Article:
Hughes, Vincent, Cardoso, Amanda Beth, Foulkes, Paul et al. (2023) Speaker-specificity in speech production: the contribution of source and filter. Journal of Phonetics. 101224. ISSN: 0095-4470

https://doi.org/10.1016/j.wocn.2023.101224

Journal of Phonetics 97 (2023) 101224

Contents lists available at ScienceDirect

Journal of Phonetics

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

Research Article

Speaker-specificity in speech production: The contribution of source and filter

Vincent Hughes [a,*], Amanda Cardoso [b], Paul Foulkes [a], Peter French [a,c], Amelia Gully [a], Philip Harrison [a]

[a] Department of Language and Linguistic Science, University of York, UK
[b] Department of Linguistics, The University of British Columbia, Canada
[c] J P French Associates, York, UK

ARTICLE INFO

Article history:
Received 3 February 2022
Received in revised form 10 January 2023
Accepted 23 January 2023

Keywords:
Speaker-specificity
Speaker recognition
Forensic speech science
Hesitation markers
Source-filter theory

ABSTRACT

This study examines the extent to which speaker-specific information is encoded in different features of vocal output and the relationships between those features. A range of acoustic features, grouped as source (laryngeal voice quality measures and fundamental frequency) and filter features (formants and Mel-frequency cepstral coefficients; MFCCs), were extracted from the vocalic portion of the hesitation marker *um* for 90 male speakers of Standard Southern British English. Little overall correlation between the sets of features was observed, suggesting no strong interdependence between source and filter in our data. Although filter features were consistently better at discriminating between same- and different-speaker pairs compared with source features, combining source and filter has the potential of producing the lowest error rates and the strongest speaker discrimination scores. Taken together, results show that source and filter provide complementary speaker-specific information. However, the extent of the improvements in speaker discrimination performance when combining source and filter varied across speakers. We explore potential explanations for this finding and discuss the implications for source-filter theory, and for applied fields such as speaker recognition and forensic speech science.

1. Introduction

Speech is a complex acoustic signal. It conveys not only semantic and pragmatic meaning but is also rich with indexical information about a speaker's regional and social background, their stance or attitudes towards a given topic or interlocutor, as well as short term factors such as intoxication and ill health. The voice also carries considerable speaker-specific information, allowing listeners to recognise individuals reliably, in certain contexts, from the acoustic signal alone (though not without the possibility of error, see e.g. Ladefoged & Ladefoged, 1980). Within the field of forensic speech science, many studies have tested the speaker discriminatory power of specific phonetic parameters and units, such as salient vowels and fundamental frequency. Work in the forensic domain has also compared the performance of phonetic and automatic (i.e. computational) methods of speech analysis, with a general focus on the supralaryngeal vocal tract (i.e. the acoustic output of the *filter* rather than the *source*) as the assumed primary carrier of speaker-specific information. However, the aims of such work are typically applied in nature, focusing on speaker discrimination rates to inform choices of features or methods for analysis in forensic casework. Little work has attempted to examine the relationships *between* features to build a comprehensive model of where within the speech signal speaker-specific information is located, and how and why individual voices differ from one another. Understanding the bases of speaker-specific behaviour in speech production therefore remains a key aim within forensic speech science, but with important implications more generally for our understanding of the limits of phonetic variation within- and between-speakers and for our theoretical models of speech production.

1.1. Towards a model of speaker-specificity in speech

An individual's voice is determined by a combination of biological factors relating to the size and shape of the vocal apparatus and behavioural factors relating to the articulatory implementation of speech production. Both biological and

* Corresponding author.
E-mail address: vincent.hughes@york.ac.uk (V. Hughes).

https://doi.org/10.1016/j.wocn.2023.101224
0095-4470/

behavioural factors can be sources of between-speaker variation. Biological factors are, by definition, speaker-specific, and include the size and mass of the vocal folds, and the length and overall shape of the vocal tract, along with other potential muscular or vocal apparatus differences. Garvin and Ladefoged (1963) further separate behavioural factors, distinguishing between group and individual factors. Group-level factors relate to the speech community in which a speaker was raised and to which a speaker belongs (for a critical discussion on the notion of speech community see Britain, 2013). Speakers can also develop patterns of speech production which are not determined by the speech community, but could be acquired through processes of individual learning, such as diphthong dynamics (McDougall, 2004).

However, as highlighted by Nolan (1983), the biological-behavioural and group-individual dichotomies oversimplify the complexity of speech production, since they imply that specific speech patterns can be neatly attributed to one or other set of factors. In reality, speaker-specific behaviour derives from an interaction between physiological and learned factors, as well as short-term factors relating to the context in which the discourse is taking place and occasion-to-occasion variability (Foulkes, Scobbie, & Watt, 2010: 706). In practice it is therefore problematic, perhaps impossible, to fully disentangle these various factors from one another. In addition, Nolan (1983) also provides a hierarchical model of the stages and components required to understand what makes it possible to distinguish between voices and speakers. Nolan's model draws links between communicative intent, divided into segmental and suprasegmental strands, leading to phonetic representation and implementation. Speaker-specific variation can be found at any level, although it is predicted to occur more at some levels than others. In particular, Nolan identifies *long-term quality* as being a key source of speaker-specific information. *Long-term quality* here refers to supralaryngeal vocal setting together with laryngeal voice quality, a distinction proposed by Laver (1980). Both vocal setting and voice quality draw on particular segmental and suprasegmental properties and are cumulative over a stretch of speech. Nolan (1983) work remains the only well-developed theoretical model of the phonetic bases of speaker-specificity in speech. However, it lacks large-scale empirical data incorporating methods and features from contemporary phonetic and forensic research in order to assess the relative contribution of the different levels and strands to speaker-specificity.

1.2. Speaker-specific features of speech production

It is within the fields of forensic speech science and automatic speaker recognition (ASR) that the greatest explicit attention has been given to individual speaker variation. This work has typically focused on identifying features of the voice that are responsible for speaker discrimination. In the main, such work has focused on *filter* output. Most ASR systems use Mel-frequency cepstral coefficients (MFCCs; see Davis & Mermelstein, 1980) as input features, potentially in combination with other measures. MFCCs are a representation of the rate of change of a signal's power spectrum across frequency. The number of MFCCs extracted by ASR systems is typically governed by overall speaker discriminatory performance (i.e. error rates), rather than any principled decisions about what phonetic information developers do or do not want systems to capture. However, claims have been made about MFCCs decoupling the filter and the source (i.e. removing source information) by smoothing over faster local changes in the spectrum caused by harmonics or noise in the signal (Jurafsky & Martin, 2009). Since the global shape of the power spectrum is governed by the shape of the supralaryngeal vocal tract, lower-order MFCCs encode information about the filter. The extent of the decoupling between source and filter is therefore dependent on the number of cepstral coefficients extracted; the fewer the cepstral coefficients, the smoother the spectral representation and so the less information about the source signal is captured.

A considerable body of research has been conducted to test the speaker discriminatory power of other features derived from supralaryngeal properties, such as vowel formants. Studies have analysed monophthongs (e.g. Rose 2007, 2010), diphthongs (e.g. McDougall 2004, 2006, Morrison, 2009), and triphthongs (e.g. Zhang, Morrison, & Thiruvaran, 2011) in a range of languages. Studies have also compared different parameterisations of vowel formants, from single measurements at the temporal midpoint to more complex representations of the entire trajectory across the duration of the vowel. Higher formants, particularly F3, have often been shown to discriminate more successfully between speakers than F1 and F2 (Hughes, 2014, McDougall, 2004). F1 and F2 furnish less between-speaker variation due to their key phonological role in contrasting between phonemes. Studies have also shown promising performance when pooling and modelling formants from across all vowels within a recording, and then analysing the data in a similar way to an ASR system. This approach is sometimes referred to as 'semi-automatic' speaker recognition (Nolan & Grigoras, 2005) and is one way of acoustically measuring supralaryngeal vocal setting.

By comparison, relatively little attention has been paid to the speaker discriminatory power of source features. The research that has been done has generally focused on f0 (e.g. Hudson, de Jong, McDougall, & Nolan, 2007, Jessen, Köster, & Gfroerer, 2005, Kinoshita, Ishihara, & Rose, 2009, Skarnitzl & Vankova, 2017). As with vowel research, attention has also been given to investigating the most effective ways of parameterising f0 to maximise speaker discrimination, such as modelling f0 trajectories (Rose, 2013). In general, f0 by itself has been shown to be relatively poor at characterising speakers, in large part because it displays relatively high within-speaker variability. For example, f0 is affected substantially by health, emotion and intoxication, as well as being highly susceptible to the Lombard effect (Braun, 1995). By contrast, little work in forensic speech science has considered the speaker discriminatory potential of laryngeal voice quality features (as an exception see Jessen, 1997), despite experts reporting the use of auditory-perceptual judgments of voice quality extensively in forensic casework (Gold and French 2011, 2019). Nolan (2005) is one of the few systematic studies of the forensic value of voice quality, reaching a somewhat pessimistic conclusion that formal analysis is likely to be limited due to the adverse recording conditions and variation in speech style typically encountered in casework.

There has been a general trend in recent years towards integrating the best elements of ASR systems and phonetic analysis. Central to this integration is the question of the independence of different voice features, and whether they capture complementary (i.e. independent) speaker-specific information. This information is critical in a forensic setting in order to avoid under- or overstating the strength of the evidence. González-Rodríguez, Gil, Pérez, and Franco-Pedroso (2014) and Hughes, Harrison, Foulkes, French, Kavanagh, and San Segundo (2017) used ASR systems based on MFCCs to perform speaker discrimination testing. Both studies found that system errors could be relatively easily resolved through auditory analysis by phoneticians, and that laryngeal voice quality was a key diagnostic. While these studies suggest that MFCCs fail to capture adequately the speaker-specific information provided by perceptually judged phonation (in line with the theoretical decoupling of source and filter in deriving MFCCs), the analyses conducted are relatively limited. This is because they focus on a small subset of the comparisons performed by the ASR system. More systematic studies have provided mixed results. Enzinger, Zhang, and Morrison (2012) analysed the performance of laryngeal voice quality measures extracted from the segment /n/ using the GLOTTEX software. Combining these measures (via a process called *fusion*) with the results of an MFCC-based ASR system revealed no marked improvement in performance, although this may be due to the segment used for analysis, small numbers of tokens for some speakers, or the features captured by the software. Conversely, improvements in the performance of MFCC-based ASR systems have been reported when fused with voice quality measures from the entire recording rather than a single segment (Farrús, Hernando, & Ejarque, 2007, Park, Sigouin, Kreiman, Keating, Guo, Yeung, Kuo, & Alwan, 2016, Hughes, Cardoso, Foulkes, French, Harrison, & Gully, 2019).

