Exploring sub-band cepstral distances for more robust speaker classification

Takashi Osanai[1], Yuko Kinoshita[2], Frantz Clermont[3]

[1] National Research Institute of Police Science, Japan
[2]College of Arts and Social Science/Asia and the Pacific, The Australian National University
[3]J.P. French Associates Forensic Lab., England
osanai@nrips.go.jp; yuko.kinoshita@anu.edu.au; dr.fclermont@gmail.com

Abstract

This paper presents the first of two-part exploration into the potential of parametric cepstral distance (PCD) as a forensic voice comparison feature, based on Japanese vowel data collected from 306 male native speakers under microphone and mobile transmission conditions. The behaviours of PCDs were closely examined by altering sub-band settings, and we found the behaviour of PCDs to correspond well to what is known about formants, which suggests that PCDs are relatable to articulatory gestures. Comparison between sub-band and full-band PCD revealed that limiting the band range to a specific frequency region makes the feature more robust against channel mismatch, encouraging further examination of this potential feature.

Index Terms: Sub-band cepstral distance, F-ratio, Speaker Classification, Channel mismatch, Japanese vowels.

1. Introduction

In the past few decades, the field of forensic voice comparison (FVC henceforth) has seen considerable development in its methodology in classification techniques and in the evaluation of the systems themselves (e.g. [1-4]). Still, the features to which such techniques are applied are mostly unchanged: formants and various types of cepstra, such as MFCC or LPCC, appear to be the two most commonly used features. Some past FVC research favours the use of cepstrum-based features; they generally outperform formants in speaker classification (e.g. [5-7]). This is unsurprising given the differences in their nature as features. Formants represent only the locations of spectral peaks in the frequency domain, whereas the cepstrum captures richer information by utilising the entire user-defined frequency range. Also, the cepstrum can be extracted automatically. Automatic formant extraction, on the other hand, is known to be highly unreliable (e.g. [8]) and often requires manual supervision and correction. This leads to two problems: introducing measurer-dependent variability to the data (e.g. [8-10]), and extreme resource intensiveness.

However, formants have two major advantages over the cepstrum: robustness and interpretability. A real-life FVC case often involves data of poor quality, recorded through different devices and transmission channels. Formants are known to be more robust than the cepstrum under such conditions. Formant frequencies also generally correspond to articulatory gestures in speech production, and it is therefore easier to communicate their meaning to the layperson, as well as for the expert to detect any unusual characteristics or irregularity in the data, which may or may not be related to speaker characteristics.

In legal proceedings, experts are tasked to assist the court to reach correct decisions. Communicating their analysis processes and outcomes in an understandable way to non-experts is thus essential. Of course discussing scientific evidence inevitably involves highly technical concepts unfamiliar to laypeople, and "what is understandable?" is arguable, but it is our view that less abstract and more intuitively understandable processes are preferable in these contexts.

Thus we believe that ideal FVC features need qualities additional to the standard requirements of discriminability: extracted automatically and reliably; robust against poor recording quality and unpredictable environments; and relatable to articulatory gestures for better interpretability.

The band-limited parametric cepstral distance (sub-band PCD) proposed in [11] was identified in [12] as a feature that potentially meets such criteria. Firstly, sub-band PCD is a cepstrum-based feature and readily extracted without human supervision. This facilitates large-scale data processing and excludes measurer-dependent variability. It also allows the analysis to exclude unwanted frequency ranges which largely carry non-speech information. This is particularly attractive in FVC contexts, as recordings in real life FVC often contain substantial background noise, such as passing cars, other peoples' voices, and television noise. The level and the characteristics of such noise sources vary from one moment to next, so the capacity to flexibly focus on relevant frequency ranges should be a significant advantage.

While the preliminary investigation by [12] was based on a very small dataset, it made a few promising observations. First, the F-ratios of sub-band PCDs appear to correspond well to those from formants. Also, the F-ratios were very similar across the mobile transmission and microphone recordings, suggesting that sub-band PCD may be robust against transmission mismatches.

