Journal of Phonetics 94 (2022) 101177

Contents lists available at ScienceDirect

Journal of Phonetics

journal homepage: www.elsevier.com/locate/Phonetics

Research Article

Sub-band cepstral distance as an alternative to formants: Quantitative evidence from a forensic comparison experiment

Yuko Kinoshita [a,*], Takashi Osanai [b], Frantz Clermont [a,c]

[a] Australian National University, Canberra, ACT 2601, Australia
[b] National Research Institute of Police Science, 6-3-1, Kashiwanoha, Kashiwa-shi, Chiba 277-0882, Japan
[c] J.P. French Associates Forensic Speech and Acoustic Laboratory, 86 The Mount, York YO24 1AR, United Kingdom

ARTICLE INFO

Article history:
Received 26 April 2021
Received in revised form 28 June 2022
Accepted 11 July 2022
Available online 10 August 2022

Keywords:
Forensic voice comparison
Sub-band
Parametric cepstral distance
Formants
LPCC
Likelihood ratio
Vowel

ABSTRACT

This paper demonstrates the potential of the sub-band parametric cepstral distance (PCD) formulated by Clermont and Mokhtari (1994), as an alternative to formants in acoustic phonetic research. As a cepstrum-based measure, the PCD is automatically and reliably extracted from the speech signal. By contrast, formants are time-consuming and often difficult to estimate, a well-known bottleneck for studies based on large-scale datasets. The PCD measure gives flexibility in selecting the frequency limits of any sub-band of interest within the available full band. We suggest that, if sub-band selection were guided by the acoustic–phonetic theory of speech production, PCD analysis could facilitate phonetically meaningful cepstral comparisons without relying directly on formants. We evaluate this idea by exploiting the PCD properties in the context of forensic voice comparison as an application example. The cepstral data were obtained from the vowels uttered by 306 male Japanese speakers. Similar patterns of results were observed using formants and sub-band PCDs, the latter yielding better performance. This suggests that sub-band PCDs are able to capture the spectral characteristics that we normally quantify through formants, but with better reliability and efficiency. The PCD results reported here are encouraging for other types of acoustic phonetic studies in which comparisons of spectral characteristics are required.

1. Introduction

Over many decades, formants have played a critical role in acoustic phonetic research. Formants link acoustic and vocal tract configuration (e.g. Fant, 1971; Stevens, 1998). Vocal tract configuration determines the frequency characteristics in the spectrum of a speech signal, and formants represent these characteristics by identifying local spectral peaks. Since the configuration of the vocal tract is determined partly by one's anatomical dimensions and partly by articulatory gestures, formants have been useful for a wide range of phonetic inquiries and, consequently, they have formed the foundation for much empirical research.

The use of formants in forensic voice comparison (FVC) was a natural extension of this. Theoretically, formants are expected to contain rich information on speaker individuality. A speaker's anatomy determines a possible range of formants. The individuality of their voice includes their phonetic and phonological choices, and their habits with articulatory gestures, all of which are also reflected in formants. Unlike identifiers such as DNA or fingerprints, a human voice is not a direct record of a person's anatomical features, and therefore any decision based on it contains an increased degree of uncertainty. However, research on the relationship between formants and speaker characteristics has been conducted for many decades (e.g. Ingram, Prandolini, & Ong, 1996; Nolan, 1983; Nolan & Oh, 1996; Rose, 1998; Sambur, 1975; Stevens, 1971), and has shown that formants are useful in distinguishing speakers.

Working with formants comes with some challenges. Automatic extraction of formant data is known to be unreliable, requiring human supervision and checking. This is very time-consuming and also prone to measurer-dependent variability (Duckworth, McDougall, de Jong, & Shockey, 2011; Zhang, Morrison, Ochoa, & Enzinger, 2013). This has been a significant obstacle in conducting large scale acoustic phonetic studies based on formants.

* Corresponding author.
E-mail address: Yuko.Kinoshita@anu.edu.au (Y. Kinoshita).

https://doi.org/10.1016/j.wocn.2022.101177
0095-4470/

In the early 2000s linguists working on FVC introduced a paradigm shift in their evaluation systems: emulating the approach established in the field of DNA analysis, they embraced the likelihood ratio (LR) as the primary evaluation method (Kinoshita, 2001; Meuwly & Drygajlo, 2001; Rose, 2002; for more details, see Morrison, 2009). The LR-based approach requires substantial datasets in addition to the speech recordings to be tested. First, LR calculation requires a sufficiently sized background population dataset from a relevant population, so that the observed difference between two speech samples can be evaluated in the light of its typicality in the population (e.g. Robertson, Vignaux, & Berger, 2016; Rose, 2002). Second, the essential calibration and validation of the resulting LR requires additional datasets with appropriate characteristics for the case (Morrison, 2018; Morrison, Enzinger, Ramos, González-Rodríguez, & Lozano-Díez, 2020; Ramos, Haraksim, & Meuwly, 2017).

Formants have featured in FVC research using LRs (e.g. Hughes & Foulkes, 2014; Hughes et al., 2018; Morrison, 2011; Morrison & Kinoshita, 2008; Morrison, Zhang, & Rose, 2011; Rose, 2017; Rose, Osanai, & Kinoshita, 2002; Rose & Winter, 2010), but preparing adequate formant datasets has been a significant challenge.

To overcome this issue, researchers in speech engineering have been using cepstral features (Garcia & Mammone, 1999; Juang, Rabiner, & Wilpon, 1987; McLaughlin, Reynolds, & Gleason, 1999; Reynolds & Rose, 1995; Tohkura, 1987). Mel-Frequency Cepstral Coefficients (MFCC) and various derivations of them have been the main acoustic feature used (c.f. Morrison et al., 2020). Cepstral coefficients are a mathematical representation of the spectral shape that acoustic signals exhibit in the selected frequency range. They can be extracted automatically and reliably, enabling fast, replicable and large-scale feature extraction. This capacity improves the statistical reliability of the LR-based assessment.

In addition, cepstral features by nature contain richer information. Cepstral features utilise all frequency information within the analysis frequency range, whereas formants represent only a discrete point in each spectral peak. Cepstral features thus reflect the acoustical effects of vocal tract configurations more completely than formants do. This makes cepstral features generally more powerful as FVC features: many speaker comparison experiments report that cepstral features generally outperform formants (e.g. Alzqhoul, Nair, & Guillemin, 2014; Enzinger & Morrison, 2017; Rose, Lucy, & Osanai, 2004; Rose et al., 2002).

Then why are cepstral features not used widely in linguistic phonetics research? This is because cepstral features lack a characteristic that is of critical importance to phoneticians: relatability to linguistic observations, such as articulatory gestures or phonetic and phonemic variations. Cepstral features are mathematically derived from the overall spectral shape. While they reflect the vocal tract configurations in total, the numerical values obtained from it do not directly relate to individual resonances created by different part of vocal tract.

To address this issue, and more broadly to seek a more efficient approach to quantifying and analysing acoustic phonetic information, we put forward the use of sub-band PCD. This approach applies band-limited analysis to linear prediction cepstral coefficients (LPCC), and was first proposed by Clermont and Mokhtari (1994). It compares the shapes of two spectral envelopes within a user-defined frequency range (sub-band) and returns the average difference between the two as a score (PCD). Its three key properties are that:

(1) it is automatically and reliably extractable;
(2) the selection of the sub-band is completely flexible within the full band range of the original LPC analysis;
(3) it is computationally efficient, as once the LPCCs are extracted, the PCD can be calculated from any sub-band range without re-analysing the original speech signals.

What these properties suggest is that, with phonetically motivated sub-band selection, sub-band PCD may detect differences in spectral characteristics (as formants do), but far more efficiently. In fact, formant measurement can be regarded as a kind of a sub-band approach. For instance, if we are looking for the F1 of a male speaker's /i/ vowel, we look for a spectral peak located around the 300 Hz region, and do not look around 1000 Hz. We focus our analysis on a particular frequency range, motivated by our phonetic knowledge. The same idea can be applied to sub-band PCDs. In selecting a suitable frequency region (sub-band), both formants and the sub-band PCD represent information on a particular aspect of a vocal tract configuration.

This motivated us to conduct a feasibility study as will be reported here. In this study, we aim to explore the question "can sub-band PCD be an alternative to formants in acoustic phonetic research?", using vowel data of 306 male Japanese speakers from the NRIPS database (Makinae, Osanai, Kamada, & Tanimoto, 2007). We conducted comparative FVC experiments and observed the LRs obtained from F1, F2 and F3 and from sub-band PCDs calculated from the equivalent frequency ranges. We examined how the results from the sub-band PCD based system relate to the formant-based system, as well as overall performance.

Our experiment adopts a common scenario in forensic casework: channel mismatch comparison (mobile phone versus microphone recordings), as we postulate that sub-band PCD may be particularly useful in this scenario (see the discussion in Section 2).

The subsequent sections of this paper are organised as follows. The methodology section is divided into three parts: Section 2 presents the mathematical formulation and development of sub-band PCD; Section 3 describes the data and the approaches to acoustic feature extraction; Section 4 discusses our approach to LR calculation. Section 5 presents the results, and Section 6 and 7 present discussions and conclusions.

2. Sub-band parametric cepstral distance (PCD)

The PCD is a band-selective formulation of the index-weighted cepstral distance (Clermont & Mokhtari, 1994). It is based on the all-pole linear-prediction (LP) model of speech production and has the ability to produce sub-band distances between two sets of standard LP-cepstral coefficients (LPCCs). Sections 2.1 and 2.2 introduce the PCD formulation and its properties. The mathematical expression for the PCD is outlined in Eq. (3) and expanded in Appendix A. Section 2.3 illustrates the PCD using LPCCs from the Japanese vowel data employed for this work. Section 2.4 reviews previous

studies based on sub-band PCD, and Section 2.5 discuss its particular relevance to FVC.

2.1. Index-weighted cepstral distance: full-band formulation

At the core of the PCD is the index-weighted cepstral distance (Yegnanarayana & Reddy, 1979) expressed below in Eq. (1). C_k and C'_k represent two LPCC vectors, and M denotes the LP-analysis order.

$$D^2_{\mathrm{NDPS}}(0,\pi) = \frac{1}{2}\sum_{k=1}^{M}\left[k\left(C_k - C'_k\right)\right]^2 \tag{1}$$

Scaling each LPCC by its index k to yield the sequence kC_k is a direct consequence of the Negative Derivative of the LP-Phase Spectrum (NDPS). This makes D^2_{NDPS} equivalent to the Euclidean distance between a pair of cepstrally-smoothed NDPS, as formulated in Eq. (2). The full-band expression in Eq. (1) follows from Eq. (2) by setting the limits θ_1 and θ_2 of the integral to 0 and half the sampling frequency π, respectively.

$$D^2_{\mathrm{NDPS}}(0,\pi) = \frac{1}{(\theta_2-\theta_1)} \times \int_{\theta_1=0}^{\theta_2=\pi}\left[\left(-\frac{d\varphi(e^{j\theta})}{d\theta}\right) - \left(-\frac{d\varphi\prime(e^{j\theta})}{d\theta}\right)\right]^2 d\theta \tag{2}$$

where:

$$-\frac{d\varphi(e^{j\theta})}{d\theta} = \sum_{k=1}^{M} kC_k\cos(k\theta)$$

$$-\frac{d\varphi\prime(e^{j\theta})}{d\theta} = \sum_{k=1}^{M} kC'_k\cos(k\theta)$$

2.2. Index-weighted cepstral distance: band-selective formulation

A general expression $D^2_{\mathrm{NDPS}}(\theta_1,\theta_2)$ is easily derived from Eq. (2) by keeping θ_1 and θ_2 as parameters for the lower and upper bounds of a desired sub-band. Eq. (3) gives such an expression in matrix form (see Appendix A for further details). Note that we use the acronym PCD also to refer to the numerical value of $D^2_{\mathrm{NDPS}}(\theta_1,\theta_2)$ for the selected sub-band.

$$D^2_{\mathrm{NDPS}}(\theta_1,\theta_2) = [\mathbf{K}\cdot(\mathbf{C}-\mathbf{C}')]^T\cdot\mathbf{W}(\theta_1,\theta_2)\cdot[\mathbf{K}\cdot(\mathbf{C}-\mathbf{C}')] \tag{3}$$

The term **K** is the index-weighting matrix which renders the PCD between **C** and **C**′ (pair of LPCC vectors) sensitive to deviations around spectral peaks. This property arises from the NDPS emphasis on spectral slopes, which has proven effective for speaker-dependent and speaker-independent speech recognition under clean and noisy conditions (Hanson & Wakita, 1987; Hunt & Lefèbvre, 1989; Juang et al., 1987; Shikano & Itakura, 1992; Tohkura, 1987).