1.3. The present study

Despite the considerable body of research in forensic speech science examining the speaker discriminatory power of different features, there remain key unanswered questions about speaker-specificity. What precisely makes voices different from one another? How can we best extract and analyse speaker-specific information encoded in the speech signal? Only Nolan (1983) has attempted to bring together the different sources and loci of speaker-specific variation to build a model of the phonetic bases of speaker-specificity. In the present study, we contribute to such a model by examining the relative speaker discriminatory power of a range of acoustic measures of source and filter, using insights about their relative independence to inform how best to capture speaker-specific phonetic information within the speech signal. Specifically, we address the following research questions:

How do individual acoustic features of source and filter output covary?
To what extent do source and filter capture complementary speaker-specific information?
Does the fusion (i.e. combination) of source and filter features improve speaker discrimination performance over either one individually?

While our research questions are principally exploratory in nature (see Roettger, 2019 for more on why the distinction between confirmatory and exploratory research matters), our choice of features is motivated by a range of factors. Here, we analyse features that are widely used and well-understood from a phonetic perspective. Therefore, much is known about what information these features capture (e.g. MFCCs, in principle, capture formant information; see 2.2 for more), allowing us to separate source and filter contributions and capture these contributions in a variety of different ways.

The aim here is not to compete with speaker recognition studies or with current state-of-the-art ASR systems that can now produce essentially no errors with good quality materials (e.g. Snyder, Garcia-Romero, Sell, Povey, & Khudanpur, 2018). By design, we use methods that do make errors, allowing us to use the error profiles to better understand the intrinsic relationships between different features. Further, although we do have an interest in its implications for forensic speech science, this is not designed as a forensically realistic study. Indeed, we intentionally extract controlled segmental data from high quality recordings, as is widely used in voice quality research, to remove the confounding effects of channel and transmission quality, issues which are common in forensic casework. Our focus is on the relative, rather than absolute, speaker discrimination performance across the features and combinations of features that we analyse.

In the following section, we provide a general overview of the methods used within this study. We then describe three experiments designed to address our research questions and include specific methods within each of these sections. In section 6, we then discuss our results and consider the implications for forensic speech science and ASR, as well as for theoretical models of speech production (including source-filter theory, Fant, 1960).

2. Method

2.1. Data

The current study initially used 93 young (18–25 years old), male speakers of Standard Southern British English from the Dynamic Variability in Speech (DyViS; Nolan & McDougall, de Jong, G. & Hudson, T., 2009) corpus. We intentionally chose a controlled corpus to remove the effect of social and demographic sources of between-speaker variation. High quality studio recordings of spontaneous speech from DyViS Tasks 1 and 2 were used. Task 1 involves a mock police interview in which participants were required to lie about a crime. The Task 1 samples were between 11 and 26 minutes in total duration (mean duration = 17 minutes). Task 2 involves a landline telephone conversation with a mock accomplice, recalling details of the police interview in Task 1. The accomplice was a researcher. Task 2 samples were between 9 and 23 minutes in total duration (mean duration = 15 minutes). They were recorded simultaneously at the near- and far-ends of the telephone line, although only the near-end recordings were used for this study. Both tasks were recorded on the same day, with Task 2 always following Task 1.

Data were extracted from the vocalic portion of the hesitation marker *um*. There are several reasons for using *um*. Hesi-

tation markers are, in principle, good speaker discriminants: they are frequent, easily measurable, and resistant to disguise; and experiments show that they yield high between-speaker but low within-speaker variation. Hughes (2014) showed that hesitation markers generally outperform lexical vowels at speaker discrimination based on formant analysis. There is, therefore, more scope for between-speaker variation across the entire vowel space, compared with lexical vowels where the variation is much more constrained by phonology. Using only the DyViS Task 1 (simulated police interview) recordings, Hughes, Wood, and Foulkes (2016) investigated the speaker discriminatory performance of different acoustic measures extracted from *um*s and *uh*s (i.e. purely vocalic hesitation markers). They found that *um* consistently outperformed *uh* and that the best performance was achieved using representations of formant trajectories, as well as the durations of the vowel and nasal portions. Other studies have considered laryngeal measures extracted from hesitations (e.g. Tschäpe, Trouvain, Bauer, & Jessen, 2005).

Hesitations are also often longer than lexical vowels and occur in relative isolation, with silence preceding and/or following. This has the benefit of limiting within-speaker variation as a function of coarticulation. Perhaps more importantly, it also helps to ensure that the acoustic laryngeal voice quality measures extracted in this study are as reliable as possible, while still using real spontaneous speech. It is often preferable to use sustained vowels with relatively stable formant trajectories and pitch to obtain accurate and more easily comparable acoustic laryngeal voice quality measures. The vowel portions of hesitations fit these requirements much more closely than lexical vowels in spontaneous speech. A single segment is analysed here because many of the measures of relative harmonics are dependent on the formant structure of the vowel being measured, being derived from the harmonic closest to a given formant. Where different phonemes are analysed, it is possible to correct for this variability in generating the acoustic measures using formant measures (see Section 2.2); however, it is likely to increase within-speaker variability, especially given that the distribution of vowels for each speaker may be different in spontaneous speech.

The aim of the present study differs from that of Hughes et al. (2016), which focused exclusively on speaker discrimination for forensic purposes. In the present study, we use speaker discrimination testing as a means of better understanding speaker-specificity. Here, we also use a much-expanded data set, in terms of the number of speakers and tokens per speakers, and a much wider range of features (see below). Whereas Hughes et al. (2016) analysed the first 20 tokens per speaker, the present study analyses all the tokens produced within each recording for each speaker. The vowel portions of each token were manually segmented using Praat TextGrids (v. 6.0.49; Boersma & Weenink, 2019).

2.2. Features

A range of acoustic features were extracted from the vocalic portion of the hesitation markers. The features that we have chosen are used routinely in phonetics, forensic speech science, and ASR to capture discrete elements of speech production. In many cases, most obviously that of vowel formants, there are well known relationships between acoustic output and articulatory implementation. Many such features are also known to be good speaker discriminants. We group features broadly into two categories: source and filter. This reflects the distinction between supralaryngeal and laryngeal vocal output proposed by Laver (1980). This distinction allows us to make some generalisations about the loci of speaker-specific behaviour and to make claims about relative independence. We do, however, recognise that in some cases it may not be easy to categorise a feature as being exclusively a measure of source or of filter. In those cases, we discuss the complexity below. We use acoustic measurements as this allows us to conduct analysis of features using the same quantitative tests and to allow for empirical fusion results to assess speaker discriminatory power. In the case of laryngeal voice quality, the acoustic features we include are intended to capture the same perceptually judged information that phoneticians have relied on in studies at the intersection of phonetics and ASR (see Section 1.2). The specific relationship between auditory percepts and acoustic measures is beyond the scope of this paper. However, Cardoso, Foulkes, French, Harrison, Hughes, Kavanagh, and San Segundo (2018) and Klug, Kirchhübel, Foulkes, and French (2019) both examine these issues using the same corpus as the present study.

The following filter features were analysed:

Formants: F1, F2, and F3 values were extracted using the LPC-based *To Formant (burg)* function in Praat identifying between 5 and 6 formants between 0 and 5000 Hz, followed by the *Formant: Track* function with reference values set at 500 Hz for F1, 1500 Hz for F2, and 2500 Hz for F3. The frequency, bandwidth and transition costs were set to 1.0.

Mel-frequency cepstral coefficients (MFCCs): MFCCs are a representation of the entire spectrum and as such will encode formant information relating to spectral peaks. MFCCs analysis was conducted using the *rastamat* toolbox (Ellis, 2005) in MATLAB. Each vowel token was divided into a series of 20 ms frames, using a Hamming window, with 10 ms overlap between adjacent frames. From each frame a vector of 12 MFCCs was extracted within a 0 to 4000 Hz range, using a pre-emphasis coefficient of 0.97. As mentioned previously, the extent to which MFCCs are exclusively a measure of filter output is determined by the number of coefficients extracted. The 12 MFCCs used here is a relatively small number (modern ASR systems typically use more like 18 MFCCs) to minimise the amount of source information that is captured. Nonetheless, it is likely that our MFCC data captures some of the source information related to overall spectral shape. We consider this issue in interpreting the results.

The following source features were analysed:

Fundamental frequency: f0 extraction was conducted in Praat using the *To Pitch (ac)* function based on autocorrelation with the frequency range set to between 75 Hz and 200 Hz (an ample range for this set of voices; see Hudson et al. 2007).

Energy: Root Mean Square energy, capturing average overall amplitude, was extracted from 20 ms frames shifted by 10 ms across the duration of the vowel using VoiceSauce (Shue, 2010). Energy is categorised as a source feature here as overall it is determined by the rate of transglottal airflow. However, as pointed out by a reviewer, energy may also be affected by supralaryngeal gestures, such as jaw opening or lowering. We

consider this later in the results. While the microphones used for recording the DyViS tasks were in roughly the same place for each speaker, there is likely to have been some within- and between-speaker variability in the distance between the speaker and the microphone. This may compromise the comparability of data from different tokens, recordings and speakers.

Additive noise: Cepstral peak prominence (CPP) and harmonics to noise ratio (HNR) up to 500 Hz (HNR05), 1500 Hz (HNR15) and 2500 Hz (HNR25) were extracted from 20 ms frames shifted by 10 ms across the duration of the vowel using VoiceSauce. This produces a four-feature vector for each frame to represent additive noise. CPP is the normalised pitch peak amplitude within the real cepstral domain. Hillenbrand, Cleveland, and Erickson (1994) and Awan, Solomon, and Helou (2013) have argued that CPP is essentially a measure of the signal-to-noise ratio in the cepstrum. HNR was calculated using the algorithm in de Krom (1993). A number of studies (e.g. Gordon & Ladefoged, 2001, Garellek, 2019) have demonstrated that HNR captures breathy voice well, with HNR higher in breathy voice than modal.