These observations call for further investigation on this feature based on a much larger dataset. The current study thus presents the first part of this investigation. Focusing on observations to the channel effect in relation to the F-ratio and the effect of differing frequency ranges and regions of the sub-bands, we aim to better understand the behaviours of sub-band PCD and explore its potential as a feature for FVC in court.

2. Data

This study selected 306 adult male speakers from the NRIPS database [13]. They are native speakers of Japanese, aged from 18 to 76 years. Their places of origin spread widely across Japan, hence so did their dialectal background. Two non-contemporaneous recordings, separated by 2 to 3 months, were made for each speaker and the recording tasks were performed twice at each recording session. Recordings were made simultaneously through 2 channels: direct microphone (Ch1), and mobile phone transmission (Ch3).

The speech material consisted of read (C)V syllables: the 5 Japanese vowel phonemes, /a/, /i/, /ɯ/, /e/, and /o/, in combination with the 11 preceding consonants, ∅ (no consonant), /k/, /s/, /t/, /h/, /r/, /g/, /z/, /d/, /b/, and /p/. We excluded the consonants /n/, /m/, /y/, and /w/, for the in order to facilitate reliable automatic segmentation.

Japanese hiragana syllabary pairs, i.e. ぢ /di/ – じ /zi/ and づ /dɯ/ – ず /zɯ/, have been merged into phonetically identical forms, [dʑi] and [dzɯ], although the writing system still maintains the distinction. Therefore, for /i/ and /ɯ/, we have the vowel data in 10 different phonological contexts and 11 for /a/, /e/, and /o/.

3. Procedures

3.1. Segmentation and full-band LPCC extraction

The target syllables were automatically segmented into a preceding consonant and a vowel based on their power and F0. The sound files were down-sampled from 44.1 kHz to 8kHz, and full-band LPCCs were extracted from the selected vowel sections (order 14, Hamming window, window length 25ms, time-step 5ms). The LPCCs was averaged across the vowel duration, and the means across different phonological contexts were calculated for each vowel. As result, we obtained the LPCCs for 5 vowels, 2 recording sessions, 2 repeats, and 2 recording channels for each speaker.

3.2. Parametric cepstral distance and F-ratio calculation

The usefulness of FVC features is reflected in the ratio of between- to within-speaker variances, which are expressed below by Equations (1) and (2), respectively. The numerators of both expressions describe parametric cepstral distances (PCDs) between pairs of full-band LPCCs that are index-weighted by the matrix $\mathbf{K}$ to emphasise spectral slope differences, and weighted by the matrix $\mathbf{W}(\omega_1, \omega_2)$ to focus on any sub-band selectable by its lower and upper limits ω_1 and ω_2. It should be noted that the formulation of $\mathbf{W}$ detailed in [1] affords the flexibility of obtaining sub-band PCDs directly from full-band LPCCs. Note also that, for $\omega_1 = 0$ and $\omega_2 = \pi$, the distances in equations (1) and (2) simply reduce to full-band PCDs.

$$\sigma^2_{between}(\omega_1, \omega_2) = \frac{\sum_{i=1}^{N} n_i \, d_i^2(\omega_1, \omega_2)}{N - 1} \qquad (1)$$

$$\sigma^2_{within}(\omega_1, \omega_2) = \frac{\sum_{i=1}^{N} \sum_{j=1}^{n_i} d_{ij}^2(\omega_1, \omega_2)}{(\sum_{i=1}^{N} n_i) - N} \qquad (2)$$

where:
$i \equiv$ speaker-session index, $N \equiv$ number of speakers
$j \equiv$ token index, $n_i \equiv$ number of tokens per i^{th} speaker

$$d_i^2(\omega_1, \omega_2) = \left(\overline{\mathbf{C}_i} - \bar{\bar{\mathbf{C}}}\right)^T \cdot \mathbf{K}^T \cdot \mathbf{W}(\omega_1, \omega_2) \cdot \mathbf{K} \cdot \left(\overline{\mathbf{C}_i} - \bar{\bar{\mathbf{C}}}\right)$$
$$\equiv \text{PCD between } \overline{\mathbf{C}_i} \text{ and } \bar{\bar{\mathbf{C}}} \qquad (3)$$