The sub-band matrix **W** is evidently the most critical term of Eq. (3). Its built-in parameters θ_1 and θ_2 give the *flexibility* of delimiting any sub-band with the only constraint that their values be contained within the sampling range ($0 \le \theta_1 < \theta_2 \le \pi$). Computational *efficiency* is another significant property, as the standard LPCCs are extracted only once and then re-used every time a sub-band is selected. Finally, *consistency* in spectral representation is also preserved since the PCD operates on the same LPCCs irrespective of the sub-band selected.

2.3. PCD illustrations

Figs. 1 and 2 demonstrate the PCD concept. In Fig. 1 the top panel shows the cepstrally-smoothed spectrum for the /a/ vowel produced twice by a single speaker. The bottom panel gives the profile of PCDs in 500 Hz contiguous sub-bands. The two spectra differ mainly around the peaks in the 2500–3000 Hz sub-band and, as expected from phonetic knowledge, this is where the PCD is the largest.

Whether the PCD should be considered as an effective feature for speaker characterisation depends on its ability to produce larger values in different-speaker (DS) than in same-speaker (SS) comparisons. Fig. 2 illustrates SS results obtained for /a/ by comparing two repetitions from the same speaker (left panel), and DS results for the same vowel from two different speakers (right panel). The PCDs from the DS comparison are indeed larger than those from the SS comparison in certain sub-bands. Furthermore, the latter happen to coincide with the frequency ranges expected for the /a/ formants. These preliminary observations suggest that PCDs could be both effective and phonetically-interpretable if the choice of sub-bands were guided by phonetic knowledge of typical formant locations.

2.4. Previous applications of sub-band PCD

Sub-band PCD was first applied to a large dataset by Khodai-Joopari, Clermont, and Barlow (2004). Using 297 male Japanese speakers recorded through a landline telephone system, they calculated sub-band PCDs and examined their F-ratios. The F-ratio calculates a between-group to within-group variability ratio. In the context of voice comparison, a high F-ratio indicates a greater speaker discrimination potential. Their study found high F-ratios in the frequency region of F3 and higher.

Sub-band PCD analysis allows its user to flexibly focus their analysis to their chosen frequency ranges. Looking at this differently, sub-band PCD can exclude unwanted frequency ranges from LPCC analysis. Clermont, Kinoshita, and Osanai (2016) proposed that FVC research can take advantage of this property. Forensic speech samples in real life often contain channel differences and background noise (e.g. Jessen, 2008). FVC performance may be improved by excluding frequency ranges which are likely to contain more noise than desired information. On this premise, they conducted a small-scale pilot study using Japanese vowels recorded through two channels: microphone and mobile phone transmission. The resulting F-ratios were not only consistent with the studies by Khodai-Joopari et al. (2004), but they also found that the F-ratio from the mobile recordings compared well to those from the microphone recordings. These findings suggest that sub-band PCD analysis has a capacity to incorporate linguistic phonetic knowledge, and to mitigate some of the negative impacts of channel mismatch by excluding those frequency ranges which contain mostly information which is unrelated to speaker characteristics.

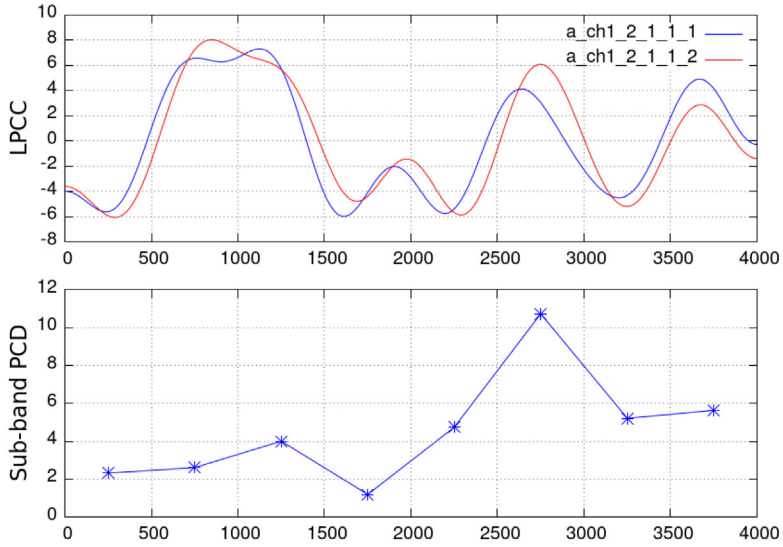

Fig. 1. Comparison of two tokens of /a/ produced by a single speaker.

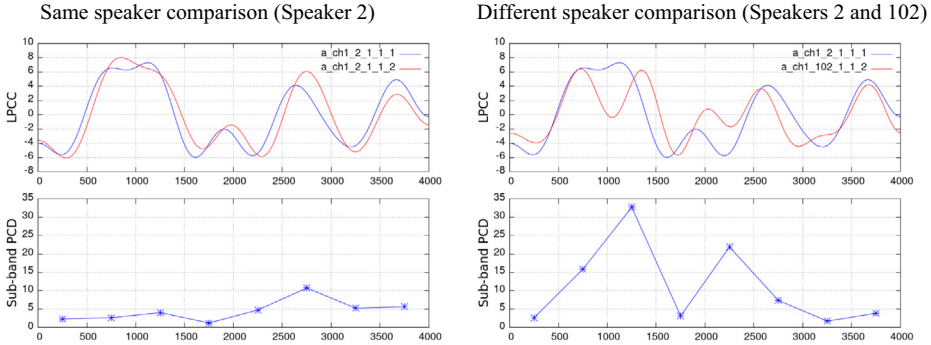

Fig. 2. Same-speaker (left) and different-speaker (right) comparisons of /a/ vowels.

Osanai, Kinoshita, and Clermont (2018) and extended this work using 306 male native speakers of Japanese in the NRIPS database used in the current study (Makinae, Osanai, Kamada, & Tanimoto, 2007). They observed F-ratios and how well the sub-band PCD itself can distinguish within-speaker comparisons from different-speaker comparisons with various sub-band ranges. They confirmed findings in Clermont et al. (2016): F-ratios were higher in the frequency regions where formants are expected to be located. Particularly high F-ratios were found in the F2 and F3 frequency ranges for the /i/ vowel, and the F3 range for /e/, mirroring the results of many of the previous formant-based FVC studies. They also found that the speaker verification rate under a channel mismatch condition degrades less for sub-band PCDs than for full-band PCDs, presumably due to exclusion of frequency ranges containing less speaker information. Building on this, Kinoshita, Osanai, and Clermont (2018) conducted an LR-based FVC experiment using sub-band PCD as the feature. The study found that, while LRs for individual sub-bands were close to 1, i.e. not strong, they appeared to be well calibrated.

Importantly, sub-band PCDs have also been shown to be useful outside of speaker comparison studies. Clermont and Kinoshita (2019) employed sub-band PCD in analysing how much and in which frequency ranges two types of information — speaker identity and phonological context — affect spectra of Japanese vowels. The results showed that context effects were most strongly present in the frequency range where we expect F2 to be, and the ratio of speaker-to-context effect was highest in the F3 frequency range or higher. This also supports classical findings that F2 is most affected by coarticulation (e.g. Lindblom, 1963; Stevens & House, 1963) and that F3 and above are most useful for forensic voice comparison (e.g. Rose, 2002).

Past research suggests that, with phonetically informed sub-band selection, sub-band PCD captures similar information to formants. Sub-band PCD analysis could be a very viable alternative methodology for capturing phonetic information in speech spectra, especially given the ease and reliability of extraction of this feature.

2.5. Relevance to forensic voice comparison

The previous section described *efficiency* and *flexibility* as the key characteristics of the sub-band PCD method. Here we describe the relevance of these properties to its application to forensic voice comparison (FVC).

The first property, efficiency, offers analysts capacity to acoustic information automatically and quickly. The feature

extraction is also easily *replicable*. This contributes to the credibility of assessments, invaluable in LR-based evaluation.

The flexibility of analysis range is also a great advantage. It is well-reported that speaker characteristic information is not evenly distributed across the spectrum. Higher frequency regions contain more speaker information, whereas lower frequency regions contain more linguistic information (e.g. Clermont & Mokhtari, 1998; Furui & Akagi, 1985; Goldstein, 1976; Lu & Dang, 2008; Mokhtari & Clermont, 1994; Pols, Tromp, & Plomp, 1973; Saito & Itakura, 1982; Stevens, 1971). This suggests that selective focus on higher frequency regions can be advantageous for FVC. Reliable detection of spectral peaks in higher frequency range is challenging, due to their lower energy. Sub-band PCD can capture higher frequency information more robustly as it takes advantage of the whole shape information. It can also easily exclude frequency ranges which contain external noise, or little voice signal.

We also speculate that this band selectivity is advantageous in dealing with a common issue in FVC: recording channel mismatch. In FVC, analysts routinely need to compare samples which were recorded under very different conditions, e.g. a telephone recording from a crime scene and a direct microphone recording made at a police interview. Over the years, much research has been done on the impact of channel mismatch (e.g. Byrne & Foulkes, 2004; Künzel, 2001; Zhang, Morrison, Enzinger, & Ochoa, 2013), and various techniques have been proposed to compensate for different conditions (e.g. Garcia & Mammone, 1999; Reynolds, 2003; Solomonoff, Campbell, & Boardman, 2005). However, such techniques require building a channel characteristics model. Crime scene recordings are often very short, so the recording in question may not provide sufficient information for building a dependable model of channel characteristics, and the quality of background noise is often not consistent across the recording (e.g. Jessen, 2008). Mobile phone transmission characteristics change continuously, as the compression rate and methods change in response to network conditions (Alzqhoul, Nair, & Guillemin, 2015; Guillemin & Watson, 2008). In addition, numerous internet voice call options have become available, with endless variations in signal processing and stability. One may attempt to retrospectively 'match' the conditions by putting a non-telephone recording (such as a police interview) through a mobile codec or a phone network, but there is no practical way of knowing if the conditions applied are indeed similar to those of the forensic samples. Such adjustments introduce extra assumptions to the analysis process, which may be problematic if the result is to be used as court evidence.

In our view, the most practical way forward is to seek features which are less susceptible to external factors, and which are reliably measurable even in poor quality recordings.

3. Database and acoustic feature extraction

3.1. Speakers, speech materials, and recordings

This study used 306 adult male speakers from the NRIPS database (Makinae et al., 2007). They are native speakers of Japanese, aged from 18 to 76 years. The same data was used in the previous studies (Kinoshita et al., 2018; Osanai et al., 2018). The speakers had varied dialectal backgrounds. However, Okuda (2005) suggests that dialectal variations affect vowel formants very little in modern Japanese, therefore it is unlikely to create unrealistically favourable conditions for the speaker classification task.

As an exploratory study, we prioritised including a large number of speakers over testing with forensically realistic data. For ease of feature extraction, we used read-out (C)V syllables as the target speech material. Each target syllable was composed of a selected consonantal environment followed by one of the five vowel phonemes of Japanese, /a/, /e/, /i/, /o/ and /u/. The consonantal environment was selected based on the ease of automatic segmentation. The selected consonantal environments are: ∅ (no consonant), /k/, /s/, /t/, /h/, /r/, /g/, /z/, /d/, /b/, and /p/; excluding the phonemes /n/, /m/, /j/, and /w/ from analysis for ease of automatic segmentation.

The Japanese *kana* syllabary writing system maintains a dis- tinction between the pairs ぢ /di/ – じ /zi/ and づ /du/ – ず /zu/, but each pair are phonetically identical, realized as [dzʲi] and [dzɯ] respectively. This resulted in vowel data of 10 different phonological contexts for /i/ and /u/; and 11 for /a/, /e/, and /o/.

This is controlled read-out speech, and thus not forensically realistic. However, the focus of this study is the comparison between performance of formants and sub-band PCDs, so we consider the current data appropriate for our purpose.

All speakers in the NRIPS database were recorded on two separate recording sessions, two to three months apart to emulate an aspect of FVC – the two speech recordings to be compared are always recorded at two different occasions. They performed the same recording tasks twice at each recording session, and the whole process was recorded simultaneously through three channels: direct microphone (channel 1: Ch1), bone-conducting microphone (channel 2: Ch2), and transmitted through a mobile phone network (channel 3: Ch3). This study focuses on the cross-channel comparisons between the direct microphone recording (Ch1) and the mobile phone recording (Ch3), chosen for their relevance to forensic casework. Further detail of the comparison setup will be discussed in the subsequent section (4.1 Comparison setups).

As a result, for each speaker, we had 40–44 tokens per vowel for each of two channels.

3.2. Acoustic feature extraction

3.2.1. LPCC

The target syllables were automatically segmented into a preceding consonant and a vowel based on their amplitude and F0. The sound files were down-sampled from 44.1 kHz to 8 kHz.

In selecting LP order, we followed the guidelines by Markel and Gray (1976) to capture the spectral characteristics for voiced speech: (1) LP order should be at least equal to the sampling frequency in kHz, with the aim of securing enough poles to capture the expected number of formant peaks; and (2) an additional 2 to 5 should be added to model other spectral-shape features resulting from glottal and lip radiation. The LP order was accordingly set to 14, and full-band LPCCs were extracted from the selected vowel sections using Hamming window, window length 25 ms, and time-step 5 ms.