Relative harmonics: A five-feature vector of measures of the distribution of energy across the spectrum were extracted from 20 ms frames shifted by 10 ms across the duration of the vowel using VoiceSauce. The measures were: H1-H2, H2-H4, H1-A1, H1-A2, and H1-A3. The codings refer to the amplitude of the harmonic (H) or formant (A), and the numbering of the relevant harmonic or formant. Thus H1-H2 captures the difference in amplitude between the first and second harmonic, H1-A1 is the difference in amplitude between the first harmonic and the amplitude of the first formant, and so on. These measures capture relative amplitudes across different frequency ranges and are commonly used in phonetic research on laryngeal voice quality. Measures were corrected using formants extracted from the Snack Sound Toolkit (Sjölander, 1997) tracking five formants between 0 and 5000 Hz (following the method in Iseli, Shue, & Alwan, 2006). This correction process, which is automatically computed within VoiceSauce, reduces the effect of the filter (formants) on the source features (harmonics). Relative harmonics are claimed to distinguish creaky and breathy voice, with low spectral tilt found in certain types of creaky voice (Keating, Garellek, & Kreiman, 2015) and high spectral tilt found in breathy voice (Garellek, 2019).

Jitter: This is a measure of cycle-to-cycle variability in duration, extracted using the *Get jitter (Local)* function in Praat. The local measure was used which is the average difference between periods divided by the average period, in this case within a 20 ms frame. The local measure was preferred to other measures, firstly, because it is the most intuitive to understand, and secondly, because it is commonly measured in studies of voice quality (e.g. Finger, Cielo, & Schwarz, 2009).

Shimmer: This is a measure of cycle-to-cycle variability in amplitude, extracted using the *Get shimmer (Local)* function in Praat. As with jitter, the local measure was used which is the average difference in amplitude between periods divided by the average amplitude.

2.3. Data tidying

The original DyViS corpus contains 100 speakers. In this study, seven speakers were removed prior to analysis due to small numbers of tokens in either sample. Acoustic data were then extracted from a total of 8,374 tokens from the remaining 93 speakers across the two tasks. A series of procedures was implemented to remove measurement errors from the data set (following Hughes et al., 2016, Foulkes, Docherty, Shattuck-Hafnagel, & Hughes, 2018). Firstly, acceptable ranges for formants and f0 were established. F1 values outside 250–900 Hz, F2 values outside 900–2000 Hz, F3 values outside 1900–3200 Hz and f0 values outside 75–200 Hz were removed. These boundaries were intentionally wide given the potential variability in *um* production across the entire vowel space. Secondly, the values for each feature were converted to z-scores, and values ±3.29 standard deviations from the mean were removed. Where measurement errors could not be resolved the entire token was removed from the analysis. Speakers with fewer than five tokens for either task were also removed. The resulting final dataset contained 6,758 tokens from 90 speakers. Overall, there were more tokens for Task 2 (phone call; median = 39 per speaker, min = 5, max = 99) than for Task 1 (interview; median = 30, min = 5, max = 71). There is some correlation between token numbers for the two tasks. However, inspection of the speaker discrimination results in section 3.3 reveals no obvious effect of sample size on performance.

3. Experiment 1: Correlations in the raw data

Research question 1 asked the extent to which source and filter features of vocal output covary (i.e. the extent to which they are independent). In order to address this research question, we performed a correlation analysis between all source and filter features, with midpoint data pooled across all speakers.

3.1. Methods

The filter features consisted of F1, F2, F3, and MFCCs. The source features were f0, additive noise (four features), relative harmonics (five features), energy, jitter, and shimmer. For the sets of additive noise and relative harmonics measures, separate principal components analyses (PCA) were performed to reduce the multivariate input to a set of univariate data (using principal component 1; PC1), accounting for the largest amount of variance in the raw data (see Lee, Keating, & Kreiman, 2019 for a similar application of PCA to acoustic voice quality measures). This allowed us to perform straightforward tests of correlation using univariate data that was as close to the interpretability of the raw data as possible. For the additive noise measures PC1 accounted for 85% of the variance in the data, while for the relative harmonics, PC1 accounted for 91% of the variance. PCA was not used to reduce the dimensionality of the MFCCs. This is because the MFCCs are already orthogonal, whereas PCA is intended for dimensionality reduction where dimensions are correlated.

Linear regression was then performed to assess the correlations between each of the source and filter features. With the exception of the MFCCs, correlations were performed between univariate data. Following Broad and Clermont (1989) (see also Darch & Milner, 2008), the 12-value MFCC vectors were used to predict the univariate source and filter features using a multiple linear regression model based on weighted sums

of the MFCCs, where weights were determined by the least squares method. All correlations are reported in terms of the R^2, which is the proportion of variation around the mean of one feature accounted for by the other feature.

3.2. Results

Fig. 1 displays the correlation matrix, using R^2 values, for each midpoint source and filter feature based on linear regression models using pooled data from Tasks 1 and 2. The strongest correlations are found within the filter features. Consistent with previous work (Högberg, 1997, Darch & Milner, 2008), individual formants are fairly strongly correlated with MFCCs, with F2 producing the largest R^2 value (i.e. the MFCCs account for 82.5% of the variation in F2). The relationship between formants and MFCCs can be explained by the fact that the MFCCs capture information about the spectrum, of which the peaks (i.e. the formants) are a dominant component. As expected, considerably weaker correlations are found within the source features.

Of most relevance to the present study are the correlations between source and filter features (green box, lower right of Fig. 1). The strongest of these correlations are between MFCCs and relative harmonics and between MFCCs and energy, with the MFCCs accounting for over 59.0% and 52.5% of the variation in each of these source features respectively. As with the formants, such relationships are predictable since MFCCs capture information about the entire spectrum. Despite the differences in their computation, relative harmonics are represented in the overall shape of the spectrum, which is itself captured by the MFCCs. In the same way, energy (i.e. overall amplitude) is an overall property of the spectrum and so should be encoded within the MFCCs (the 0th MFCC coefficient, which explicitly captures overall energy, was itself not included in the MFCC vectors for the purposes of analysis here). A weak correlation (R^2 = 0.324) was also found for MFCCs and f0. Aside from these, correlations between source and filter features were all extremely close to 0.

Experiment 1 provides evidence that source and filter features are largely independent of each other as demonstrated by the generally very weak correlations between the source and filter measures. The exception to this is the predictable incomplete decoupling of source and filter in extracting MFCCs, leading to correlations with some source features (i.e. measures related to amplitude).

4. Experiment 2: Correlations in between-speaker distances

While research question 1 focused on individual source and filter features, in research question 2, we were interested in the overall extent to which source and filter generally capture complementary speaker-specific information. To do this, it was necessary to reduce the complex multivariate data for both source and filter into a single value which captured speaker-specificity. We did this by using the raw data to calculate distances between each pair of speakers in our data set.

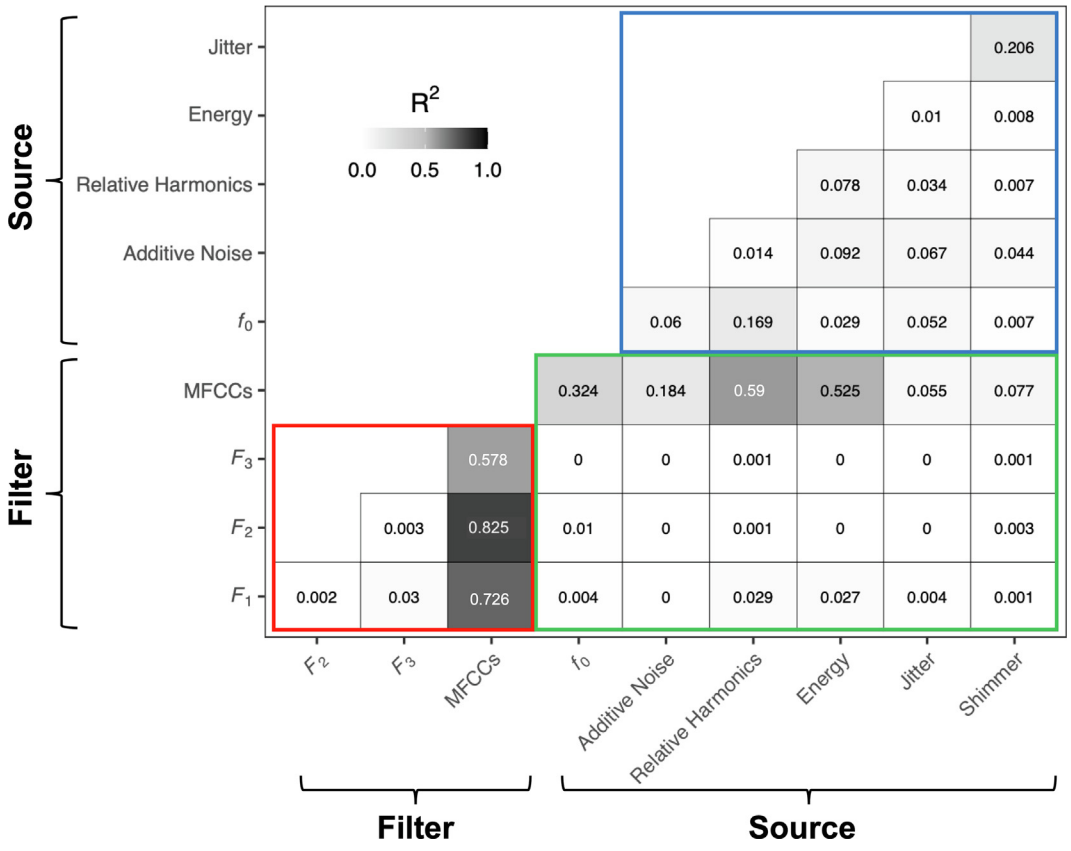

Fig. 1. Correlations (R^2 values rounded to three decimal places) within source (blue box) and filter (red box) features and between source and filter features (green box) based on data pooled across all 90 speakers across Tasks 1 and 2 (values of 0 indicate that the variable on the x-axis accounts for less than 0.01% of the variation in the variable on the y-axis).