$$d_{ij}^2(\omega_1, \omega_2) = \left(\mathbf{C}_{ij} - \overline{\mathbf{C}_i}\right)^T \cdot \mathbf{K}^T \cdot \mathbf{W}(\omega_1, \omega_2) \cdot \mathbf{K} \cdot \left(\mathbf{C}_{ij} - \overline{\mathbf{C}_i}\right)$$
$$\equiv \text{PCD between } \mathbf{C}_{ij} \text{ and } \overline{\mathbf{C}_i} \qquad (4)$$

$\mathbf{C}_{ij} \equiv$ mean LPCC for i^{th} speaker's j^{th} token across the vowel duration
$\overline{\mathbf{C}_i} \equiv$ mean LPCC for i^{th} speaker across all tokens
$\bar{\bar{\mathbf{C}}} \equiv$ grand-mean LPCC over all speakers

$$\mathbf{K} = \begin{bmatrix} \mathbf{1} & \mathbf{0} & \cdots & \cdots & \mathbf{0} \\ \mathbf{0} & \ddots & \ddots & & \vdots \\ \vdots & \ddots & \boldsymbol{k} & \ddots & \vdots \\ \vdots & & \ddots & \ddots & \mathbf{0} \\ \mathbf{0} & \cdots & \cdots & \mathbf{0} & \boldsymbol{M} \end{bmatrix} \equiv \text{index-weighting matrix}$$

$\mathbf{W}(\omega_1, \omega_2) \equiv$ band-selective matrix (see [1])
$\boldsymbol{M} \equiv$ LPCC order
$\omega_1 \equiv$ lower limit of sub-band selected within $[0, \pi]$
$\omega_2 \equiv$ upper limit of sub-band selected within $[0, \pi]$

To observe interactions between PCDs and sub-band ranges, we divided the range from 100 Hz to 1000 Hz in 100 Hz increments, and the range from 1000 Hz to 3500 Hz in 500-Hz increments. In search for the frequency regions that are relatively rich in speaker information, we also shifted sub-bands by 100-Hz steps, and calculated PCDs for each step.

3.3. Comparisons

The four recordings from 306 speakers yielded 1224 patterns of same-speaker (SS) pairs and 373,320 patterns of different-speaker (DS) pairs. The 2 different recording channels allowed us to make comparisons under 3 different channel conditions: Ch1 (microphone), Ch3 (mobile phone), and Ch1-3 (channel mismatch).

After calculating the sub-band and full-band PCDs, we examined their F-ratios and conducted simple verification tests based on the size of PCDs, so that we can observe the speaker-classification potential of each vowel-and-sub-band combination. PCDs are already a distance measure, so we pooled PCDs from 1224 SS pairs and 373,320 DS pairs, and separately plotted for their distributions. Using the Equal Error Rate (EER) as the threshold, we classified the PCDs under the three different channel conditions.

4. Results and discussion

4.1. F-ratios

We observed higher F-ratios for the between-Ch1 comparison than for between-Ch3, with occasional exceptions. This was expected, as microphone recordings contain more information and less of the unpredictable variability caused by mobile phone and telephone transmission.

The F-ratio plots for various sub-band ranges revealed that sub-band PCDs generally behave similarly in both channels, although how much they vary across two channels depends on vowel, with /a/ and /o/ revealing the greater variation than the rest. In general, the difference between two channels were greater in the higher-frequency regions.

Also, for all vowels, the greatest peak F-ratios were found where the sub-band range was set at 100Hz, the narrowest of all. The frequency region affects F-ratio less as the sub-band range becomes greater, and gets close to the baseline F-ratio, obtained from the full-band PCD, as expected.

In observing their relation to the full-band PCDs, we find that the sub-band PCDs outperform the full-band PCDs in certain frequency regions. This suggests that the sub-band PCDs are likely to outperform the full-band PCDs in speaker classification, when multiple of them are combined.