The selection of the LP order is known to impact formant measurements (Dissen, Goldberger, & Keshet, 2019; Harrison, 2013). This is not such a critical issue as far the reliability of LPCCs is concerned. As long the LP order is defined according to the criteria mentioned above, the PCD is able to capture essential features of the smoothed spectra implied by the LPCCs, including the formant peaks that determine their shape.

The LPCCs were averaged across the vowel duration, and further averaged across 10 (for /i/ and /u/) or 11 (for /a/, /e/, and /o/) different phonological contexts for each vowel. As a result, for each speaker we obtained the four sets of mean LPCCs (two recording sessions, two repeats) for five vowels and two recording channels.

We note that our methodology is based on the use of LPCCs, though MFCCs are more commonly used in speech science. Our choice of LPCCs is motivated by three factors. First, we have adopted the all-pole LP-model for its well-known ability to capture the spectral peak characteristics of non-nasalised voiced sounds. LPCCs are therefore expected to carry acoustic–phonetic information that is particularly relevant to forensic voice comparison based on vowels as attempted here. Second, our standard LPCCs are derived on a linear-frequency scale by contrast with the MFCC which suppresses potentially important information in the high-frequency range. Third, the PCD formulation itself is tightly linked to the mathematical basis for deriving LP-cepstral distances as discussed in Section 2.

LPCC is a feature based on vocal tract modelling, and thus it is inherently linked to the speaker's anatomy and articulatory habits. MFCC, on the other hand, is more reflective of our perceptual response, being represented on an auditory scale. We thus consider the use of LPCC theoretically more appropriate for our purpose: characterising speakers.

3.2.2. Formants

Since the target segments for this study are monophthongs, we postulate that the target formants can be reasonably represented by the most typical values of those extracted from multiple sampling points across the duration of a given vowel. Based on this premise, we developed a new procedure for semi-automatic formant extraction, which employs a systematic and replicable approach for correcting the outputs of automatic peak extraction, simulating human intervention. We present brief descriptions of the six steps of our formant extraction procedure (Fig. 3 illustrates the first three steps, and see Appendix for the full descriptions of the methodology):

(1) **Automatic spectral peak extraction.** The frequencies of the spectral peaks for each utterance were extracted using the *formant listing* function of *Praat* (Boersma & Weenink, 2017). The analysis range was 0–4000 Hz, sampled every 5 ms, using the Burg formant algorithm with 50db pre-emphasis. Although only the first four formants are of interest for this analysis, we extracted additional peaks in case one was a false reading, thus ensuring that we had extracted the necessary peaks. We tried extracting five and six peaks. Inspecting formant tracking overlayed on spectrographs for several speakers, we decided five peaks setting produces more reliable tracking (Fig. 3-1).

(2) **Filtering outliers.** We produced histograms from the values from step 1 for each spectral peak. Values which fall outside the most populated bin and its adjacent ones were excluded from the analyses as outliers (Fig. 3-2).

(3) **Selecting the representative values.** The density function of the data obtained from step 2 was produced (Fig. 3-3a). Then, the peak and the two adjacent points were identified, and their x- and y- coordinates were extracted (Fig. 3-3b). Then, a quadratic function was fitted, and its maximum value was identified. The x-coordinate of this point was recorded as the spectral peak frequency for the utterance (Fig. 3-3c). At the end of this step,

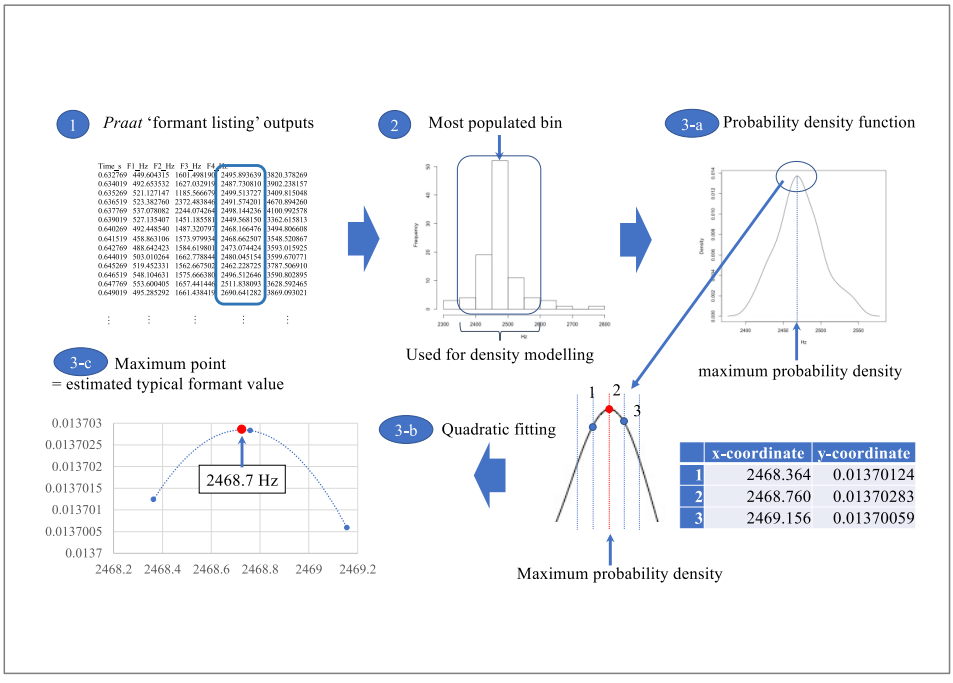

Fig. 3. Identifying the typical frequency value for each spectral peak (steps 1–3).

the list of time-series frequency values for each peak from step 1 was summarised into a single frequency value, deemed to be the most representative of the spectral peak.

(4) **Setting initial values.** In preparation for selecting F1, F2 and F3 from the four peaks, we first pooled all speakers' peak frequency values from step 3, for each peak separately. We then produced a histogram with 100 Hz bins for each combination of peaks 1–4 and each vowel (e.g. peaks 1 to 4 for /i/). We used the same process for the formant extraction for both Ch1 and Ch3 recordings, except we assigned the peaks from the Ch3 data to formants using the initial values from the Ch1 data. We chose to do so because the Ch1 data contained fewer acoustic artifacts and, therefore, is likely to provide more reliable reference points.

(5) **Assigning peaks to formants.** Assuming that there are no missing values, the four peaks can be assigned to the three formants in the four patterns as shown in Table 1. For each token, we selected the pattern in which the three peaks together showed the smallest distance from the initial values for each formant. After performing this process to all tokens, we went back to step 4, updating the overall means and replacing the initial values with these updated means. We repeated the peak assignment process and re-examined which of the four patterns were the closest to these model population formant values, until the population mean formant values stabilised.

(6) **Assigning missing values.** As the final step, we calculated the population mean and standard deviation for F1, F2 and F3, and assigned NA to any values which were outside of ±4 standard deviations — that is, any values outside of 99.99% of the distribution.

Table 2 presents the resulting means and standard deviations of F1, F2 and F3 across 306 speakers for the microphone recordings (Ch1) and the mobile recordings (Ch3). We visually examined Praat formant tracks for some speakers, and our methods appeared to return values which corresponded well to what we observed in formant tracks. Our approach was guided by the idea that, for monophthongs, taking the most typical formant value is theoretically a better representation of the target vowel than taking a midpoint value or an average across duration. As the "typical formants" are something that we mathematically calculated over many sampling points, they are not directly comparable to manual measurements.

However, it is reassuring to note that the central tendency and the variability reflected in these summary statistics are comparable to those reported by Mokhtari (2000) for the same Japanese vowels embedded in consonantal contexts and recorded by a heterogeneous speaker set. It is also worth observing that the ratios of standard deviations to means in Table 2 lie comfortably within the 6–13% range of difference limens found for human perception of formant frequencies of coarticulated vowels (Mermelstein, 1978; Nakagawa, 1982).

We add that we do not expect manual correction to improve our formant extraction. We believe that human manual corrections will not be consistent across data of this size and would instead result in introducing noise to the data. It is prohibitively time consuming to manually correct data of this size.

Fig. 4 shows the distributions of the extracted formants. Ch1 and Ch3 are presented in black and red lines, respectively. It reveals that some formant-vowel combinations were affected by mobile phone transmission more than others. /a/ appeared to be affected significantly across all three formants, whereas for /u/ none of the three formants showed much impact from the transmission channels at all. It is noteworthy that, apart from the /a/ vowel, the difference in recording channel appeared not to affect F1 greatly, unlike previous studies which reported F1 measurements to be particularly susceptible to the effect of mobile phone transmission (Byrne & Foulkes, 2004; Guillemin & Watson, 2008).

To statistically examine the effects of mobile transmission, we built linear mixed effect models. We assigned recording channel, recording session and phonological contexts as the fixed effects; and speaker as the random effect. The phonological contexts effect was nested in the recording sessions effect in this model. We modelled each vowel and formant combination separately, making 15 models in total. For every model, the recording channel was found to have a very strong impact; all of them produced extremely low p-values ($<2.2e^{-16}$). When a statistical test is repeated many times, as repetition of the test increases the possibility of achieving statistical significance by chance. We thus need to be mindful in interpreting outputs here. However, in the case of current study, the extremely low p-values suggest that the statistical significance is

Table 1
Possible patterns of peak-to-formant assignment.

Assignment patterns	peak1	peak 2	peak 3	peak 4
pattern 1	F1	F2	F3	
pattern 2	F1	F2		F3
pattern 3	F1		F2	F3
pattern 4		F1	F2	F3

Table 2
Mean formants and standard deviations, averaged across 306 speakers, in Hertz.

Vowels		Ch1			Ch3		
		F1	F2	F3	F1	F2	F3
/a/	Mean	693	1280	2615	671	1297	2504
	SD	88	101	200	59	143	147
/e/	Mean	491	1906	2548	511	1949	2704
	SD	45	168	228	47	198	267
/i/	Mean	376	2148	2964	363	2131	2819
	SD	39	182	213	39	175	245
/o/	Mean	490	916	2673	494	881	2424
	SD	51	109	207	46	78	243
/u/	Mean	394	1395	2357	397	1425	2369
	SD	36	160	224	33	158	170

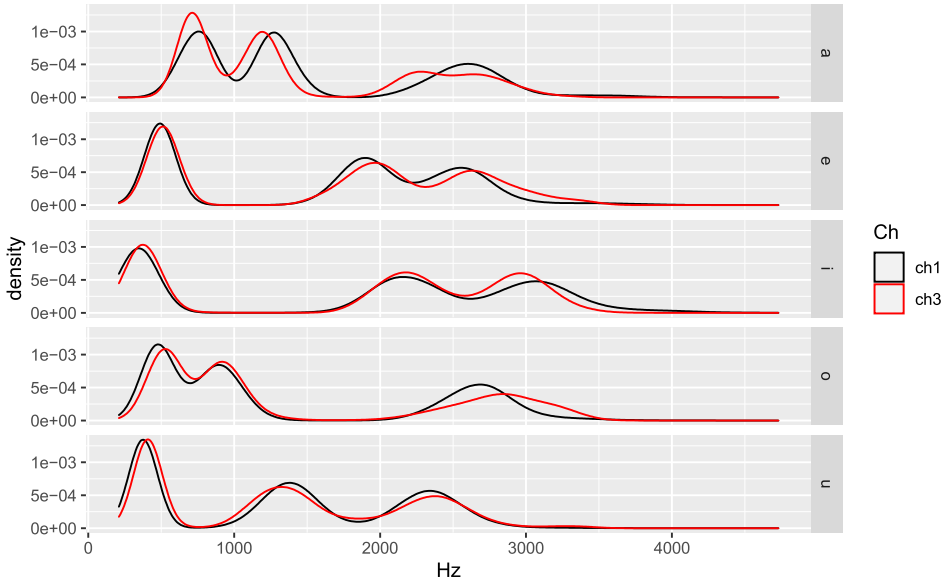

Fig. 4. Population distributions of F1–3 extracted from the two recording channels.

unlikely to be accidental. The mobile phone transmission appears to have a significant impact on formants – a concerning result for forensic phoneticians, as raised by Byrne and Foulkes (2004).

4. Likelihood ratio calculation

4.1. Comparison setups

We divided 306 speakers into three datasets: testing data (100 speakers), development data (100 speakers) and background population (106 speakers) to create experimental datasets independent of each other. Both testing and development data produced 100 same-speaker (SS) pairs and 4950 different-speaker (DS) pairs each. All speakers in this database were recorded at two non-contemporaneous recording sessions and, at each session, they read out the same word list twice. The word list contained each of the target (C)V syllables once. This enabled us to compare each of the SS pairs in four different ways, and each of the DS pairs in eight different ways, as shown in Fig. 5. This produced 400 SS comparisons (100 speakers * 4) and 39 600 DS comparisons (4950 speaker pairs * 8) for both test and development datasets. Since this study focuses on cross-channel comparisons between the direct microphone recording (Ch1) and the mobile phone network recording (Ch3), all comparisons were made between the Ch1 and Ch3 recordings.