4.1. Methods

Each speaker's midpoint data were fitted with a Gaussian distribution (more complex Gaussian mixture models were also tested but did not provide a better fit to the data). Given the nature of the MFCCs, additive noise, and spectral tilt measures, multivariate Gaussian distributions were fitted to these sets of raw data, whereas univariate Gaussians were fitted to the other features. Within each source and filter feature, Kullback-Leibler (KL; Campbell & Karam, 2010) divergences between speaker models were computed. KL is a measure of the similarity between two distributions; in this case, the similarity between data from two speakers. A KL divergence of zero means that the two distributions are identical, while the bigger the value the bigger the divergence between the distributions, and therefore between speakers. KL divergences between all pairs of speaker models were computed for each of the filter (formants and MFCCs) and source (f0, additive noise, spectral tilt, energy, jitter, and shimmer) features. Normalisation was then performed within features by first taking the natural logarithm of the values to reduce the skew in the data and then converting to z-scores to make distances comparable across features. Finally, the mean z-normalised KL divergence between each pair of speakers was calculated for the source and filter features separately.

4.2. Results

Fig. 2 displays a scatterplot of normalised KL divergences between each of the 90 speakers (4005 data points in total) based on source and filter features, fitted with a linear trend line. The filter results explain just 0.65% of the variation in the source results, indicating that the information for separating speakers captured by the source and filter measures is essentially independent. As was found in Experiment 1, these results support the idea that source and filter features provide mostly independent information about voices and, therefore, using both may give us a better understanding of speaker-specific characteristics of voices.

The following section examines whether the findings based on the raw data are replicated in terms of speaker discrimination, and whether adding source features to filter features improves speaker discrimination performance over filter features alone.

5. Experiment 3: Speaker discrimination

Experiment 3 complements experiment 2 in addressing our research question related to the overall extent to which source and filter capture speaker-specific information. In this experiment, however, we extend our analysis to conduct more formal speaker discrimination testing, in order to examine whether combining source and filter features can produce better speaker discrimination performance (i.e. capture more speaker-specific information), compared with source or filter in isolation.

5.1. Methods

The performance of each feature was evaluated using likelihood ratio (LR)-based testing. These procedures are standard within the fields of forensic speech science and ASR (see Morrison et al., 2021). In line with Morrison (2013), we refer to each set of tests conducted with different features and combinations of features as a *system*.

The LR is a measure of the probability of the evidence (in this case, the acoustic measurements) under two competing propositions: that the two samples contain the voice of the same speaker, and that the two samples contain the voices of different speakers. In the forensic setting, it indicates the relative extent to which the evidence supports the prosecution

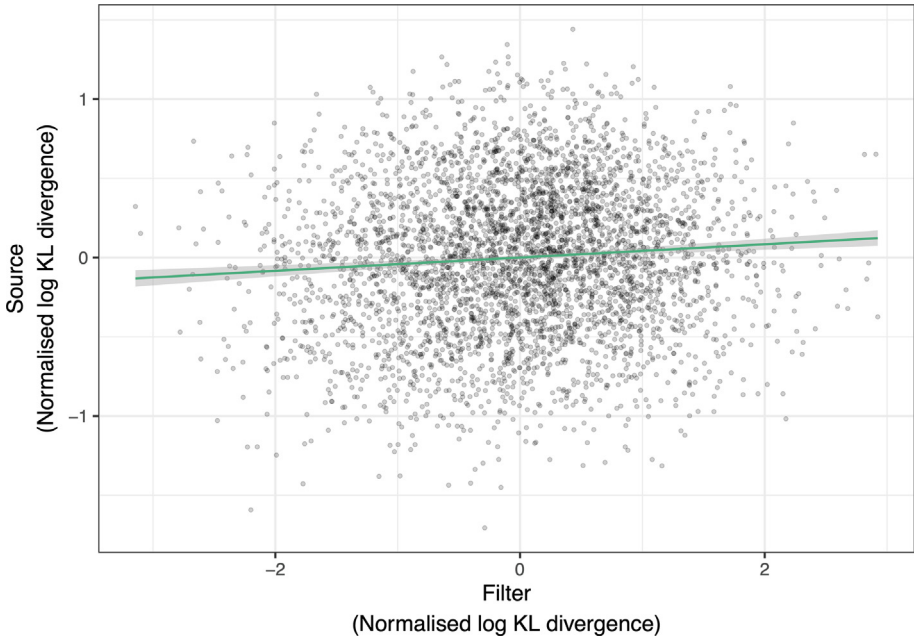

Fig. 2. Normalised and log transformed Kullback-Leibler divergences between each pair of speakers (4005 data points) based on filter features and source features fitted with a linear trend line (R^2 = 0.0064); the larger the value, the bigger the dissimilarity between speakers.

and defence (see Aitken & Taroni, 2004; Robertson, Vignaux, & Berger, 2016). To test speaker discrimination performance, pairs of samples are required where the ground truth is known about whether the samples came from the same- (SS) or different-speakers (DS). The 90 speakers were divided randomly into training, test, and reference sets containing 30 speakers each. SS and DS comparisons were performed for the training and test speakers using data from Task 1 (interview) and Task 2 (phone call). Each comparison generates a LR-like score (i.e. a numerical value) which captures the similarity between the two sets of data, and the typicality of the data relative to the wider population, based on a model generated using data from speakers in the reference set. The test scores were then converted to log LRs (LLRs) using logistic-regression calibration (see Morrison, 2013). This process involves shifting and scaling the test scores based on calibration coefficients learned from the training scores to improve interpretability, comparability, and performance. Scores from different features were combined using logistic-regression fusion (Pigeon, Druyts, & Verlinde, 2000), a method of combining results that accounts for the underlying correlations in the scores.

System performance was evaluated using the equal error rate (EER) as a measure of the absolute proportion of contrary-to-fact LLRs (i.e. *errors*), and the log LR cost (C_{llr}; Brümmer & du Preez, 2006). C_{llr} penalises the system based on the magnitude of contrary-to-fact LLRs (i.e. how big the error is). For both EER and C_{llr}, values closer to zero indicate better performance, indicating fewer and less severe errors respectively. A C_{llr} of 1 or greater indicates that the system provides no speaker discriminating information (a system that consistently produced LRs of 1, i.e. no support for either side, would also produce a C_{llr} of 1). As outlined in Wang, Hughes, and Foulkes (2019a, 2019b), speaker discrimination performance is susceptible to variation depending on the configuration of speakers in each of the three speaker sets. Therefore, each set of tests was replicated 20 times using randomly selected configurations of training, test, and reference speakers. Overall performance is therefore assessed based on the distribution of EER and C_{llr} values across the 20 replications.

In addition to the midpoint data used in Experiment 1, speaker discrimination testing was performed using dynamic and holistic representations of some features. This allows us to assess the best ways in which to capture speaker-discriminatory information for each feature. The models for computing LRs also allow us to preserve the multivariate nature of some of the features (e.g. combinations of three formants, additive noise, and relative harmonics):

- Dynamic data were analysed for each of the formants and f0 using time-normalised measurements at + 10% steps across the duration of the vowel. These data were then fitted with quadratic polynomials of the form $y = f(x) = ax^2 + bx + c$, which capture information about the shape of the trajectory. Curve fitting via polynomials has the benefit of reducing dimensionality, with each curve of nine raw data points reduced to three coefficients. Such parametric representations of trajectories have also been shown to perform better at speaker discrimination than the equivalent raw data (McDougall, 2006, Morrison, 2009). Quadratic representations were chosen based on visual inspection of the raw data, expectations about the shape of the trajectories, and because they proved optimal in Hughes et al. (2016).
- Holistic data were analysed for each of the features extracted at the frame level, i.e. MFCCs, energy, additive noise, and relative harmonics. This involves pooling and then modelling the data from all frames across the duration of the vowel. This methodology is used in ASR, where measurements are taken across an entire recording, rather than a single vowel.
- For jitter and shimmer, only data from the frame at the temporal midpoint was analysed in terms of speaker discrimination because of the well-known unreliability of measures taken at different points in time, especially where the speech material is uncontrolled (see Labuschagne & Ciocca, 2016).

LR-like scores were computed using either multivariate kernel density (MVKD; Aitken & Lucy, 2004) or Gaussian mixture model-universal background model (GMM-UBM; Reynolds, Quatieri, & Dunn, 2000). Different methods were chosen based on their use with different types of features in previous work. MVKD has generally been used for linguistic variables and token-based features, with small numbers of correlated dimensions (e.g. formant midpoints or dynamics), whereas GMM-UBM was developed for ASR and so is more suitable for frame-based analysis. MVKD models the reference data with a kernel density made up of equally weighted Gaussians from each reference speaker and the target speaker data with a normal distribution. GMM-UBM was, at one time, the state-of-the-art approach in ASR systems, although has now been superseded by approaches which utilise deep learning (e.g. xVectors), but which are more opaque in terms of the speaker-specific information that the systems capture. GMM-UBM is more suited to the aims of the present study as it provides the most transparency (i.e. minimal transformation of data) in the relationship between the input features and output probabilities. In GMM-UBM, the reference data are pooled to build a speaker-independent, universal background model (UBM). The target speaker model is built by copying the UBM and then adapting it towards the target speaker data in a process called maximum a posteriori (MAP) adaptation. MAP adaptation is particularly appropriate where there are small amounts of target speaker data; the more target data there is available, the closer the target model is to the target data.

MVKD and GMM-UBM have both been used extensively to examine linguistic and ASR data. These approaches also have a transparent relationship between the input features and the output scores, with relatively little data transformation in between (unlike modern state-of-the-art ASR systems). Table 1 shows the measurement methods for each feature tested and the approach used to compute scores. For the GMM-UBM systems, the number of Gaussians was determined based on pretesting. The best performing measure for each feature was used to generate the best combination of source and filter features, respectively. The best source features were then fused with the best filter features to assess whether their combination outperformed either source or filter features in isolation. Improvement in performance when combined indicates that features capture independent speaker discriminatory information.

Table 1
Source and filter features, measurement methods, and LR score computation methods (for the purposes of speaker discrimination testing).