Figure 1 presents the results for one of the sub-band sizes, 300 Hz, as an example. It summarises the relationship between F-ratios and frequency regions. The results from Ch1 are shown in red, and Ch3 in blue. The horizontal lines present the baseline F-ratio that was produced form the full-band PCDs.

One of the characteristics we seek in FVC features are interpretability. If we can observe the correspondence between the expected formant values and the frequency regions which produced the high F-ratios, it suggests sub-band PCDs' interpretability similar to formants. Here we present Japanese vowel formants data obtained from 11 male speakers [14] to compare with Figure 1. It shows that the frequency regions where higher F-ratios was observed are, in many cases, vicinity of the expected locations of the formants.

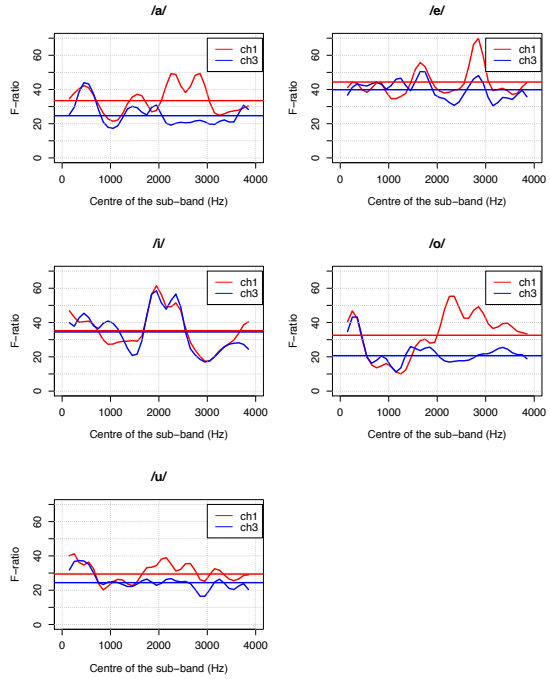

Figure 1: *F-ratios for varying frequency regions (sub-band width 300Hz shifted by 100Hz step), horizontal lines indicating full-band results*

Table 1: *Mean and SD of vowel formants from 11 male speakers of Japanese*

	F1		**F2**		**F3**	
	mean	*sd*	*mean*	*sd*	*mean*	*sd*
/a/	666	53	1414	114	2532	190
/e/	452	108	2009	136	2666	166
/i/	306	35	2174	184	2886	189
/o/	462	94	1125	115	2508	163
/u/	334	45	1550	154	2404	202

4.2. Verification results

Next we conducted a simple verification experiment to gain some insights into the relative effectiveness of each sub-band and channel condition as a speaker classifying feature. The verification rate was calculated from the 2 PCD distributions (1,224 SS pairs and 373,320 DS pairs), the EER point used as the threshold, and the frequency regions with a higher speaker discriminability were sought. Each sub-band size and vowel combination was tested by altering sub-band ranges and frequency regions. The whole experiment was then repeated under 3 different channel conditions: microphone (Ch1), mobile phone (Ch3), and mismatch condition (Ch1-3).

Figure 2 presents a sample of the results taken from the tests using the sub-band range of 300Hz as Figure 1. Here, what immediately notable is the impact of channel mismatch. With every vowel and band range including the full-band, verification rate deteriorated considerably under mismatch conditions. Although this was expected, its significance is noteworthy.

The relationship between the two matching conditions, on the other hand, was somewhat unexpected. With all vowels but /o/, the Ch3 condition produced a better outcome than the Ch1 condition. Although it is counterintuitive, it has been reported that speech put through a mobile phone codec performs better in speaker classification than un-coded speech does [15], perhaps because the processing applied by mobile codec has the effect of reducing within-speaker variability. This does not agree with what we observed with F-ratios, however. Why the result for the vowel /o/ should be opposite to those of the other vowels remains unclear. However, with the /o/ vowels in Figure 1, F-ratio for Ch3 is notably worse than that for Ch1 at around 2200-3500Hz. We could speculate that the mobile transmission strongly affected the frequency region particularly relevant to speaker information for /o/ vowel (appears to be F3 region), and the positive effect of within-speaker variation reduction could not compensate this loss. Further investigation is needed, however.