For each of these comparisons, FVC scores were calculated using three different methodologies:

1. Sub-band PCD scores produced with two different widths of sub-band ('1SD-PCD' and '2SD-PCD'),
2. PCD obtained from full-band LPCC ('Full'),
3. the Multivariate Kernel Density (MVKD) likelihood ratio (Aitken & Lucy, 2004) from the formants.

The parameters are summarised in Table 3 below.

4.2. PCD calculation

4.2.1. Selection of sub-band frequency ranges

To align the sub-band ranges to the frequency regions where F1, F2 and F3 are likely to be, we referred to the formant measurements presented in Table 2. For each vowel-formant combination, the mean and the standard deviation were calculated across the 306 speakers. PCD was then calculated using two different sub-band range calculations: mean ± 1 standard deviation ('1SD-PCD'); and mean ± 2 standard deviations ('2SD-PCD'). We have labelled the sub-band frequency ranges that correspond to F1, F2 and F3 as 'subF1′, 'subF2′ and 'subF3′. This is summarised in Table 4 below.

4.2.2. PCD to LR

A PCD is a numerical summary of the distance between two spectral slopes in a given frequency band. In other words, it is a form of a comparison score based on the similarity. Unlike formants, PCDs cannot be used as features for MVKD, as they are already a distance measure. Thus, we took PCD as a distance-based comparison score and calculated LRs from their distributions. This is one of the well-established approaches to LR calculation (e.g. Robertson et al., 2016). For our relatively large datasets, we postulate that the PCD values pooled across all the comparisons would provide us with a fair estimation of the population distributions: SS comparison PCDs for within-speaker variations, and DS comparisons PCDs for between-speaker variations. From these two distributions, we can derive an LR for each PCD that reflects both the similarity between the testing pair and their typicality in a given population.

In order to produce the evaluation scores in this approach, however, we first need to model the distributions. To find a suitable model, we tested the fit of four different distributions: Gaussian, Gamma, Weibull, and Log-normal. For both SS and DS comparisons of each of the combinations of vowel and analysis band, we examined the fit of their PCD distribu-

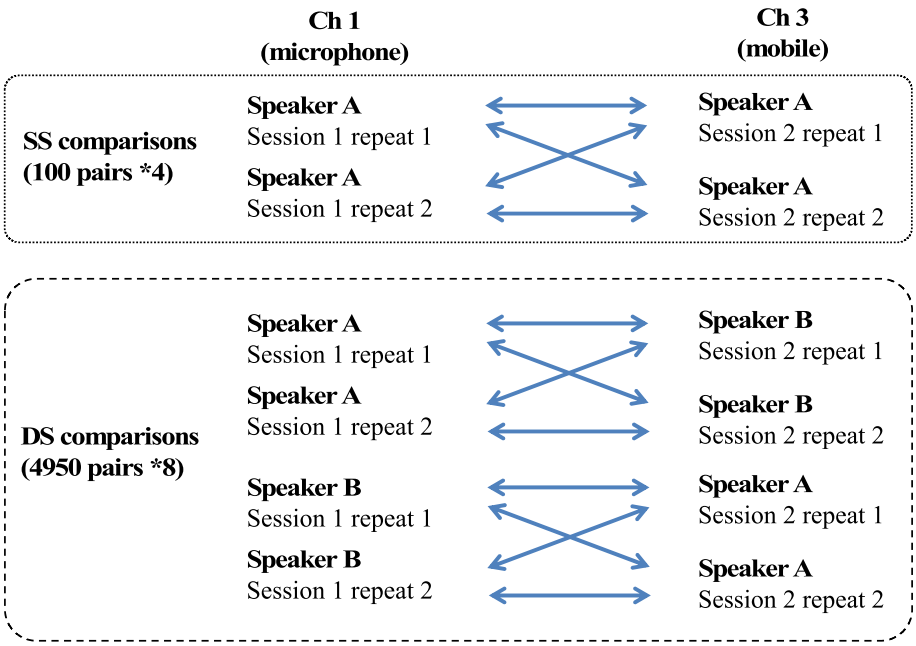

Fig. 5. Same speaker and different speaker comparisons made for this study.

Table 3
Summary of the variables used in the score calculations.

Label		PCD		Formants
Score calculation	1SD-PCD	2SD-PCD	Full	MVKD
Frequency range	subF1, subF2, subF3	subF1, subF2, subF3	N/A	F1, F2, F3
Target segments	5 vowels	5 vowels	5 vowels	5 vowels

Table 4
The sub-band frequency ranges used for sub-band PCD calculations (Hz).

		1SD-PCD			2SD-PCD		
Vowels		SubF1	SubF2	SubF3	SubF1	SubF2	SubF3
/a/	from	605	1179	2415	517	1078	2215
	to	781	1381	2815	869	1482	3015
/e/	from	337	1966	2751	298	1784	2538
	to	415	2330	3177	454	2512	3390
/i/	from	358	1235	2133	322	1075	1909
	to	430	1555	2581	466	1715	2805
/o/	from	446	1738	2320	401	1570	2092
	to	536	2074	2776	581	2242	3004
/u/	from	439	807	2466	388	698	2259
	to	541	1025	2880	592	1134	3087

Table 5
Number of instances each model was selected as the best fit. The maximum score for each cell is five (for five target vowels).

	Gaussian		Gamma		Weibull		Log-normal	
	SS	DS	SS	DS	SS	DS	SS	DS
Full	0	0	4	5	0	0	1	0
subF1	0	0	4	2	1	3	0	0
subF2	0	0	4	3	1	2	0	0
subF3	0	0	5	4	0	1	0	0
total	0	0	17	14	2	6	1	0

tions to these four models, using Akaike's Information Criterion (AIC) and Bayesian Information Criterion (BIC). Although AIC and BIC measure fitting slightly differently (Vrieze, 2012), their assessments were in complete agreement across all vowel and band combinations. Table 5 presents the counts of how often each model was found to be the best fit. Gamma distribution was found by far the best for modelling PCD distributions. As before SubF1, SubF2 and SubF3 denote the sub-bands corresponding to the frequency ranges of F1 to F3, and Full represents full-band LPC. SS and DS indicate the comparison types (see Table 6).

Thus, we adopted the Gamma distribution as our model. The PCD calculated for each vowel and sub-band combination was pooled separately for SS and DS comparisons and fitted with a Gamma distribution. The Gamma distribution is defined as below:

$$f(x) = \frac{1}{\Gamma(k)\theta^k} x^{k-1} e^{-x/\theta} \quad (4)$$

Evaluation scores were derived by taking the ratio between the probability of a given PCD belonging to a within-speaker PCD distribution; and the probability of the same PCD belonging to a between-speaker PCD distribution: an LR.

An example is shown in Fig. 6. Here, 'd' denotes the original distribution of PCDs, and 'g' denotes the fitted Gamma distribution. 'SS and 'DS' indicates the SS and DS comparisons, respectively. The PCD for this test pair was 2.5. Referring to the Gamma distributions modelled from the SS and DS com-

Table 6
C_{llr} produced from the full-band PCD system, the 1SD-PCD and 2SD-PCD systems with three sub-bands fused, and the formant-based system.

Vowel	1SD-PCD	2SD-PCD	Full-PCD	Formant
a	0.888	0.817	0.908	0.919
e	0.803	0.768	0.862	0.893
i	0.731	0.677	0.845	0.844
o	0.949	0.904	0.976	0.956
u	0.695	0.659	0.806	0.899
mean	0.813	0.765	0.880	0.902

parison PCDs, the probability of this PCD belonging to the SS and DS distributions is estimated as 0.157203 and 0.072693, respectively. Taking the ratio between these two probabilities, we obtain 2.162 as the LR for this particular comparison.

We note that the LRs we produced here are based solely on the distance — i.e. similarity — between two speech spectra, and do not contain information on typicality. Thus the resulting LR does not fully capture the nature of the observed evidence. While this approach has been used widely, it has been reported that omission of typicality makes the resulting LR less accurate (c.f. Garton, Ommen, Niemi, & Carriquiry, 2020; Morrison & Enzinger, 2018). However, the aim of the current study is to examine whether or not sub-band PCDs can capture the information for which we thus far used formants, not the absolute performance of the sub-band PCD. Therefore, we consider this not to be a critical issue for the purpose of this study.

4.2.3. Formant-based comparison

Using exactly the same comparison setup, we conducted a formant-based FVC analysis by using data on the first three formants in the MVKD formula, using the '*comparison*' package of R (Lucy, Curran, & Martyna, 2020). MVKD produces an LR-like score from multiple variables, discounting the correlation among them. We calculated MVKD scores from F1, F2 and F3 for all five Japanese vowels separately.

4.3. Calibration and fusion

Next, we applied linear logistic calibration (Morrison, 2013) to the four sets of LRs obtained from the four FVC systems (1SD-PCD, 2SD-PCD, Full-band PCD and formant-MVKD systems), and converted to $\log_{10}$ LRs (LLR). The two sub-band PCD systems produced LLRs for three different sub-bands (subF1, subF2 and subF3). The LLRs from the three sub-bands were then fused by applying linear logistic regression fusion (Morrison, 2013). Since subF1 to subF3 are extracted from frequency regions broadly corresponding to F1 to F3, we deemed the LLRs produced by fusing them to be comparable to the LLRs from the formant-MVKD system, in which LLRs are calculated by combining F1 to F3. The full-band PCD system was compared to two sub-band systems to examine whether exclusion of certain frequency ranges is effective as we predicted.

To compare the FVC performance of these systems, we used C_{llr}. C_{llr} is a metric of the goodness of an FVC system (Brümmer & Du Preez, 2006; van Leeuwen & Brümmer, 2007). C_{llr} evaluates the cost of erroneous verification that the experimental scores produce, considering the strength of the scores that point to erroneous verification and penalising them accordingly. Being a metric of costs, a lower C_{llr} indicates a better system: closer to 0 is better, and greater than 1 suggests that the system is useless in classifying speakers. Calibration and calculation of C_{llr} requires a training dataset with known origin (i.e. whether the scores came from SS comparisons or DS comparisons), separate from the testing datasets. We therefore split our data; we used one half of the scores as testing data, and the other as the training data as in Fig. 7.

5. Results

To properly understand the behaviour of sub-band PCDs and the comparability of this method to formant analysis, this section provides detailed examinations of the relationship between the C_{llr}s and the contributing variables: sub-band frequency region, vowel, and sub-band width. As the first step, we

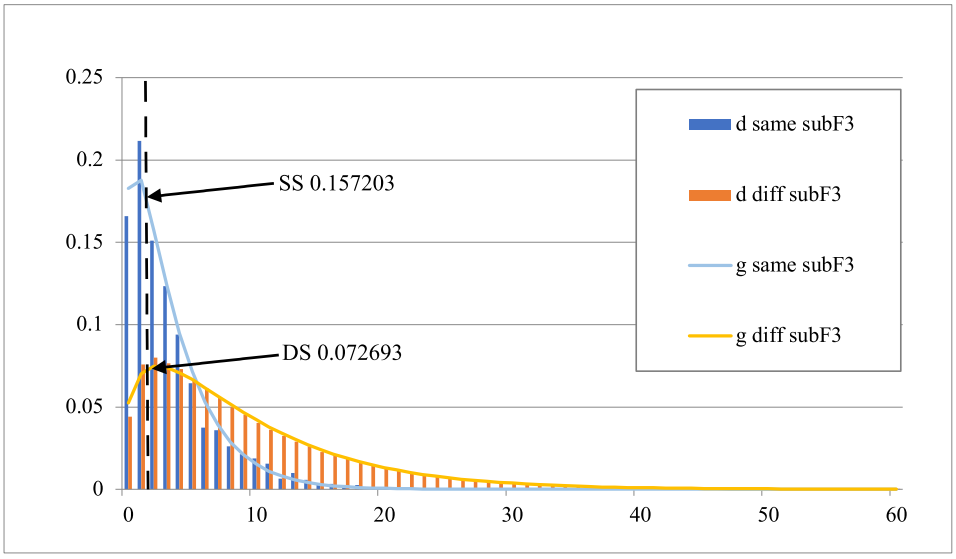

Fig. 6. Example score calculation based on the PCD distributions for /i/ subF3.

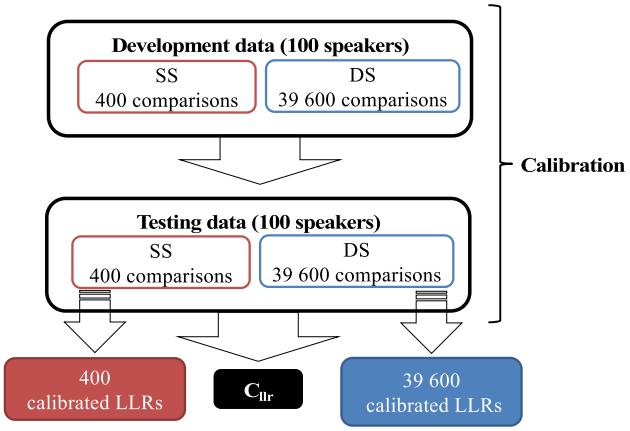

Fig. 7. Process of calibration and C_{llr} calculations.

examine the performance of the individual sub-bands (subF1, subF2, subF3) in relation to full-band PCD. We then combine the three sub-bands for each vowel, and finally fuse the LLRs from all five vowels to assess the overall FVC performance.