	Feature	Measurement method	LR score method
Filter	Formants	Midpoint	MVKD
		Dynamic	MVKD
	MFCCs	Midpoint	MVKD
		Holistic	GMM-UBM (64 Gaussians)
Source	f0	Midpoint	MVKD
		Dynamic	MVKD
	Additive noise	Midpoint	MVKD
		Holistic	GMM-UBM (128 Gaussians)
	Relative harmonics	Midpoint	MVKD
		Holistic	GMM-UBM (128 Gaussians)
	Energy	Midpoint	MVKD
		Holistic	GMM-UBM (128 Gaussians)
	Jitter	Midpoint	MVKD
	Shimmer	Midpoint	MVKD

5.2. Results

Table 2 shows the speaker discriminatory performance of the different measurement methods for each source and filter feature across the 20 replications. A reminder here that the closer the EER and C_{llr} are to 0, the better the performance. Overall, the filter features perform markedly better than the source features, with holistic MFCCs extracted from across the duration of the vowel providing the best overall performance (mean EER = 3.1%, mean C_{llr} = 0.14). As in Hughes et al. (2016), the dynamic formant information substantially outperforms the midpoint data, indicating that the shape of formant trajectories encodes speaker-specific information beyond what is captured by absolute frequency alone. These findings provide further evidence that vowel trajectory shape carries useful cues for speaker differences. It also supports the recent trend in production studies to analyse vowel trajectory rather than midpoint measurements, even when phonological monophthongs are the variable of interest (e.g. Haddican, Foulkes, Hughes, & Richards, 2013, Docherty, Gonzalez, & Mitchell, 2015). The best performing source feature is the relative harmonics, followed by additive noise. f0 performs best when using dynamic data, although the improvement in performance over the midpoint-only measure is relatively small. Finally, across all 20 replications, energy performed worst, providing little to no speaker discriminatory information (the C_{llr} was consistently close to or greater than 1).

All possible combinations were fused together to find the best performing combination of filter features and the best combination of source features. Fusing the holistic MFCCs and dynamic formants improved filter performance relative to either MFCCs or formants individually, producing a mean EER of 0.72% and mean C_{llr} of 0.05. The best source performance was generated by fusing dynamic f0, midpoint additive noise, holistic relative harmonics, and midpoint energy measures. This combination produced a mean EER of 8.25% and mean C_{llr} of 0.32, marginally better (by 0.0041 in terms of mean C_{llr}) than the performance with energy excluded. This shows that combining features can help improve overall performance, even when those features individually provide little speaker discriminating information. However, overall performance was worse when including jitter and shimmer.

The best source combination and the best filter combination (MFCCs, formants, f0, additive noise, relative harmonics, and energy; see Table 2) were then fused and compared with the best filter-only features (MFCCs and formants) to assess potential improvements in performance when adding source information. Fig. 3 displays EER and C_{llr} values for the filter features and the combination of source and filter features for each of the 20 replications. Overall, the addition of source features reduced EER; for the source-filter combination mean EER was 0.55%, compared with 0.72% for filter features only. The addition of the source features had essentially no effect on C_{llr}. However, different replications, using different configura-

Table 2
Speaker discrimination performance (EER and Cllr; in both cases the closer to zero the better the performance) for each source and filter feature across the 20 replications, as well as the best performing combination within the set of source features and within the set of filter features (bold).

			EER (%)				C_{llr}			
			Mean	Min	Max	Std	Mean	Min	Max	Std
Filter	Formants	Midpoint	13.2	9.9	17.4	2.5	0.45	0.33	0.57	0.06
		Dynamic	9.2	4.2	13.2	2.1	0.35	0.21	0.45	0.07
	MFCCs	Midpoint	16.8	9.9	26.7	4.3	0.56	0.39	0.84	0.13
		Holistic	3.1	0.1	5.9	1.1	0.14	0.06	0.23	0.05
	Best filter combination:formants (dynamic) & MFCCs (holistic)		**0.72**	**0**	**3.39**	**1.09**	**0.05**	**0.02**	**0.13**	**0.03**
Source	f0	Midpoint	27.0	20.1	33.8	3.8	0.76	0.65	0.89	0.08
		Dynamic	21.8	16.8	26.8	2.5	0.67	0.55	0.77	0.06
	Additive noise	Midpoint	17.2	10.6	25.8	3.6	0.61	0.49	0.78	0.08
		Holistic	17.7	13.3	23.2	2.7	0.62	0.82	0.49	0.08
	Relative harmonics	Midpoint	23.8	17.1	29.9	3.4	0.76	0.63	0.97	0.08
		Holistic	16.4	12.8	20.0	2.3	0.61	0.50	0.78	0.06
	Energy	Midpoint	46.0	33.0	66.8	11.1	1.02	0.99	1.24	0.06
		Holistic	47.7	36.0	64.1	9.5	1.01	0.99	1.13	0.03
	Jitter	Midpoint	34.8	26.7	42.7	5.0	0.93	0.82	1.12	0.07
	Shimmer	Midpoint	34.2	29.0	42.5	3.5	0.91	0.83	0.98	0.04
	Best source combination:f0 (dynamic), additive noise (midpoint), relative harmonics (holistic), & energy (midpoint)		**8.25**	**3.85**	**13.33**	**2.41**	**0.32**	**0.19**	**0.49**	**0.09**

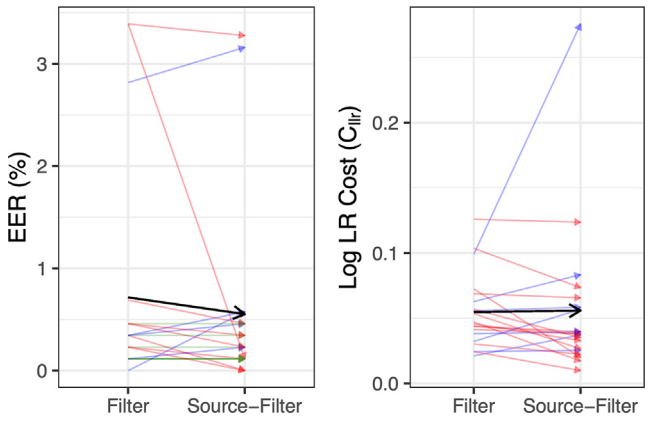

Fig. 3. Speaker discrimination performance (EER, left; C_{llr}, right) based on filter features and a combination of source and filter features across the 20 replications (each arrow is a separate replication; red = better performance when adding source, blue = worse performance when adding source, green = no difference when adding source; black = mean filter and source-filter performance across the 20 replications).

tions of speakers, showed more or less improvement. In eight of the 20 replications, EER was lower (by an average of 57%) when combining source and filter, compared with filter features alone. In two of these replications, EER was reduced to 0% (i.e. no errors) with the addition of source features. There was no change in EER with addition of source features for seven replications, while for the remaining five replications, EER was slightly worse when source and filter were combined.

In these replications, source features provide little to no speaker-specific information. In 13 of the 20 replications, the addition of source information reduced C_{llr} by an average of 31%, with some replications producing reductions of as much as 73% or as little as 2%. In the remaining seven replications, C_{llr} was higher (by an average of 54%, but largely driven by one replication where there was a marked increase in C_{llr}) for the combination of source and filter features compared with filter features alone. The variability across replications (shown in Fig. 3) therefore suggests that the relative discriminatory value of source and filter information differs across individual speakers – i.e. there are some speakers where the addition of source information will be beneficial to speaker discrimination and others where it will be detrimental. Examining overall performance alone may be masking such speaker-specific effects.

To investigate this issue further, mean same-speaker (SS) and different-speaker (DS) LLRs for each speaker across all 20 replications were calculated based on the source features, filter features, and source and filter features combined. As well as outperforming the source features in terms of overall performance, the filter features also produced considerably stronger LLRs by around three orders of magnitude on average. Fig. 4 displays the difference between the mean SS and DS LLRs for the source and filter features combined, relative to the filter features alone for each speaker. Positive values on the SS dimension (x-axis) and negative values on the DS dimension (y-axis) indicate stronger LLRs when combining source and filter information. From the perspective of source-filter combination, the red quadrant (top right) represents the best performance. It contains speakers for whom the mean SS and DS LLRs increased in magnitude when combining source and filter features (i.e. where the addition of source features helped performance). For example, for a speaker who produces a mean SS LLR of +2 for filter-only features and +3 for source-filter combined, the value on the x-axis would be +1, reflecting an improvement in the magnitude of the evidence by including source information. For a speaker who produces a mean DS LLR of −2 for filter-only features and −4 for source-filter, the value on the y-axis would be −2. The green quadrants (top left and bottom right) demonstrate improvement in one direction (either SS or DS) and worsening performance in the other. They contain speakers displaying either an increase in SS magnitude and decrease in DS magnitude (bottom right; although no speakers displayed this pattern) or increase in DS magnitude and decrease in SS magnitude (top left). The empty blue quadrant (bottom left) would contain speakers whose LLRs were weaker when combining source and filter features compared with filter features alone. It is noteworthy that no speakers were in this category (appear in this quadrant).

For 57 of the 90 speakers (63.3%), the addition of source information increased the mean SS LLR. This increase was on average one order of LLR magnitude (i.e. one unit on the scales in Fig. 4), although for one speaker (#4) the mean SS LLR increased by over two orders of magnitude. However, for over one third of speakers (36.7%), SS LLRs were weaker when adding source features. This could be due to considerable homogeneity across the DyViS speakers in terms of laryngeal VQ (see San Segundo et al., 2019). In the case of one speaker (#88), SS LLRs were on average over two orders of magnitude weaker when combining source and filter compared with using filter features alone. The effects of adding source information were more marked for the DS comparisons, with all speakers displaying stronger mean DS LLRs; an average increase of 2.7 orders of magnitude compared with the filter features alone. For two speakers (#19 and #53), the difference between the source-filter and filter-only input was over five orders of magnitude. Taken together, the data in Fig. 4 show that adding source information to the filter information affects individuals in different ways. However, overall, the speaker-specific information captured by the source features was sufficient to aid discrimination either in SS or DS comparisons, with no speaker producing both weaker SS and DS LLRs when combining source and filter (all speakers are positioned in the upper two quadrants of Fig. 4).

This experiment suggests that while filter-features alone capture more speaker-specific information than source-features alone, in general the combination captures the most speaker-specific information. However, the extent of the improvement varies substantially across speakers.