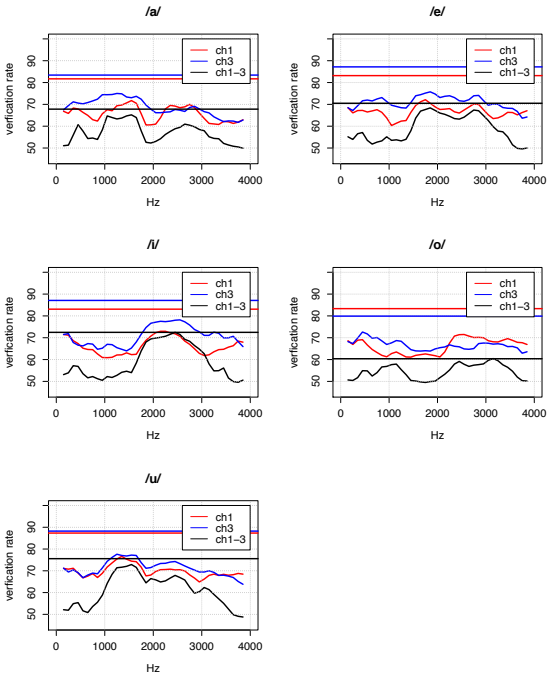

Figure 2: *Verification rates for varying frequency regions (sub-band 300Hz shifted by 100Hz step), horizontal lines indicating full-band results*

What is most notable with this experiment, however, is the relationship between the verification rates of the sub-band PCDs and the base-line full-band PCDs. The verification rate difference between two appears to be smaller under the mismatch condition in almost all the comparisons, suggesting that sub-band PCDs are more effective under mismatch conditions. This is a particularly exciting finding for FVC, as in most forensic cases, the two recordings to be compared

have been recorded on the different devices and transmitted through different channels. Telephone speech recordings are often compared to direct microphone recordings of police interviews. This channel mismatch has been long recognised as a hindrance to effective speaker classification. Various techniques for channel compensation have been proposed (e.g. [16-18]), but they all appear to require building a channel-characteristics model. However, since mobile transmission technology has an inherently highly variable signal processing path [15], the effectiveness of such approach may be limited. Further, crime-scene recordings are often very short. Thus the recording in question may not contain sufficient information to build a usable channel-characteristics model. Given all these constraints, it seems more practical to seek robust features against channel mismatch, rather than to attempt to compensate for this, at least in FVC casework contexts. The results from the current study seem to indicate that sub-band PCDs are promising features in this regard.

In observing the relationship between the full-band PCDs and the sub-band PCDs, we also found that, in particular combinations of the sub-band range and the frequency regions, the sub-band PCDs perform almost as well as the full-band PCDs, or occasionally better. This confirms the observation made in relation to F-ratios: the sub-band PCDs are likely to outperform the full-band PCDs in speaker classification, especially when multiple sub-bands are combined as partially independent sources of information.

5. Conclusion

This study explored the potential of the sub-band PCD as a speaker classification feature, using a large Japanese vowel dataset. Observations of F-ratios and verification rates revealed some promising characteristics of the sub-band PCDs. Firstly, the behaviour of sub-band PCDs is mostly predictable from our knowledge of articulatory and acoustic phonetics. This is all the more significant because of PCDs band-limited to formant ranges afford more direct articulatory interpretations than the typically-measured full-band cepstra. Secondly, sub-band PCDs seems more robust against channel mismatch than full-band PCDs. This is a welcome finding as most FVC casework involves speech data recorded under mismatch conditions.

The findings reported here warrant us to proceed to the next step: LR-based evaluation and FVC experiments based on this feature, which is presented as the second part of this study.

6. Acknowledgements

The work presented here was partly supported by JSPS KAKENHI Grant Number JP18H01671, JP25350488.