5.1. Band specific observations

At this stage, we do not know the optimal width of sub-band for capturing speaker information. This may not be the same across frequency regions and vowels. Thus we first compared C_{llr}s for each sub-band, produced under the two systems using different sub-band widths (1SD-PCD and 2SD-PCD). Once the best performing band width for each sub-band plus vowel combination was identified, we fuse the scores produced by these conditions, to attain the best possible final results.

Fig. 8 below summarises C_{llr} for the PCD produced from each sub-band range, corresponding roughly to expected frequency ranges for F1, F2 and F3 ('subF1', 'subF2', and 'subF3'). The results for the 1SD-PCD and 2SD-PCD system are presented in pairs for the ease of comparison. The horizontal line is the results from the full-band (0–4000 Hz) PCD. As C_{llr} is a cost measure, the lower is the better.

Across five vowels, the two systems reveal very similar tendencies, with slightly better results for the 2SD system. SubF1 seems to contribute very little. SubF2 and subF3 have lower C_{llr}s than subF1, in some cases comparable to the full-band results. SubF2 show generally better results for the front vowels /i/ and /e/. This finding corroborates a previous study on Japanese vowel formants; Kinoshita (2001) reported higher F-ratios for F2 of /i/, F2 and F3 of /e/. The strength of /i/ as a speaker discrimination feature has also been reported in a study on Australian vowels (Rose, 2007).

A high back vowel, /u/, appears to perform well too. Phonetically, F3 is often associated with lip rounding. Traditionally, lip rounding for /u/ in Japanese is associated with regional dialectal differences. However, Okuda (2005)'s study based on 3,771 Japanese speakers reports that there is no statistically significant regional difference in vowel formants in contemporary Japan. Our result agrees with the previous study by Khodai-Joopari et al. (2004) that examined cepstral F-ratios of Japanese vowels from 296 speakers, in which they reported the greatest F-ratio in the frequency region of F2 for /i/, relatively strong F-ratios in the F2 and F3 regions for /e/, and in the F3 region for /u/.

These observations seem to suggest that, with phonetically motivated sub-band selection, PCD allows us to extract similar information to that captured by formants, but more efficiently and reliably.

Compared to individual sub-bands, the full-band system performed better. However, in most cases the performance of subF2 and subF3 seem to be comparable to the full-band PCD. This suggests two things: sub-band PCD behaves in accordance with our theoretical understanding; and combining the information of multiple sub-bands — where more speaker information is contained — is likely to produce better FVC outcomes than individual sub-bands do, very much like with formants.

The comparison of the two systems, 1SD-PCD and 2SD-PCD, the 2SD-PCD system consistently produced lower (i.e. better) C_{llr}.

5.2. Sub-bands fused for each vowel

Next, the LLRs from the three sub-bands were fused to examine the performance of vowels as a whole. In theory,

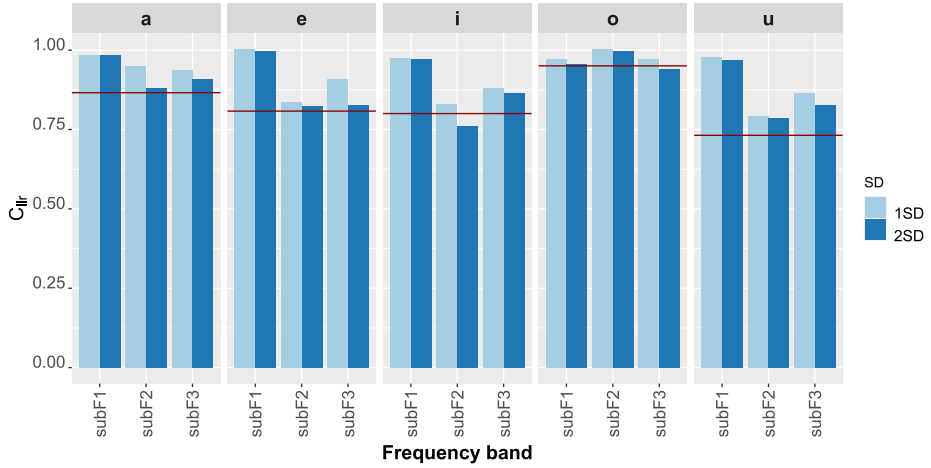

Fig. 8. Summary of C_{llr} for each frequency range (sub-band) and vowel. 1SD and 2SD represent the two different sub-band widths. The results from the full-band PCD system are presented as a horizontal line as the reference point.

Table 7
C_{llr} and EER for the three systems, five vowels fused.

	1SD	2SD	Formant
C_{llr}	0.567	0.470	0.738
EER	0.182	0.142	0.213

we can choose to fuse the LLRs produced with different sub-band ranges (1SD or 2SD) for each sub-band to get the best outcomes. However, the previous section found that the 2SD-PCD system consistently outperformed the 1SD-PCD system. So, we present the results of fusion for each system, rather than mixing the two systems. This section also presents the results from the full-band PCD system and the formant-MVKD system as the baseline for the comparison. Table 7 and Fig. 9 present C_{llr} for the five vowels produced by the four systems.

The 2SD-PCD system performed best for all vowels. They demonstrate that, with suitable band selection, the sub-band PCD system outperforms the full-band system, which validates our prediction that excluding the frequency ranges with little speaker information will improve the FVC performance.

Comparing the results for the sub-band PCD systems to formants, we see that the sub-band PCD systems outperform the formant system consistently across all five vowels. Importantly, the sub-band PCD-based systems and the formant-based system show a very similar pattern with respect to the relative performances of five vowels, i.e. which vowels are more effective in FVC in a given system, with the exception of /u/ vowel. The similarity of the pattern suggests that, with phonetically motivated sub-bands, PCD can capture similar information to formants.

As for the difference observed for /u/, we postulate that this was caused by the difficulty of formant measurements in the higher frequency region. Khodai-Joopari et al. (2004) found that for the /u/ vowel, the frequency region surrounding its expected F3 is particularly rich in speaker information. As the lower amplitude in this higher frequency region makes reliable detection of the spectral peak less accurate, the F3 measurements for /u/ in this study may have failed to fully reflect this. Sub-band PCD, on the other hand, is less affected, as it is derived from an entire spectral shape within the defined frequency range.

These observations suggest that sub-band PCD is not only capable of capturing the information for which we have used formants, but does so more robustly, especially in the frequency region where accurate formants measurements are difficult.

5.3. Five vowels fused for system comparison

As the final step, we fused the five vowels to compare the overall FVC performance. Here, we focus our observations on the three systems: 1SD-PCD, 2CD-PCD and formant-MVKD. Table 7 presents C_{llr} and Equal Error Rate (EER) for the three systems. EER is the point where the same-speaker and different-speaker comparisons produce an equal error rate, and this rate suggests how well the system can discriminate between speakers – the lower EER is, the better the discrimination capacity of the system. EER is not a preferred metric in current FVC research, but it has been used widely in speaker recognition research. Thus, we present EER as well as C_{llr} for ease of comparison. The two sub-band systems clearly outperform the formant-based system. The 2SD-PCD system had a better performance than the 1SD-PCD, consistent with the observations made in the previous two stages of the examination.

To gain further insights, we produced a Tippet plot (Fig. 10). It presents cumulative distributions of the LLRs obtained from SS and DS comparisons (SS and DS in red and black lines, respectively). The x-axis presents $\log_{10}$LR (LLR), and the y-axis is the cumulative probability density. The value read from the y-axis shows the probability of observing an LLR smaller (for SS comparison) or greater (for DS comparisons) than the given LLR indicated on the x-axis. The point of intersection of the SS and DS comparison

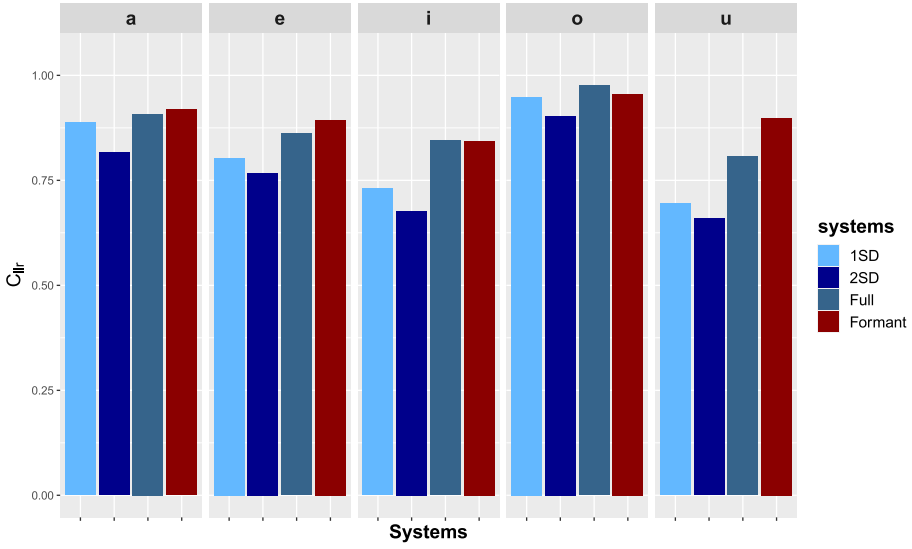

Fig. 9. C_{llr} for the four systems: 1SD-PCD, 2SD-PCD, full-band, and formant-MVKD.

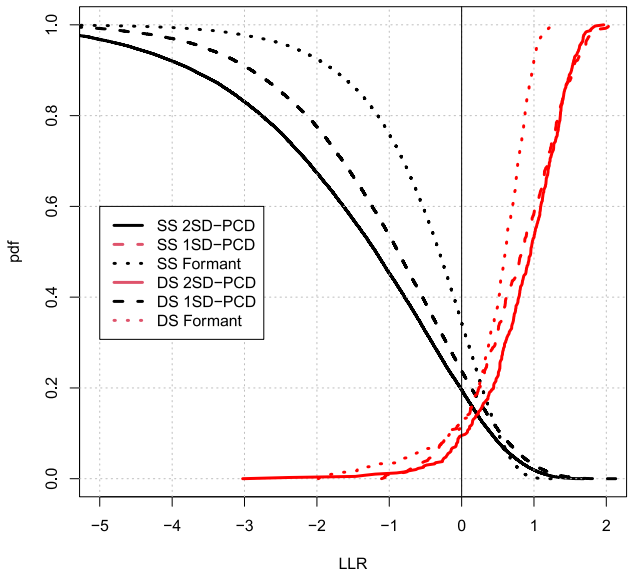

Fig. 10. Tippet plot of the LLRs, five vowels fused. The three systems were represented by different line types as seen in the legend. SS and DS comparisons are colour-coded in red and black respectively.

lines indicates the EER. The EER of a well-calibrated system should fall somewhere close to LLR 0. It is worth pointing out that the 2SD system has the best calibration, and the formant system had the worst.

Fig. 10 shows that the 2SD-PCD system separated the SS and DS comparisons the best, and the formant-based system worst. It indicates the 2SD-PCD is capable of producing stronger evidence than the formant-based system. The probability density of the intersection of the SS and DS comparison lines shows the EER. This is lowest for the 2SD-PCD system. While the LLR calculation approach used in this study has the limitation of not incorporating typicality, these results indicate promise for further development.

6. Discussion

6.1. Relevance to acoustic phonetic research

This study examined the potential of sub-band PCD as an alternative acoustic feature to formants, using FVC experiments as a testing ground. Sub-band PCD incorporates more complete spectral information than formants do, as it is derived from entire spectral shapes within the user defined frequency range, hence the differences in bandwidth and relative amplitude are fully accounted in PCD. This leads to a theoretical prediction that sub-band PCD would be as useful, or more useful than, formants in speech signal comparison.

Examining the performance of each combination of vowels and frequency ranges, we found that sub-band PCD gives a similar pattern of results to formants, but better C_{llr}. The previous studies based on formants reported that Japanese /i/ F2 and /e/ F2 and F3 discriminate speakers better. We found this also to be the case with sub-band PCD. The PCDs extracted from frequency ranges where formants would be expected to discriminate speakers better than the formants do. The observations for the /u/ vowel highlighted another advantage of sub-band PCD: its capacity to reliably characterise and compare two spectra, regardless of the existence of a prominent spectral peak. This suggests that sub-band PCD would be useful for comparing spectral information in the higher frequency range where reliable formant detection is difficult.