6. Discussion

This paper set out to examine which parts of the speech signal encode the greatest speaker-specific information. We did this by analysing a range of acoustic features grouped broadly as source and filter, to make generalisations about groups of features and to relate speaker-specific information to models of speech production. Our research questions aimed to (1) examine the covariation between different acoustic source and filter features, (2) assess the extent to which source and filter capture complementary speaker-specific information, and (3) evaluate whether combining source and filter features

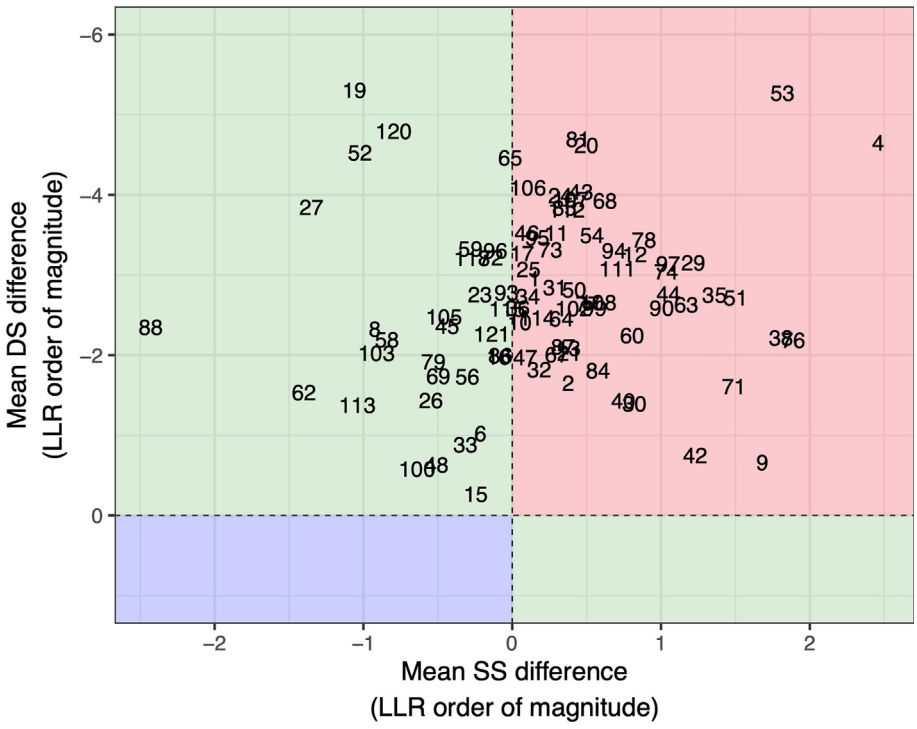

Fig. 4. Differences in mean same-speaker (SS) and different-speaker (DS) log likelihood ratios (LLRs; i.e. strength of evidence) between the source and filter feature combination and filter-only features for each speaker (numbers refer to speaker codes in the DyViS corpus). Larger positive values on the x-axis indicate an increase in the strength of SS results; larger negative values on the y-axis indicate an increase in the strength of DS results (note the inversion of the y-axis to capture this).

improves speaker discrimination performance. Below we discuss our results in light of these research questions.

6.1. Contributions to a model of speaker-specificity

Given the focus here on data extracted from a single phonetic segment, to maximise control and comparability, the features we include underlie what Nolan (1983) refers to as *long term quality* specifically within the segmental strand of his multi-layered model of speech. Our findings contribute to this model by demonstrating that more speaker-specificity lies in filter rather the source features within the segmental strand and that these sets of features can be treated, in general, as being mutually independent of each other.

In terms of complementarity, we find some correlations within the filter features, consistent with the findings of previous work (Darch & Milner, 2008). The results indicate that MFCCs, as a representation of the overall shape of the spectrum, unsurprisingly encode information about formants (i.e. spectral peaks). This interdependence is also consistent with the findings of Hughes et al. (2017), who report very little improvement when combining an MFCC-based ASR system with long term formants, where MFCCs and formants were extracted from frames across all vowels within a recording. Within the source features, there were no meaningful correlations, with R^2 values reaching no higher than 0.206. The lack of interdependence between source features is consistent with the fact that the combination of f0, additive noise, relative harmonics and energy produced the best speaker discrimination performance (i.e. each captures independent speaker-specific information). Given the lack of correlations, the fact that jitter and shimmer did not improve source performance may be attributed to the uncertainty around how to measure them most accurately (see e.g. Woubie, Koivisto, & Bäckström, 2021).

The results of the midpoint analysis in 3.2 do reveal a small number of correlations between source and filter, but in all cases, as predicted, these correlations involve MFCCs and the source features that capture some overall property of the spectrum (i.e. relative harmonics and energy). Given what MFCCs are in principle modelling within the speech signal, such findings are unsurprising. However, MFCCs were also able to predict 32.4% of the variation in f0. The relationship between MFCCs and f0 is consistent with previous work (Darch & Milner, 2008), but, as with the results for formants, the correlation is weaker than in Hughes, Clermont, and Harrison (2020) due to the speaker-independent methods used in the present study. Although MFCCs involve smoothing over harmonic information, there is still likely to be some sensitivity to harmonic spacing especially when the number of MFCCs is high. Hughes et al. (2020) show that the strength of correlations between MFCCs and f0 (and, indeed, formants) increases with the number of cepstral coefficients. With these exceptions, in the present data, there were only very weak correlations between source and filter features. This is the case even where correlations may be predicted, for example, between individual formants and f0 (Assmann & Nearey, 2007, Assmann, 2008). This may be because previous studies into the relationship between formants and f0 have tended to use data averaged by-speaker. The use of pooled data here may be masking some underlying relationship between measures.

The results of speaker discrimination testing show that filter features generally capture more speaker-specific information

than source features. Our results are also largely consistent with the underlying independence of source and filter features demonstrated within the raw data, albeit not in a straightforward way. In many of the 20 replications in 5.2, speaker discrimination performance was improved when combining source and filter, compared with using filter features alone. The logistic-regression fusion technique used to combine results accounts for correlations in the underlying scores. Therefore, any improvement in performance indicates that the two sets of scores (derived from source and filter) are providing complementary speaker-specific information. However, some replications produced essentially the same performance whether source features were included or not, while other replications led to poorer performance. One explanation for the replications with poorer performance is the sensitivity of EER and C_{llr} when the performance of the filter features alone is already so good. In most replications, the filter features produced EERs of less than 1%. This reflects a floor effect where it is difficult to improve performance any further, due, in large part, to the use of high-quality recordings and the use of MFCCs, which are known to be very good speaker discriminatory features. Meanwhile, if the source information adds even a single error (*contrary-to-fact* LLR) to the system, this can have a drastic effect on relative performance.

As well as the issue of the floor effect, a key finding of 5.2 is that the behaviour of individual speakers plays a role in the lack of systematic effects in speaker discrimination when combining source and filter information. Thus, for some configurations of speakers, the source information is helpful in separating speakers and leads to an overall improvement in system performance. By contrast, for other configurations of speakers it can lead to no improvement or even poorer system performance. Further, Fig. 4 highlights that even if overall performance is unaffected (e.g. because performance is already very good), the addition of source features can increase the magnitude of LLRs (i.e. the strength of the evidence) in SS or DS comparisons, or indeed in both. The extent of such increases is dependent on the speaker. Thus, taken together, our results demonstrate that different speakers differ on different dimensions (i.e. the addition of source information is only likely to be of use for certain speakers and maybe only in certain contexts).

6.2. Implications

In this section we outline the implications of our findings for both phonetic theory and for the applied fields of forensic speech science and ASR.

6.2.1. Models of speech production

For over six decades, source-filter theory (Fant, 1960) has been the dominant model of speech production in phonetics and speech science. An important assumption of source-filter theory is that the two components are independent of each other. This assumption successfully underpins many analytical methods of acoustic analysis (e.g. linear prediction) and applications of speech technology (e.g. speech synthesis). On a general level, the considerable independence between features in our study, particularly when considering the raw data, is consistent with the theoretical assumption of independence, even in spite of the incomplete decoupling of source and filter in deriving MFCCs. However, within our data we do find some evidence of interrelationships between source and filter, particularly at the level of the individual. This finding is, in fact, consistent with claims in the literature on non-linear source-filter interactions (Titze, 2008, Titze, Riede, & Popolo, 2008). Indeed, Maxfield, Palaparthi, and Titze (2017: 149) go further, claiming that “the degree to which the source and filter are coupled is variable, both between and within individuals”. The question is then, what factors determine the degree to which source and filter are related between and within individuals?

Titze (2008, p. 2733) claims that “it has been recognised all along . . . that the linear theory (of source-filter independence) is more applicable to male speech than female or child speech”. The results of our study show that this also applies to variability across speakers of the same sex and age. Thus, it is probable that variation in the size and shape of the vocal apparatus, and particularly deviations from the average male 17.5 cm tube often used as a vocal tract model, are likely to produce variation in source-filter interaction. Specifically, according to Titze (2008), greater source-filter interaction should be produced with narrower vocal tracts and constriction of the vocal tract closer to the vocal folds (i.e. in the velar, uvular and pharyngeal regions of the vocal tract). This, in turn, may relate to differences in vocal setting and have knock on effects for voice quality and thus also acoustic measures of source features. This is because a narrower vocal tract results in an increased acoustic reactance which reduces the impedance difference between source and filter, affecting the transglottal pressure and hence glottal airflow (Titze, 2008). Further, the closer the f0 or any other harmonics to F1 or indeed any formant, the greater the predicted interaction. Finally, the functioning of the vocal folds is also likely to determine, to some extent, the degree to which source and filter are related. Modal vocal fold vibration (e.g. in loud speech where the vocal folds are vibrating most efficiently) is likely, all else being equal, to produce the greatest source-filter independence. Non-modal laryngeal voice quality should, therefore, produce greater interaction. In reality, however, the degree of interaction between source and filter between and within speakers is likely caused by a complex interaction of both short- and long-term factors relating to physiology, acquired patterns of speech production (which may or may not be related to group-level behaviour; e.g. long-term voice quality and vocal setting), and the specific linguistic content at a given point in time.

6.2.2. Forensic speech science

An essential goal of forensic speech science research is to identify the features of the voice that carry the greatest speaker-specificity, which then inform the choice of features for analysis in forensic casework. Our study has shown that under optimal conditions the addition of source information can improve speaker discrimination over filter features alone (although in forensic casework, it is also necessary to consider a variety of other confounding factors, such as channel, background noise, duration etc.). However, the fact that such improvements are not evenly distributed across speakers shows that the choice of features in speaker discrimination (at least when using linguistic methods of analysis) is itself speaker-specific (and likely variety-specific as well, see

Foulkes & Hughes, in press). A key question for forensics, then, is whether we might be able to predict, based on acoustic data, which speakers are likely to produce stronger or weaker speaker discrimination results when combining source and filter features, which in turn could inform the choice of features in a forensic case.

