, A. (2011) Filter-bank design based on dependencies between frequency components and phoneme characteristics. *Proceedings of the 19th European Signal Processing Conference*: 2142-2145.

Mokhtari, P. and Clermont, F. (1994) Contributions of selected spectral regions to vowel classification accuracy. *Proceedings of The 3rd International Conference on Spoken Language processing*: 1923-1926.

Mokhtari, P. and Clermont, F. (1996) A methodology for investigating vowel-speaker interactions in the acoustic-phonetic domain. *Proceedings of the 6th Australian International Conference on Speech Science & Technology*: 127-132.

Morrison, G. S. (2013) Tutorial on logistic-regression calibration and fusion: Converting a score to a likelihood ratio. *Australian Journal of Forensic Sciences 45*(2): 173-197. https://dx.doi.org/10.1080/00450618.2012.733025

Osanai, T., Kinoshita, Y. and Clermont, F. (2018) Exploring sub-band cepstral distances for more robust speaker classification. *Proceedings of the 17th Australasian International Conference on Speech Science and Technology*: 41-44.

Osanai, T., Tanimoto, M., Kido, H. and Suzuki, T. (1995) Text-dependent speaker verification using isolated word utterances based on dynamic programming [in Japanese]. *National Research Institute for Police Science Report 48*(1): 15-19.

Peterson, G. E. (1959) The acoustics of speech–part II: Acoustical properties of speech waves. In L. E. Travis (ed.), *Handbook of Speech Pathology:* 1st Edition ed., 137-173. New York: Appleton-Century-Crofts, Inc.

Pols, L. C., Tromp, H. R. and Plomp, R. (1973) Frequency analysis of Dutch vowels from 50 male speakers. *The Journal of the Acoustical Society of America 53*(4): 1093-1101. https://doi.org/10.1121/1.1913429

Rabiner, L. R. and Juang, B. H. (1993) *Fundamentals of Speech Recognition* (1st ed.). Englewood Cliffs, N.J.: Prentice Hall.

Rose, P. (2022) Likelihood ratio-based forensic semi-automatic speaker identification with alveolar fricative spectra in a real-world case. In R. Billington (ed.), *Proceedings of the 18th Australasian International Conference on Speech Science and Technology*: 6-10.

Rose, P., Osanai, T. and Kinoshitaa, Y. (2003) Strength of forensic speaker identification evidence: multispeaker formant-and cepstrum-based segmental discrimination with a Bayesian likelihood ratio as threshold. *Forensic Linguistics 10*: 179-202. https://doi.org/10.1558/sll.2003.10.2.179

Saito, S. and Itakura, F. (1982) Personal characteristics of the frequency spectrum of vowels. *Annual Bulletin Research Institute of Logopedics and Phoniatrics 16*: 73-79.

Salesky, E., Abdullah, B. M., Mielke, S. J., Klyachko, E., Serikov, O., Ponti, E., … Vylomova, E. (2021) SIGTYP 2021 shared task: Robust spoken language identification. *Proceedings of the 3rd Workshop on Computational Typology and Multilingual NLP*: 122–129.

Sambur, M. (1975) Selection of acoustic features for speaker identification. *IEEE Transactions on Acoustics, Speech, and Signal Processing 23*(2): 176-182. https://dx.doi.org/10.1109/TASSP.1975.1162664

Soni, M. H., Patel, T. B. and Patil, H. A. (2016) Novel subband autoencoder features for detection of spoofed speech. *Proceedings of Interspeech 2016*: 1820-1824.

van den Heuvel, H., Cranen, B. and Rietveld, A. C. M. (1993) Speaker-Variability in Spectral Bands of Dutch Vowel Segments. *Proceedings of the 3rd European Conference on Speech Communication and Technology*: 635-638.

Wang, L., Wang, J., Li, L., Zheng, T. F. and Soong, F. K. (2016) Improving speaker verification performance against long-term speaker variability. *Speech Communication 79*: 14-29. http://dx.doi.org/10.1016/j.specom.2016.02.004