Sub-band PCD has an obvious application to speaker comparison tasks, as it is a distance score between two LPCC vectors. However, it has many other applications in acoustic phonetics. For example, Clermont and Kinoshita (2019) studied the relative contributions of speaker and co-articulation effects by calculating sub-band PCDs between individual speakers and the population mean LPCCs. As sub-band PCD is automatically extractable and flexible in sub-band selection, it can facilitate statistical analyses of observed phonetic differences in large datasets. Sub-band PCD may particularly be useful in studies on some articulatory gestures, such as lip rounding, which relates to higher formants. Also, we anticipate that cepstral mean subtraction, a technique developed for compensating for channel differences in the automatic speech recognition field (Westphal, 1997), could be used for normalisation of sub-band PCDs.

6.2. Sub-band PCD as a potential feature for FVC

The sub-band PCD systems outperformed the full-band PCD when the information from three sub-bands were fused. This indicates that FVC benefits from the exclusion of frequency ranges that do not contain much speaker information. The channel-mismatch comparisons in this study may have contributed to this. Conducting an FVC experiment in channel-matched conditions will be useful to test this assumption.

Sub-band PCDs in this study were calculated with two different width settings: mean ± 1 SD and mean ± 2 SD. We then compared results to find optimum bandwidths for each vowel and frequency region, since there is no theoretical imperative to use uniform bandwidths. However, we found that the 1SD-PCD system never performed as well as the 2SD-PCD system. We postulate that this was because mean ± 1 SD was not wide enough to fully capture between-speaker variability. The examination of Fig. 10 seems to support this hypothesis, as the difference between the 1SD-PCD and 2SD-PCD systems is more apparent in DS comparisons (black lines). Given that mean ± 1 SD and ± 2 SD cover 68% and 95% of their distributions respectively, further widening of the sub-band width is unlikely to improve FVC performance.

However, we note that the optimum sub-band width may differ from language to language. Japanese has a relatively sparse vowel space, having only five vowel phonemes. Languages with a more crowded vowel space, such as English, may require speakers to keep their vowel production variation tighter (c.f. Fletcher & Butcher, 2003). In such languages, it is possible that narrower sub-bands, such as 1SD-PCD, or even 1.5SD-PCD, perform better. In any case, the flexibility of sub-band selection is one of the advantages of this approach. Empirical examination of language-specific optimum sub-band width would be very useful.

Acoustic-phonetic FVC does not often use formants higher than F3, as it is difficult to reliably identify and extract them. Since sub-band PCD does not rely on detection of any particular point in the spectral shape, it opens a way to use these ranges. Cao and Dellwo (2019)'s study on Mandarin Chinese vowels, /i/, /y/, and /r/, reported that F4 and F5 were the most

effective formants in speaker classification. Cavalcanti, Eriksson, and Barbosa (2021)'s study on Brazilian Portuguese vowels also reported that F4 had the strongest speaker discriminatory power among the first four formants. Our study investigated only up to the frequency range equivalent to F3 to make it comparable to the baseline formant-MVKD system. However, there is no technical difficulty in including a higher frequency region in sub-band PCD calculations if the recordings contain higher frequency information, and this is likely to improve FVC performance. Also, formant detection for female speech is often more difficult than male speech due to wider spacing of harmonics. It is thus possible that a sub-band PCD system outperforms formant-based systems more strongly in the comparison of female speakers.

The current consensus in the FVC research community is that any system which produces a C_{llr} smaller than 1 contributes useful information (Morrison et al., 2021). In this light, our results for C_{llr} of 0.470 for 2SD-PCD suggests considerable promise in our approach, but how does this compare to the C_{llr}-s reported in other FVC studies? Here, we compare our results to some of the existing studies.

Differences in experimental conditions demand caution in comparisons. Meaningful comparison with results from early studies such as Rose et al. (2002) and Rose et al. (2004) is difficult, since these studies were conducted before calibration, fusion and C_{llr} became an integral part of FVC experiments. Thus the work by Alzqhoul et al. (2014) would be a good starting place. They analysed vowels /aɪ/, /eɪ/, and /i:/ extracted from the words "nine", "eight" and "three" read by 130 male speakers on two different occasions, using four different types of cepstral features – Complex Cepstral Coefficients, Real Cepstral Coefficients, MFCC, and LPCC – and the coefficients of Discreet Cosine Transformation (DCT) of formant trajectories as acoustic features. They simulated mobile phone conditions, by applying two of the most commonly used mobile codecs, CDMA and GSM, at high and low quality modes. They calculated LLRs using PCAKLR (Nair, Alzqhoul, & Guillemin, 2014). They found that MFCC performed best regardless of the codec types, with C_{llr} ranging from 0.108 to 0.127, depending on the type and the quality of codec, and the C_{llr} for the formant trajectories of 0.341 to 0.401.

These are substantially better than our results. However, part of the performance difference may be explained by channel conditions, since they compared speech samples processed under matching codec settings. The cepstral features used by them are likely to perform less well under channel mismatch conditions, as their approach does not distinguish speech signals from acoustic artefacts originated from the recording channels. Also, their speakers were recorded at four non-contemporaneous recording sessions at one-month intervals. This may have contributed to the performance, as it would permit better modelling of within-speaker variability.

Enzinger and Morrison (2017) also compared acoustic–phonetic approaches to automatic speaker recognition techniques. Their speech data were more forensically realistic: channel-mismatched spontaneous speech data. They used Australian female speakers. They produced evaluation scores from a Gaussian Mixture Model – Universal Background Model (GMM-UBM) built on MFCC as features and converted the scores to LRs by applying linear regression calibration. They compensated for channel differences with feature warping and probabilistic feature mapping techniques, with a resulting C_{llr} of 0.401 for the MFCC system. For the acoustic phonetic approach, they used the DCT of the trajectory of F2 of /o/ extracted from the word "no" and mean F0 as features, and MVKD were used for LR calculations. They reported a C_{llr} of 0.834.

Enzinger and Morrison (2017)'s approach for handling channel differences is far more technical than ours in which we simply excluded the frequency ranges unlikely to contain speaker information. Given our data was far more controlled and therefore favourable to FVC experiments than Enzinger & Morrison's, we cannot draw a conclusion that our approach is similarly effective to theirs. However, comparison to their results seems to suggest our technique has some promise and is worth further exploration.

Using long term distribution of formants for F1-F4, Coy, Hughes, Harrison, and Gully (2021) compared the performance of various formant extraction techniques under various channel conditions. Their best performing system under a channel-mismatch condition involving mobile phone transmission produced a C_{llr} of about 0.85. Our study did not explore long-term sub-band PCD nor the F4 range. Considering the difficulty of accurately measuring F4, we speculate that sub-band PCD may extract these long-term features more effectively, and perhaps we could improve the performance by incorporating long-term features.

The techniques in the field of automatic speaker recognition are progressing rapidly, and some excellent results have been reported (for reviews of the development and state-of-art techniques in ASR, see Greenberg, Mason, Sadjadi, & Reynolds, 2020; Kabir, Mridha, Shin, Jahan, & Ohi, 2021; Morrison et al., 2020). In 2016, a special issue of Speech Communication called for papers that test and report the performance of different FVC systems, using the prescribed evaluation dataset, *forensic_eval_01* (Morrison & Enzinger, 2016). This dataset consists of testing and training data, which reflect realistic conditions of FVC: noisy, channel-mismatched, non-contemporaneous and spontaneous. This special issue received six submissions, testing ten different systems based on various techniques which had been developed over the previous two decades, from GMM-UBM, i-vector, Deep Neural Network (DNN) bottle neck, to x-vector (Morrison & Enzinger, 2019). The results presented clearly demonstrated that automatic speaker recognition works well even under forensically realistic conditions, and the newer techniques outperform old ones. The C_{llr} ranged from 0.593 for a GMM-UBM based system to 0.246 and 0.208 for x-vector based systems (Jessen, Bortlík, Schwarz, & Solewicz, 2019; Kelly et al., 2019).

Their experiments were conducted under far more realistic conditions than our study, and yet the latest techniques produced considerably lower C_{llr}s than those we observed in this study. However, we believe that it is worthwhile to further explore the use of band-limited cepstral features, such as the one proposed in this study. First, this approach still has scope for refinement which could lead to improved performance. Second, in some forensic cases, the available speech material is very limited in its phonological content as well as quantity. In such a situation, segmentally focused evaluation maybe more appropriate than the global approach used in automatic

speaker recognition. Finally, and perhaps most importantly, any FVC systems need to be validated and calibrated with a training dataset which appropriately reflects the conditions of the forensic speech samples, but defining the characteristics of the forensic samples and determining which of those should be reflected in the training data – and to what extent – are far from clear. We need more investigation into effects of factors such as linguistics variety, speaking styles and social settings, emotion, recording conditions and equipment, background noise, to name a few, in order to select an 'appropriate' training dataset effectively and objectively. Sub-band cepstral features can facilitate a large scale linguistically informed investigation on this issue, as they are automatically extractable and can flexibly focus on any frequency ranges.

7. Conclusion and future tasks

Sub-band PCD opens a way to combine linguistic phonetic knowledge with an automatically extractable feature, LPCC, in a transparent way. This study examined its potential as an alternative to formants in acoustic phonetics research, and the results were promising. Our comparative FVC experiments demonstrated that sub-band PCD captures the spectral information within desired frequency ranges more completely and robustly than formants do.

We propose sub-band PCD analysis as a useful alternative to formants. It enables fast processing of acoustic data, which enhances our capacity to work with larger datasets, and assists statistical validation of phonetic observations. We also conjecture that the sub-band PCD can mitigate channel effects with a simpler measure: simply exclude the frequency ranges which are unlikely to be useful.

However, for sub-band PCD to be used in FVC in future, a few further investigations are required. Firstly, we need to test this approach on naturally spoken data, which is known to have a reduced vowel space (e.g. on Japanese vowels, see Okuda, 2005). Secondly it would be worth examining the effect of averaging LPCC across the 10–11 phonological contexts. We speculate that the performance may improve if we produce PCD for each phonological environment and fuse them.

Also, while our C_{llr} results appear promising, we converted sub-band PCD into LLR without incorporating typicality information, as noted earlier. Thus the LLRs produced here are likely not to be sufficiently accurate (Morrison & Enzinger, 2018). Solutions to this methodological issue are currently under development by the third author of this study.

CRediT authorship contribution statement

Yuko Kinoshita: Conceptualization, Methodology, Writing – original draft, Visualization, Software, Investigation, Formal analysis, Project administration. **Takashi Osanai:** Data curation, Methodology, Software, Investigation. **Frantz Clermont:** Investigation, Writing, Methodology.

Acknowledgement

We would like to thank our anonymous reviewers for their insightful feedback. We would also like to thank Dr Debbie Loakes for her helpful comments on our draft. However, we are solely responsible for any issues and errors found in this work.

Appendix A

The PCD expression outlined in Eq. (3) is recalled here in Eq. (A.1). A complete mathematical description of its terms is provided below.

$$D^2_{\mathrm{NDPS}}(\theta_1,\theta_2) = [\mathbf{K}\cdot(\mathbf{C}-\mathbf{C}')]^T \cdot \mathbf{W}(\theta_1,\theta_2)\cdot[\mathbf{K}\cdot(\mathbf{C}-\mathbf{C}') \tag{A.1}$$

C and **C′** are the pair of standard LPCCs to be compared. These are column vectors $(M\mathrm{x}\,1)$, where M is the LP-analysis order. The superscript T is the transpose operator.