7. References

[1] G. S. Morrison, "Forensic voice comparison and the paradigm shift," *Science and Justice,* vol. 49, pp. 298-308, 2009.

[2] G. S. Morrison, "Tutorial on logistic-regression calibration and fusion:converting a score to a likelihood ratio," *Australian Journal of Forensic Sciences,* vol. 45, pp. 173-197, 2013/06/01 2012.

[3] G. S. Morrison, "Distinguishing between forensic science and forensic pseudoscience: Testing of validity and reliability, and approaches to forensic voice comparison," *Science & Justice,* vol. 54, pp. 245-256, 5// 2014.

[4] D. A. van Leeuwen and N. Brümmer, "An introduction to application - Independent evaluation of speaker recognition system," in *Speaker Classification*. vol. 1, C. Müller, Ed., ed Berlin: Springer, 2007, pp. 330--353.

[5] P. J. Rose, T. Osanai, and Y. Kinoshita, "Strength of forensic speaker identification evidence: Multispeaker formant and cepstrum-based segmental discrimination with a Bayesian likelihood ratio as threshold," in *The 9th Australian International Conference on Speech Science & Technology* Melbourne, 2002, pp. 303-308.

[6] P. J. Rose, D. Lucy, and T. Osanai, "Linguistic-acoustic forensic speaker identification with likelihood ratios from a multivariate hierarchical effects model: A “non-idiot’s bayes” approach," in *the 10th Australian International Conference on Speech Science & Technology*, Sydney, 2004, pp. 402-407.

[7] E. A. Alzqhoul, B. B. Nair, and B. J. Guillemin, "Comparison between Speech Parameters for Forensic Voice Comparison Using Mobile Phone Speech," in *The 15th Australasian International Conference on Speech Science & Technology*, Christchurch, 2014.

[8] C. Zhang, G. S. Morrison, F. Ochoa, and E. Enzinger, "Reliability of human-supervised formant-trajectory measurement for forensic voice comparison," *The Journal of the Acoustical Society of America,* vol. 133, pp. EL54-EL60, 2013.

[9] M. Duckworth, K. McDougall, G. de Jong, and L. Shockey, "Improving the consistency of formant measurement," *International Journal of Speech, Language & the Law,* vol. 18, pp. 35-51, 2011.

[10] G. K. Vallabha and B. Tuller, "Systematic errors in the formant analysis of steady-state vowels," *Speech Communication,* vol. 38, pp. 141-160, 9// 2002.

[11] F. Clermont and P. Mokhtari, "Frequency-band specification in cepstral distance computation," in *The 5th Australian International Conference on Speech Science & Technology* 1994, pp. 354-359.

[12] F. Clermont, Y. Kinoshita, and O. Takashi, "Sub-band cepstral variability within and between speakers under microphone and mobile conditions: A preliminary investigation," in *The 16th Australasian International Conference on Speech Science & Technology*, Sydney, 2016.

[13] H. Makinae, T. Osanai, T. Kamada, and M. Tanimoto, "Construction and preliminary analysis of a large-scale bone-conducted speech database," *IEICE technical report,* vol. Speech 107, pp. 97–102, 2007.

[14] Y. Kinoshita, "Testing Realistic Forensic Speaker Identification In Japanese: A Likelihood Ratio Based Approach Using Formants," PhD, Linguistics, The Australian National University, Canberra, 2001.

[15] E. A. Alzqhoul, B. B. Nair, and B. J. Guillemin, "Impact of dynamic rate coding aspects of mobile phone networks on forensic voice comparison," *Science & Justice,* vol. 55, pp. 363-374, 2015.

[16] A. A. Garcia and R. J. Mammone, "Channel-robust speaker identification using modified-mean cepstral mean normalization with frequency warping," in *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*, 1999, pp. 325-328.

[17] D. A. Reynolds, "Channel robust speaker verification via feature mapping," ed: I E E E, 2003, pp. II-53-6.

[18] A. Solomonoff, W. M. Campbell, and I. Boardman, "Advances in channel compensation for SVM speaker recognition," in *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on*, 2005, pp. I/629-I/632 Vol. 1.