The factors outlined in 6.2.1 lead to a set of specific predictions, which deserve further empirical testing. Source-filter interaction should, in principle, be greatest for speakers who (i) have high f0 and/or low F1, (ii) have raised larynx and/or backed tongue body or pharyngeal constriction as long-term vocal settings, and (iii) have non-modal laryngeal voice qualities. More systematic work is ongoing to identify which sets of source and filter features are appropriate for the given voice(s) in a forensic case (Hughes, Harrison, Foulkes, Wormald, Xu, van der Vloed, & Kelly, 2022; Hughes et al., 2022–25).

6.2.3. Automatic speaker recognition (ASR)

As outlined in 1.2, there has previously been some success in integrating linguistic analysis into ASR systems. Indeed, the work presented in the present study provides further evidence that voice quality features can help to improve speaker discrimination over the use of MFCCs alone (see also Hughes et al. 2019). However, much of this work, as is the case here, uses old paradigms within ASR. State-of-the-art ASR systems (specifically xVector approaches) now perform extremely well, even under more challenging, forensically realistic conditions (see Morrison & Enzinger, 2019). Given this, what is the role of phonetics within ASR? We still believe that phonetic analysis can help improve ASR performance, as well as to provide checks on the output of ASR analyses. While overall error rates may be very low, not all speakers will be equally well discriminated by an ASR system. As highlighted in the present study, the addition of source features may be of considerable benefit to some speakers. In applications of ASR, such as verification for security or in forensic contexts, the performance of the system with the specific voice(s) under analysis is of central concern, rather than the general performance of the system over many trials. If phonetic analysis can help in even one such case, this can have potentially life-changing implications and is thus worthwhile. Beyond speaker discrimination performance, phonetic analysis can help us understand what information ASR systems capture. This is important given the increasingly 'black box' nature of state-of-the-art ASR systems that involve machine learning.

7. Conclusions

The present study examined speaker-specific behaviour in speech production, through the analysis of a range of acoustic source and filter features. Overall, we find considerable evidence in support of the independence of source and filter in the context of speaker discrimination. This underlying independence between source and filter is capable of improving speaker discrimination performance (see also Gonzalez-Rodriguez et al. 2014, Hughes et al. 2017). However, in line with work such as Titze (2008), as well as predictions about the covariance of source and filter output due to physiological and articulatory factors, we do find some, albeit limited, correlation between source and filter features. Therefore, the extent of the improvement in speaker discrimination is dependent on the individual speakers under analysis. Some speakers produce stronger speaker discrimination results when combining source and filter, compared with filter alone, while others do not. These results also point to the benefit of including both source and filter measures in linguistic studies that attempt to understand individual and/or community-level variation as these features are not as independent as previously suggested for all speakers.

CRediT authorship contribution statement

Vincent Hughes: Conceptualization, Methodology, Software, Formal analysis, Writing – original draft, Writing – review & editing, Visualization. **Amanda Cardoso:** Conceptualization, Investigation, Writing – review & editing. **Paul Foulkes:** Conceptualization, Writing – review & editing. **Peter French:** Conceptualization. **Amelia Gully:** Conceptualization, Writing – original draft, Writing – review & editing. **Philip Harrison:** Conceptualization, Software, Writing – review & editing.

Acknowledgements

This work was supported by the UK Arts and Humanities Research Council (Project Reference: AH/M003396/1). The authors would like to thank Dr Frantz Clermont for all his guidance and insight on this project, particularly with regard to the analysis of the correlations in the raw data and resolving issues relating to cepstral mean and variance normalisation.

References

Aitken, C. G. G., & Lucy, D. (2004). Evaluation of trace evidence in the form of multivariate data. *Applied Statistics, 54*, 109–122.

Aitken, C. G. G., & Taroni, F. (2004). *Statistics and the Evaluation of Evidence for Forensic Scientists* (2nd edition). Chichester: John Wiley.

Assmann, P. F. (2008). Developmental study of the relationship between f0 and formant frequencies. *Journal of the Acoustical Society of America, 124*, 2556.

Assmann, P. F., & Nearey, T. M. (2007). Relationship between fundamental and formant frequencies in voice preference. *Journal of Acoustical Society of America, 122*, EL35-43.

Awan, S. N., Solomon, N. P., & Helou, L. B. (2013). Spectral-cepstral estimation of dysphonia severity: External validation. *European Archives of Otorhinolarynology, 122*, 40–48.

Boersma, P., & Weenink, D. (2019). Praat: Doing phonetic by computer. *Version, 6*, 49 http://www.praat.org/.

Braun, A. (1995). Fundamental frequency: How speaker specific is it? *Studies in Forensic Phonetics (BEIPHOL, 64)*, 9–23.

Britain, D. (2013). Space, diffusion and mobility. In J. Chambers & N. Schilling (Eds.), *Handbook of Language Variation and Change* (2nd edition, pp. 471–500). Oxford: Wiley-Blackwell.

Broad, D. J., & Clermont, F. (1989). Formant estimation by linear transformation of the LPC cepstrum. *Journal of the Acoustical Society of America, 86*, 2013–2017.

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

Cardoso, A., Foulkes, P., French, J. P., Harrison, P., Hughes, V., Kavanagh, C. & San Segundo, E. (2018). Voice quality of hesitations: acoustic measures and VPA ratings. Paper presented at annual conference of the *International Association for Forensic Phonetics and Acoustics*, University of Huddersfield.

Campbell, W. M., & Karam, Z. N. (2010). Simple and efficient speaker comparison using approximate KL divergence. In *Proceedings of Interspeech* (pp. 362–365).

Darch, J., & Milner, B. (2008). Analysis and prediction of acoustic speech features from mel-frequency cepstral coefficients in distributed speech recognition architectures. *Journal of the Acoustical Society of America, 124*, 3989–4000.

Davis, S., & Mermelstein, P. (1980). Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing, 28*, 357–366.

de Krom, G. (1993). A cepstrum-based technique for determining a harmonics-to-noise ratio in speech signals. *Journal of Speech, Language and Hearing Research, 36*, 254–266.

Docherty, G., Gonzalez, S., & Mitchell, N. (2015). Static vs dynamic perspectives on the realisation of vowel nucleii in West Australian English. *Proceedings of the 18th International Congress of Phonetic Sciences*.

Ellis, D. P. W. (2005). PLP and RASTA (and MFCC, and inversion) in Matlab. Online web resource: https://www.ee.columbia.edu/~dpwe/resources/matlab/rastamat/.

Enzinger, E., Zhang, C., & Morrison, G. S. (2012). Voice source features for forensic voice comparison - an evaluation of the GLOTTEX software package. In *Proceedings of Odyssey: The Speaker and Language Recognition Workshop* (pp. 78–85).

Fant, G. (1960). *Acoustic Theory of Speech Production*. The Hague: Mouton.

Farrús, M., Hernando, J., & Ejarque, P. (2007). Jitter and shimmer measures for speaker recognition. In *Proceedings of Interspeech* (pp. 778–781).

Finger, L. S., Cielo, C. A., & Schwarz, K. (2009). Acoustic vocal measures in women without voice complaints and with normal larynxes. *Brazilian Journal of Otorhinolaryngology, 75*, 432–440.

Foulkes, P., Scobbie, J. M., & Watt, D. (2010). Sociophonetics. In W. Hardcastle, J. Laver, & F. Gibbon (Eds.), *Handbook of Phonetic Sciences* (2nd edition, pp. 703–754). Oxford: Blackwell.

Foulkes, P., Docherty, G., Shattuck-Hafnagel, S., & Hughes, V. (2018). Three steps forward for predictability: Consideration of methodological robustness, indexical and prosodic factors, and replication in the laboratory. *Linguistics Vanguard (special edition on The Role of Predictability in Shaping Human Language Sound Patterns), 4*(2).

Foulkes, P. & Hughes, V. (in press). Dialectological and sociolinguistic foundations of forensic speaker comparison. To appear in Nolan, F., McDougall K. & Hudson, T. (eds.) *Oxford Handbook of Forensic Phonetics*. Oxford: Oxford University Press.

Garellek, M. (2019). The phonetics of voice. In W. F. Katz & P. F. Assmann (Eds.), *The Routledge Handbook of Phonetics* (pp. 75–106). Abingdon: Taylor and Francis.

Garvin, P. L., & Ladefoged, P. (1963). Speaker identification and message identification in speech recognition. *Phonetica, 9*, 163–199.

Gold, E., & French, J. P. (2011). International practices in forensic speaker comparison. *International Journal of Speech, Language and the Law, 18*, 293–307.

Gold, E., & French, J. P. (2019). International practices in forensic speaker comparisons: Second survey. *International Journal of Speech, Language and the Law, 26*, 1–20.

González-Rodríguez, J., Gil, J., Pérez, R., & Franco-Pedroso, J. (2014). What are we missing with i-vectors? A perceptual analysis of i-vector-based falsely accepted trials. In *Proceedings of Odyssey: The Speaker and Language Recognition Workshop* (pp. 33–40).

Gordon, M., & Ladefoged, P. (2001). Phonation types: A cross-linguistic overview. *Journal of Phonetics, 29*, 383–406.

Haddican, B., Foulkes, P., Hughes, V., & Richards, H. (2013). Interaction of social and linguistic constraints on two vowel changes in northern England. *Language Variation and Change, 25*(3), 371–403.

Hillenbrand, J., Cleveland, R. A., & Erickson, R. L. (1994). Acoustic correlates of breathy vocal quality. *Journal of Speech Language and Hearing Research, 37*, 769–778.

Högberg, J. (1997). Prediction of formant frequencies from linear combinations of filterbank and cepstral coefficients. *Department for Speech, Music and Hearing Quarterly Progress and Status Report (TMH-QPSR), 4*, 41–49.

Hudson, T., de Jong, G., McDougall, K. & Nolan, F. (2007). f0 statistics for 100 young male speakers of standard Southern British English. In Trouvain, J. & Barry, W. J. (eds.) *Proceedings of the 16th International Congress of Phonetic Sciences*, Saarbrucken, Germany, pp. 1809–1812.