The term **K** is a diagonal matrix $(M\mathrm{x}\,M)$ where the cepstral-coefficient indices $k = 1, 2, \cdots, M$ are assigned to its non-zero elements, as shown in Eq. (A.2):

$$\mathbf{K} = \begin{bmatrix} 1 & 0 & \cdots & \cdots & 0 \\ 0 & 2 & \ddots & \vdots & \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 & \\ 0 & \cdots & \cdots & 0 & M \end{bmatrix} \tag{A.2}$$

The term **W** is a matrix $(M\mathrm{x}\,M)$ with diagonal elements α_{kk} and upper-triangular elements β_{kl}, as shown in Eq. (A.3):

$$\mathbf{W} = \frac{1}{(\theta_2-\theta_1)} \begin{bmatrix} \alpha_{11} & \beta_{12} & \beta_{13} & \cdots & \beta_{1,M} \\ 0 & \alpha_{22} & \beta_{23} & \cdots & \beta_{2,M} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \beta_{(M-1),M} & \\ 0 & \cdots & \cdots & 0 & \alpha_{M,M} \end{bmatrix} \tag{A.3}$$

where:

$$\alpha_{kk}(\theta_1,\theta_2) = \frac{(\theta_2-\theta_1)}{2} + \frac{\sin(2k\theta_2)-\sin(2k\theta_1)}{4k}$$

$$\beta_{kl}(\theta_1,\theta_2) = \frac{\sin(k-l)\theta_2-\sin(k-l)\theta_1}{(k-l)} + \frac{\sin(k+l)\theta_2-\sin(k+l)\theta_1}{(k+l)}$$

The sub-band parameters $\theta_1$andθ_2 depend on the sampling frequency f_s (Hertz) as follows:

$$\theta_1 = \frac{2\pi f_1}{f_s}$$

$(\text{radians}),\text{where } f_1 \equiv \text{lower bound of sub-band } (\text{Hertz})$

$$\theta_2 = \frac{2\pi f_2}{f_s}(\text{radians}),\ \text{where } f_2 \equiv \text{upper bound of sub-band } (\text{Hertz})$$

Appendix B

The novelty of the approach to formant extraction presented here lies in our attempt to replicate — as a structured automatic process — how experienced phoneticians would extract formants manually. First of all, we extracted spectral peaks for each target vowel using the *formant listing* function of *Praat* (Boersma & Weenink, 2017). This was followed by a sequence of steps to counter the unreliable nature of automatic formant extraction. These steps can be grouped in two phases: identifying the typical frequency values of the spectral peaks which constitute candidates for formants and determining which of these candidates are to be assigned to formants for each

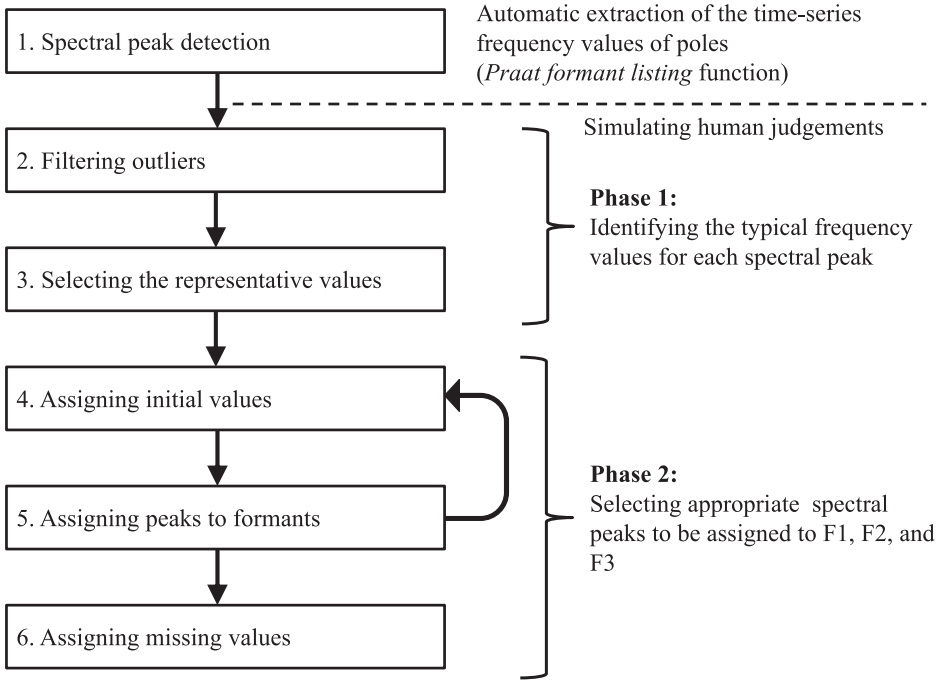

Fig. 11. Overall flow of formant extraction process.

token. The overall flow of the process is summarised in Fig. 11, and we describe each process in detail below.

Peak extraction

We extracted five spectral peaks for each target vowel using the *formant listing* function of *Praat*. This function detects and lists a user-defined number of spectral peaks from the lowest to the highest, over the selected duration (the analysis range as 0–4 kHz, sampled every 0.005 seconds). (We note that recent version of Praat now has the function called FormantPath, which appear to work much better in tracking formants.) We chose to extract five peaks even though our targets were F1 to F3 for two reasons. Firstly, it is much easier to remove extra peaks than to systematically find the peaks which are missed. Secondly, our trial formant extraction revealed that *Praat* differentiated F3 and F4 better when five poles were extracted.

At this point, we call the identified poles peaks 1–5, not formants, as we are not yet certain that those poles are indeed formants.

Phase 1: Identifying the typical frequency values of the spectral peaks for each token uttered by 306 individuals

The outputs of the *Praat* formant listing can fluctuate wildly even within the same token, because *Praat* often misidentifies formants. To exclude these as errors, we tried to simulate the process by which experienced phoneticians select formants manually. Human formant measurers would examine the suitability of the *formant listing* outputs using two criteria: the likely frequency range for a given vowel-formant combination; and the continuity of the formant track. If any automatically-identified value did not fit these criteria, it would be excluded from the data. Since the focus of this study is on monophthongs, we postulate that selecting the most typical values among the listed 'formant' outputs by *Praat* for each peak would have a similar outcome to this human judgement process. As such, we created following steps.

Filtering outliers

We first produced a histogram with bin 100 Hz for peak 1 measurements. We then identified the most populated bin and its immediately adjacent bins, one on each side. The range represented by these selected bins was considered as the likely F1 range. We selected all the peak 1 measurements which fell in this range as the updated peak 1. This process removed the values that are clearly outliers, presumably caused by the misidentification of formants by *Praat*.

We then also looked at peak 2 measurements. If the particular sampling points did not have a peak 1 measurement that fit in the likely F1 range, but its peak 2 measurement fitted in this range, we moved this to the updated peak 1 data.

We applied very similar processes to the measurements from peaks 2 to 4. In identifying the updated peak 2, we gathered peak 2 measurements that had higher frequencies than the F1 range. We produced a histogram from them, and identified the potential F2 range in the same way as we did with the F1 range. All measurements within the potential F2 range were identified as updated peak 2, as were the peak 3 measurements that also fell in the F2 range. For F3 and F4, the process was exactly the same, except that the potential frequency range was set wider – two adjacent bins on both sides of the most populated bin, to reflect the naturally greater variations in higher formants. (Note that the measurements of peak 5 were used only to supplement peak 4 measurements.) This

process gave us time series lists of the measurements of the updated peaks 1–4 which are within the likely formant ranges.

Selecting the representative values

Next, we fitted a kernel density to these lists of updated peak measurements, using the *density* function of the statistical package *R*. The density function of *R* automatically assigns x-coordinates by dividing the distance between the minimum and the maximum values into equal intervals. The default of 1024 coordinates were used in this study. We took the coordinates of the maximum point and immediately adjacent points (i.e. a total of three sets of coordinates) and fitted a quadratic function. We then recorded the maximum value of this quadratic function as the most representative frequency value of the formant of a given token. This summarised the four sets of time-series measurements of peaks into four representative frequency values. Thus, at this point, we came to have four peak frequency values for each token of the five vowels spoken by 306 speakers.

Phase 2: Assigning peaks to formants

Setting initial values

In manually selecting the most likely peaks to be F1, F2 and F3, phoneticians use their knowledge of the likely formant range for the particular formant-vowel combination. To simulate this process, we first pooled all 306 speakers' peak values obtained in the previous step. We then produced a histogram with 100 Hz bins for each peak and vowel combination (peaks 1 to 3 for the five vowels, 15 in total). For each peak, we took the most populated bins to include the typical values of the population, so the starting values of these bins were set as the initial values for each vowel-formant combination (i.e. /e/ F1, /i/ F2, etc).

These initial values were set using the Ch1 (microphone) recordings, as they were likely to be closer to the true formant values: they are less contaminated by external factors, such as mobile phone codecs, codec switching, noise etc.

Assigning peaks to formants

In formant measurements, phoneticians would choose the spectral peak closest to the expected frequency range of the target formant, but in doing so, they would also examine neighbouring peaks and make a holistic judgement as to which choice is best.

To simulate this process, we first identified the four possible patterns of peak-to-formant assignment, as shown in Table 1 in the main text of this article. For each token, we selected the pattern in which the three peaks together produced the smallest distance from the population initial values.

Once we applied this process to all tokens, we recalculated the population means and updated the initial values with these population means. We then re-examined which of the four patterns are the closest to this updated population mean, until the population mean formant values no longer change. We accepted these final values as formants.

Assigning missing values

The obtained formant values may still include some spurious values even after all these processes. To exclude these, we calculated population means and standard deviations for F1, F2 and F3 for each vowel and assigned NA to any values which were outside of ±4 standard deviations of the population mean formants: that is, any values outside of 99.99% of the distribution.

References

Aitken, C., & Lucy, D. (2004). Evaluation of trace evidence in the form of multivariate data. *Applied Statistics, 53*(4), 109–122.

Alzqhoul, E. A., Nair, B. B., & Guillemin, B. J. (2014). Comparison between speech parameters for forensic voice comparison using mobile phone speech. *Paper presented at the The 15th Australasian International Conference on Speech Science & Technology, Christchurch*.

Alzqhoul, E. A., Nair, B. B., & Guillemin, B. J. (2015). Impact of dynamic rate coding aspects of mobile phone networks on forensic voice comparison. *Science & Justice, 55*(5), 363–374.

Boersma, P., & Weenink, D. (2017). Praat: doing phonetics by computer. Retrieved from http://www.praat.org/.

Brümmer, N., & Du Preez, J. (2006). Application independent evaluation of speaker detection. *Computer Speech and Language, 20*(2–3), 230–275.

Byrne, C., & Foulkes, P. (2004). The 'mobile phone effect' on vowel formants. *International Journal of Speech Language and the Law, 11*(1), 83–102.

Cao, H., & Dellwo, V. (2019). The role of the first five formants in three vowels of mandarin for forensic voice analysis.

Cavalcanti, J. C., Eriksson, A., & Barbosa, P. A. (2021). Acoustic analysis of vowel formant frequencies in genetically-related and non-genetically related speakers with implications for forensic speaker comparison. *PLoS One, 16*(2), e0246645.

Clermont, F., & Kinoshita, Y. (2019). Analysis of speaker and co-articulation effects based on sub-band cepstral variances in the Japanese vowels of 300 male speakers. *Paper presented at the 14th Biennial Conference of the International Association of Forensic Linguists Melbourne*.

Clermont, F., Kinoshita, Y., & Osanai, T. (2016). Sub-band cepstral variability within and between speakers under microphone and mobile conditions: A preliminary investigation. *Paper presented at the The 16th Australasian International Conference on Speech Science & Technology, Sydney*.

Clermont, F., & Mokhtari, P. (1994). Frequency-band specification in cepstral distance computation. *Paper presented at the The 5th Australian International Conference on Speech Science & Technology*.

Clermont, F., & Mokhtari, P. (1998). Acoustic-articulatory evaluation of the upper vowel-formant region and its presumed speaker-specific potency. *Paper presented at the Fifth International Conference on Spoken Language Processing*.

Coy, T., Hughes, V., Harrison, P., & Gully, A. (2021). A comparison of the accuracy of Dissen and Keshet's (2016) DeepFormants and traditional LPC methods for semi-automatic speaker recognition. *Paper presented at the INTERSPEECH 2021, Brno, Czechia*.

Dissen, Y., Goldberger, J., & Keshet, J. (2019). Formant estimation and tracking: A deep learning approach. *The Journal of the Acoustical Society of America, 145*(2), 642–653.

Duckworth, M., McDougall, K., de Jong, G., & Shockey, L. (2011). Improving the consistency of formant measurement. *International Journal of Speech, Language & the Law, 18*(1), 35–51. https://doi.org/10.1558/ijsll.v18i1.35.

Enzinger, E., & Morrison, G. S. (2017). Empirical test of the performance of an acoustic-phonetic approach to forensic voice comparison under conditions similar to those of a real case. *Forensic Science International, 277*, 30–40. https://doi.org/10.1016/j.forsciint.2017.05.007.

Fant, G. (1971). *Acoustic theory of speech production: With calculations based on X-ray studies of Russian articulations*. Berlin/Boston, Germany: De Gruyter Inc..

Fletcher, J., & Butcher, A. (2003). *Local and Global Influences on Vowel Formats in Three Australian Languages*. Causal Productions.

Furui, S., & Akagi, M. (1985). Perception of voice individuality and physical correlates. 音響学会聴覚研資, H 85-18.

Garcia, A. A., & Mammone, R. J. (1999). *Channel-robust speaker identification using modified-mean cepstral mean normalization with frequency warping.* Paper presented at the Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on.

Garton, N., Ommen, D., Niemi, J., & Carriquiry, A. (2020). Score-based likelihood ratios to evaluate forensic pattern evidence. *arXiv preprint arXiv:2002.09470*.

Goldstein, U. G. (1976). Speaker-identifying features based on formant tracks. *The Journal of the Acoustical Society of America, 59*(1), 176–182.

Greenberg, C. S., Mason, L. P., Sadjadi, S. O., & Reynolds, D. A. (2020). Two decades of speaker recognition evaluation at the national institute of standards and technology. *Computer Speech & Language, 60*. https://doi.org/10.1016/j.csl.2019.101032 101032.

Guillemin, B. J., & Watson, C. (2008). Impact of the GSM mobile phone network on the speech signal: Some preliminary findings. *International Journal of Speech, Language & the Law, 15*(2).

Hanson, B., & Wakita, H. (1987). Spectral slope distance measures with linear prediction analysis for word recognition in noise. *IEEE Transactions on Acoustics, Speech, and Signal Processing, 35*(7), 968–973.

Harrison, P. (2013). *Making accurate formant measurements: An empirical investigation of the influence of the measurement tool, analysis settings and speaker on formant measurements*. University of York.

Hughes, V., & Foulkes, P. (2014). Variability in analyst decisions during the computation of numerical likelihood ratios. *International Journal of Speech Language and the Law, 21*(2), 279–315.

Hughes, V., Harrison, P. T., Foulkes, P., French, J. P., Kavanagh, C., & San Segundo Fernandez, E. (2018). The individual and the system: Assessing the stability of the output of a semi-automatic forensic voice comparison system. *Paper presented at the Proceedings of Interspeech 2018*.

Hunt, M. J., & Lefèbvre, C. (1989). *Distance measures for speech recognition*. Retrieved from.

Ingram, J. C. L., Prandolini, R., & Ong, S. (1996). Formant trajectories as indices of phonetic variation for speaker identification. *Forensic Linguistics, 3*(1), 129–145.

Jessen, M. (2008). Forensic phonetics. *Language and Linguistics Compass, 2*(4), 671–711.

Jessen, M., Bortlík, J., Schwarz, P., & Solewicz, Y. A. (2019). Evaluation of Phonexia automatic speaker recognition software under conditions reflecting those of a real forensic voice comparison case (forensic_eval_01). *Speech Communication, 111*, 22–28. https://doi.org/10.1016/j.specom.2019.05.002.

Juang, B.-H., Rabiner, L., & Wilpon, J. (1987). On the use of bandpass liftering in speech recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing, 35*(7), 947–954.

Kabir, M. M., Mridha, M., Shin, J., Jahan, I., & Ohi, A. Q. (2021). *A survey of speaker recognition: Fundamental theories, recognition methods and opportunities*. IEEE Access.

Kelly, F., Fröhlich, A., Dellwo, V., Forth, O., Kent, S., & Alexander, A. (2019). Evaluation of VOCALISE under conditions reflecting those of a real forensic voice comparison case (forensic_eval_01). *Speech Communication, 112*, 30–36. https://doi.org/10.1016/j.specom.2019.06.005.

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

Kinoshita, Y. (2001). *Testing realistic forensic speaker identification in Japanese: A likelihood ratio based approach using formants* (PhD). Canberra: The Australian National University.

Kinoshita, Y., Osanai, T., & Clermont, F. (2018). *FVC using sub-band cepstral distances as features: A first attempt with vowels from 306 Japanese speakers under channel mismatch conditions. Paper presented at the 17th Speech Science and Technology Conference (SST2018), Sydney*.

Künzel, H. J. (2001). Beware of the 'telephone effect': The influence of telephone transmission on the measurement of formant frequencies. *Forensic Linguistics, 8*(1), 80–99.

Lindblom, B. (1963). Spectrographic study of vowel reduction. *The Journal of the Acoustical Society of America, 35*(11), 1773–1781.

Lu, X., & Dang, J. (2008). An investigation of dependencies between frequency components and speaker characteristics for text-independent speaker identification. *Speech Communication, 50*(4), 312–322.

Lucy, D., Curran, J., & Martyna, A. (2020). comparison: Multivariate Likelihood Ratio Calculation and Evaluation. 1.0-5. Retrieved from https://CRAN.R-project.org/package=comparison.

Makinae, H., Osanai, T., Kamada, T., & Tanimoto, M. (2007). Construction and preliminary analysis of a large-scale bone-conducted speech database Retrieved from. *IEICE Technical Report, Speech, 107*(165), 97–102 http://ci.nii.ac.jp/naid/40015600747/.

Markel, J. D., & Gray, A. J. (1976). *Linear prediction of speech* (Vol. 12). Springer Science & Business Media.

McLaughlin, J., Reynolds, D. A., & Gleason, T. P. (1999). A study of computation speed-UPS of the GMM-UBM speaker recognition system. *Paper presented at the EUROSPEECH*.

Mermelstein, P. (1978). Difference limens for formant frequencies of steady-state and consonant-bound vowels. *The Journal of the Acoustical Society of America, 63*(2), 572–580.

Meuwly, D., & Drygajlo, A. (2001). Forensic speaker recognition based on a Bayesian Framework. *Paper presented at the A Speaker Odyssey 2001, Crete, Greece*.

Mokhtari, P., & Clermont, F. (1994). Contributions of selected spectral regions to vowel classification accuracy. *Paper presented at the Third International Conference on Spoken Language Processing*.

Morrison, G. S. (2009). Forensic voice comparison and the paradigm shift. *Science and Justice, 49*(4), 298–308.

Morrison, G. S. (2011). A comparison of procedures for the calculation of forensic likelihood ratios from acoustic–phonetic data: Multivariate kernel density (MVKD) versus Gaussian mixture model–universal background model (GMM–UBM). *Speech Communication, 53*(2), 242–256. https://doi.org/10.1016/j.specom.2010.09.005.

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://doi.org/10.1080/00450618.2012.733025.

Morrison, G. S. (2018). The impact in forensic voice comparison of lack of calibration and of mismatched conditions between the known-speaker recording and the relevant-population sample recordings. *Forensic Science International, 283*, e1–e7. https://doi.org/10.1016/j.forsciint.2017.12.024.

Morrison, G. S., & Enzinger, E. (2016). Multi-laboratory evaluation of forensic voice comparison systems under conditions reflecting those of a real forensic case (forensic_eval_01) – Introduction. *Speech Communication, 85*, 119–126. https://doi.org/10.1016/j.specom.2016.07.006.

Morrison, G. S., & Enzinger, E. (2018). Score based procedures for the calculation of forensic likelihood ratios–Scores should take account of both similarity and typicality. *Science & Justice, 58*(1), 47–58.

Morrison, G. S., & Enzinger, E. (2019). Multi-laboratory evaluation of forensic voice comparison systems under conditions reflecting those of a real forensic case (forensic_eval_01) – Conclusion. *Speech Communication, 112*, 37–39. https://doi.org/10.1016/j.specom.2019.06.007.

Morrison, G. S., Enzinger, E., Hughes, V., Jessen, M., Meuwly, D., Neumann, C., ... Anonymous, B. (2021). Consensus on validation of forensic voice comparison. *Science & Justice, 61*(3), 299–309. https://doi.org/10.1016/j.scijus.2021.02.002.

Morrison, G. S., Enzinger, E., Ramos, D., González-Rodríguez, J., & Lozano-Díez, A. (2020). Statistical models in forensic voice comparison. In D. L. Banks, K. Kafadar, D. H. Kaye, & M. Tackett (Eds.), *Handbook of forensic statistics*. Milton, United Kingdom: CRC Press LLC.

Morrison, G. S., & Kinoshita, Y. (2008). Automatic-Type Calibration of Traditionally Derived Likelihood Ratios: Forensic Analysis of Australian English /o/ Formant Trajectories. *Paper presented at the Interspeech 2008, Brisbane*.

Morrison, G. S., Zhang, C., & Rose, P. J. (2011). An empirical estimate of the precision of likelihood ratios from a forensic-voice-comparison system. *Forensic Science International, 208*(1–3), 59–65. https://doi.org/10.1016/j.forsciint.2010.11.001.

Nair, B., Alzqhoul, E., & Guillemin, B. J. (2014). Determination of likelihood ratios for forensic voice comparison using Principal Component Analysis. *International Journal of Speech Language and the Law, 21*(1), 83–112.

Nakagawa, T. (1982). Tonal difference limens for second formant frequencies of synthesized Japanese vowels. *Annual Bulletin, Research Institute of Logopedics and Phoniatrics, University of Tokyo, 16*, 81–88.

Nolan, F. (1983). *The phonetic bases of speaker recognition*. Cambridge: Cambridge University Press.

Nolan, F., & Oh, T. (1996). Identical twins, different voices. *Forensic Linguistics, 3*(1), 39–49.

Okuda, K. (2005). 発話スタイルの変動に頑健な音響モデル構築法に関する研究. (Doctoral dissertation). 大阪市立大学,

Osanai, T., Kinoshita, Y., & Clermont, F. (2018). Exploring sub-band cepstral distances for more robust speaker classification. *Paper presented at the 17th Speech Science and Technology Conference (SST2018), Sydney*.

Pols, L. C., Tromp, H. R., & Plomp, R. (1973). Frequency analysis of Dutch vowels from 50 male speakers. *The Journal of the Acoustical Society of America, 53*(4), 1093–1101.

Ramos, D., Haraksim, R., & Meuwly, D. (2017). Likelihood ratio data to report the validation of a forensic fingerprint evaluation method. *Data in Brief, 10*, 75–92.

Reynolds, D. A. (2003). Channel robust speaker verification via feature mapping. In (pp. II-53-56): IEEE.

Reynolds, D. A., & Rose, R. C. (1995). Robust text-independent speaker identification using Gaussian mixture speaker models Retrieved from. *IEEE Transactions on Speech and Audio Processing, 3*(1), 72–83 https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=365379.

Robertson, B., Vignaux, G. A., & Berger, C. E. (2016). *Interpreting evidence: Evaluating forensic science in the courtroom* (2nd ed.). John Wiley & Sons.

Rose, P. J. (1998). Difference and discriminability in the acoustic characteristics of words in voices of similar-sounding speakers - a forensic phonetic investigation. *Paper presented at the ICSLP '98, Sydney*.

Rose, P. J. (2002). *Forensic speaker identification*. Taylor & Francis.

Rose, P. J. (2007). Forensic speaker discrimination with Australian English vowel acoustics. *ICPhS XVI Saarbrucken, 6*(10).

Rose, P. J. (2017). Likelihood ratio-based forensic voice comparison with higher level features: Research and reality. *Computer Speech & Language, 45*, 475–502.

Rose, P. J., Lucy, D., & Osanai, T. (2004). Linguistic-acoustic forensic speaker identification with likelihood ratios from a multivariate hierarchical effects model: A "non-idiot's bayes" approach. *Paper presented at the the 10th Australian International Conference on Speech Science & Technology, Sydney*.

Rose, P. J., Osanai, T., & Kinoshita, Y. (2002). Strength of forensic speaker identification evidence: Multispeaker formant and cepstrum-based segmental discrimination with a Bayesian likelihood ratio as threshold. *Paper presented at the The 9th Australian International Conference on Speech Science & Technology Melbourne*.

Rose, P. J., & Winter, E. (2010). Traditional forensic voice comparison with female formants: Gaussian mixture model and multivariate likelihood ratio analyses. *Paper presented at the SST2010, Melbourne*.

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

Sambur, M. R. (1975). Selection of acoustic features for speaker identification. *IEEE Transactions on Acoustics, Speech, and Signal Processing, ASSP-23(2)*, 178–182.

Shikano, K., & Itakura, F. (1992). Spectrum distance measures for speech recognition. *Advances in Speech Signal Processing*, 419–452.

Solomonoff, A., Campbell, W. M., & Boardman, I. (2005). Advances in channel compensation for SVM speaker recognition. *Paper presented at the Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP'05). IEEE International Conference on*.

Stevens, K. N. (1971). Sources of inter- and intra-speaker variability in the acoustic properties of speech sounds. In *Proceedings of the Seventh International Cons. Phonetic Sciences* (pp. 206–232).

Stevens, K. N. (1998). *Acoustic phonetics*. Cambridge: MIT Press.
Stevens, K. N., & House, A. S. (1963). Perturbation of vowel articulations by consonantal context: An acoustical study. *Journal of Speech and Hearing Research, 6*(2), 111–128.
Tohkura, Y. (1987). A weighted cepstral distance measure for speech recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing, 35*(10), 1414–1422.
van Leeuwen, D. A., & Brümmer, N. (2007). An introduction to application – Independent evaluation of speaker recognition system. In C. Müller (Ed.). *Speaker classification* (Vol. 1, pp. 330–353). Berlin: Springer.
Vrieze, S. I. (2012). Model selection and psychological theory: A discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). *Psychological Methods, 17*(2), 228–243.
Westphal, M. (1997). The use of cepstral means in conversational speech recognition. *Paper presented at the EUROSPEECH*.
Yegnanarayana, B., & Reddy, D. (1979). A distance measure based on the derivative of linear prediction phase spectrum. *Paper presented at the ICASSP'79. IEEE International Conference on Acoustics, Speech, and Signal Processing*.
Zhang, C., Morrison, G. S., Enzinger, E., & Ochoa, F. (2013). Effects of telephone transmission on the performance of formant-trajectory-based forensic voice comparison – Female voices. *Speech Communication, 55*(6), 796–813. https://doi.org/10.1016/j.specom.2013.01.011.
Zhang, C., Morrison, G. S., Ochoa, F., & Enzinger, E. (2013). Reliability of human-supervised formant-trajectory measurement for forensic voice comparison. *The Journal of the Acoustical Society of America, 133*(1), EL54-EL60. https://doi.org/10.1121/1.4773223.