Hughes, V. (2014). *The Definition of the Relevant Population and the Collection of Data for Likelihood Ratio-Based Forensic Voice Comparison*. University of York, UK. PhD Thesis.

Hughes, V., Wood, S., & Foulkes, P. (2016). Strength of forensic voice comparison evidence from the acoustics of filled pauses. *International Journal of Speech, Language and the Law, 23*, 99–132.

Hughes, V., Harrison, P., Foulkes, P., French, J. P., Kavanagh, C. & San Segundo, E. (2017). Mapping across feature spaces in forensic voice comparison: the contribution of auditory-based voice quality to (semi-)automatic system testing. *Proceedings of Interspeech*, Stockholm, Sweden, pp. 3892–3896.

Hughes, V., Cardoso, A., Foulkes, P., French, J. P., Harrison, P. & Gully, A. (2019). Forensic voice comparison using long-term acoustic measures of voice quality. In S. Calhoun, P. Escudero, M. Tabain & P. Warren (eds.) *Proceedings of the 19th International Congress of Phonetic Sciences,* Melbourne, Australia, pp. 1455-1459. Canberra, Australia: Australasian Speech Science and Technology Association Inc.

Hughes, V., Clermont, F. & Harrison, P. (2020). Correlating cepstra with formant frequencies: implications for phonetically-informed forensic voice comparison. *Proceedings of Interspeech*. Shanghai, China, pp. 1857-1862.

Hughes, V., Harrison, P., Foulkes, P., Wormald, J., Xu, C., van der Vloed, D. & Kelly, F. (2022) Person-specific automatic speaker recognition: understanding the behaviour of individuals for applications of ASR. Poster presented at IAFPA 2022, Charles University, Prague, Czechia. 10-13 July 2022.

Hughes, V., Harrison, P. & Foulkes, P. (2022–25) Person-specific automatic speaker recognition: understanding the behaviour of individuals for applications of ASR. ESRC-funded project: ES/W001241/1. https://pasr.york.ac.uk.

Iseli, M., Shue, Y-L. & Alwan, A. (2006). Age- and gender-dependent analysis of voice source characteristics. *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, Toulouse, France, pp. 389-392.

Jessen, M. (1997). Speaker-specific information in voice quality parameters. *Forensic Linguistics, 4*, 84–103.

Jessen, M., Köster, O., & Gfroerer, S. (2005). Influence of vocal effort on average and variability of fundamental frequency. *International Journal of Speech, Language and the Law, 12*, 174–213.

Jurafsky, D., & Martin, J. H. (2009). *Speech and Language Processing: Computational Linguistics and Speech Recognition* (2nd edition). New Jersey: Prentice-Hall.

Keating, P., Garellek, M. & Kreiman, J. (2015). Acoustic properties of different kinds of creaky voice. *Proceedings of the 18th International Congress of Phonetic Sciences*, Glasgow, Scotland. ISBN 978-0-85261-941-4.

Kinoshita, Y., Ishihara, S., & Rose, P. (2009). Exploring the discriminatory potential of F0 distribution parameters in traditional forensic speaker recognition. *International Journal of Speech, Language and the Law, 16*, 91–111.

Klug, K., Kirchhübel, C., Foulkes, P. & French, J. P. (2019). Analysing breathy voice in forensic speaker comparison: using acoustics to confirm perception. In S. Calhoun, P. Escudero, M. Tabain & P. Warren (eds.) *Proceedings of the 19th International Congress of Phonetic Sciences,* Melbourne, Australia, pp. 795-799. Canberra, Australia: Australasian Speech Science and Technology Association Inc.

Labuschagne, I. B., & Ciocca, V. (2016). The perception of breathiness: Acoustic correlates and the influences of methodological factors. *Acoustical Science and Technology, 37*, 191–201.

Ladefoged, P., & Ladefoged, J. (1980). The ability of listeners to identify voices. *UCLA Working Papers in Phonetics, 49*, 43–51.

Laver, J. (1980). *The Phonetic Description of Voice Quality*. Cambridge University Press.

Lee, Y., Keating, P., & Kreiman, J. (2019). Acoustic voice variation within and between speakers. *Journal of the Acoustical Society of America, 146*, 1568–1579.

Maxfield, L., Palaparthi, A., & Titze, I. (2017). New evidence that nonlinear source-filter coupling affects harmonic intensity and f0 stability during instances of harmonics crossing formants. *Journal of Voice, 31*, 149–156.

McDougall, K. (2004). Speaker-specific formant dynamics: An experiment in Australian English /aɪ/. *International Journal of Speech, Language and the Law, 11*, 103–130.

McDougall, K. (2006). Dynamic features of speech and the characterisation of speakers: Towards a new approach using formant frequencies. *International Journal of Speech, Language and the Law, 13*, 89–126.

Morrison, G. S. (2009). Likelihood-ratio-based forensic speaker comparison using parametric representations of vowel formant trajectories. *Journal of the Acoustical Society of America, 125*, 2387–2397.

Morrison, G. S. (2013). Tutorial on logistic-regression calibration and fusion: Converting a score to a likelihood ratio. *Australian Journal of Forensic Sciences, 45*, 173–197.

Morrison, G. S., Enzinger, E., Hughes, V., Jessen, M., Meuwly, D., Neumann, C., ... Zhang, C. (2021). Consensus on validation of forensic voice comparison. *Science and Justice, 61*, 299–309.

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.

Nolan, F. (1983). *The Phonetic Bases of Speaker Recognition*. Cambridge: Cambridge University Press.

Nolan, F. (2005). Forensic speaker identification and the phonetic description of voice quality. In W. J. Hardcastle & J. Mackenzie Beck (Eds.), *A Figure of Speech* (pp. 385–411). Mahwah, NJ: Lawrence Erlbaum Associates.

Nolan, F., & Grigoras, C. (2005). A case for formant analysis in forensic speaker identification. *International Journal of Speech, Language and the Law, 12*, 143–173.

Nolan, F., & McDougall, de Jong, G. & Hudson, T. (2009). The DyViS database: Style-controlled recordings of 100 homogeneous speakers for forensic phonetic research. *International Journal of Speech, Language and the Law, 16*, 31–57.

Park, S. J., Sigouin, C., Kreiman, J., Keating, P., Guo, J., Yeung, G., Kuo, F-Y. & Alwan, A. (2016). Speaker identity and voice quality: modelling human responses and automatic speaker recognition. *Proceedings of Interspeech*, San Francisco, USA, pp. 1044-1048.

Pigeon, S., Druyts, P., & Verlinde, P. (2000). Applying logistic regression to the fusion of the NIST'99 1-speaker submissions. *Digital Signal Processing, 10*, 237–248.

Reynolds, D. A., Quatieri, T. F., & Dunn, R. B. (2000). Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing, 10*, 19–41.

Robertson, B., Vignaux, G. A., & Berger, C. (2016). *Interpreting Evidence: Evaluating Forensic Science in the Court Room* (2nd edition). Chichester: John Wiley.

Roettger, T. (2019). Researcher degrees of freedom in phonetic research. *Journal of the Association for Laboratory Phonology, 10*, 1–27.

Rose, P. (2007). Forensic speaker discrimination with Australian English vowel acoustics. *Proceedings of the 16th International Congress of Phonetic Sciences*, Saarbrücken, Germany, pp. 1817–1820.

Rose, P. (2010). Bernard's 18 - vowel inventory size and strength of forensic voice comparison evidence. *Proceedings of the 12th Speech Science and Technology Conference*, Melbourne, Australia, pp. 30–33.

Rose, P. (2013). Where the science ends and the law begins: Likelihood ratio-based forensic voice comparison in a $150 million telephone fraud. *International Journal of Speech, Language and the Law, 20*, 277–324.

San Segundo, E., Foulkes, P., French, J. P., Harrison, P., Hughes, V., & Kavanagh, C. (2019). The use of the vocal profile analysis for speaker characterisation: A methodological proposal. *Journal of the International Phonetic Association, 49*, 353–380.

Shue, Y.-L. (2010). *The Voice Source in Speech Production: Data, Analysis and Models*. UCLA. PhD Thesis,.

Sjölander, K. (1997). The Snack Sound Toolkit. Retrieved from https://www.speech.kth.se/snack/.

Skarnitzl, R., & Vankova, J. (2017). Fundamental frequency statistics for male speakers of common Czech. *Philologica, 3*, 7–17.

Snyder, D., Garcia-Romero, D., Sell, G., Povey, D. & Khudanpur, S. (2018). x-vectors: robust DNN embeddings for speaker recognition. *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, Calgary, Canada, pp. 5329–5333.

Titze, I. R. (2008). Nonlinear source-filter coupling in phonation: Theory. *Journal of the Acoustical Society of America, 123*, 2733–2749.

Titze, I. R., Riede, T., & Popolo, P. (2008). Nonlinear source-filter coupling in phonation: Vocal exercises. *Journal of the Acoustical Society of America, 123*, 1902–1915.

Tschäpe, N., Trouvain, J., Bauer, D. & Jessen, M. (2005). Idiosyncratic patterns of filled pauses. Paper presented at annual conference of the *International Association for Forensic Phonetics and Acoustics*, Marrakesh, Morocco.

Wang, B., Hughes, V. & Foulkes, P. (2019a). Effect of score sampling on system stability in likelihood ratio based forensic voice comparison. In S. Calhoun, P. Escudero, M. Tabain & P. Warren (eds.) *Proceedings of the 19th International Congress of Phonetic Sciences,* Melbourne, Australia, pp. 3065-3069. Canberra, Australia: Australasian Speech Science and Technology Association Inc.

Wang, B., Hughes, V., & Foulkes, P. (2019). The effect of speaker sampling in likelihood ratio based forensic voice comparison. *International Journal of Speech, Language and the Law, 26*, 97–120.

Woubie, A., Koivisto, L., & Bäckström, T. (2021). Voice-quality Features for Deep Neural Network Based Speaker Verification Systems. *Proceedings of the 29th European Signal Processing Conference (EUSIPCO)*, Dublin, Ireland, pp. 176-180. IEEE.

Zhang, C., Morrison, G. S. & Thiruvaran, T. (2011). Forensic voice comparison using Chinese /iau/. *Proceedings of the 17th International Congress of Phonetic Sciences*, Hong Kong, China, pp. 2280-2283.