This is a repository copy of *Assessing the effects of accent-mismatched reference population databases on the performance of an automatic speaker recognition system*.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/id/eprint/167493/

Version: Accepted Version

Article:

Watt, Dominic orcid.org/0000-0003-0978-9282, Harrison, Philip orcid.org/0000-0003-2419-2388, Hughes, Vincent orcid.org/0000-0002-4660-979X et al. (4 more authors) (2020) Assessing the effects of accent-mismatched reference population databases on the performance of an automatic speaker recognition system. International Journal of Speech, Language and the Law. pp. 1-34. ISSN: 1748-8885

https://doi.org/10.1558/ijsll.41466

Reuse

Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

Takedown

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk/

Assessing the effects of accent-mismatched reference population databases on the performance of an automatic speaker recognition system

Dominic Watt, Carmen Llamas*, Peter French*†, Almut Braun**, Duncan Robertson*, Philip Harrison*† and Vincent Hughes**
*University of York † JP French Associates **Bundeskriminalamt, Wiesbaden

Abstract

Automatic Speaker Recognition (ASR) systems are designed to provide the user with statistics relating to the similarity of two or more speech samples and to the typicality of those shared features in the wider population. When an ASR system is used as part of a forensic investigation, the user must decide what counts as the appropriate 'wider population' and select a reference database accordingly. While it has generally been held that the voices populating the reference database should be similar in accent to that of the samples under consideration, the degree to which the accents should correspond has not been investigated empirically.

In this article we test the effects on ASR system output of using reference population voice sets that are (a) closely matched in accent to that of questioned sample, (b) more loosely matched, (c) markedly different from the questioned sample, and finally (d) of mixed accent composition. In testing these effects we draw on two speech corpora that are highly similar in terms of content but which represent phonologically divergent varieties of British English: Standard Southern British English vs. North-Eastern English. While the first corpus may be considered internally homogeneous, the second is comprised of speech drawn from the three conurbations of Newcastle, Sunderland and Middlesbrough. Although these are in many ways very similar, they are nevertheless distinguishable from one another phonetically and phonologically in respects that have been well documented in sociophonetic studies of the region. Our results show that when using good-quality, contemporaneous samples, the ASR system is able to successfully separate same- and different-speaker pairs irrespective of the reference data used to assess typicality. As such, in terms of discrimination, the system might be considered insensitive to accent variation in the reference data. However, that is not to say that accent did not play any role in system output. Accent mismatch between the questioned and reference samples produced scores that were more poorly calibrated than those where the accent was closely matched. Further, accent mismatch produced much stronger same-speaker evidence, observed as a rightward shift of the same-speaker curves in the Tippett plots we use to visualise our results, owing to the fact that same-speaker samples appear more unusual relative to a population of speakers of a different accent.[1]

1. Introduction

Interest in the viability of using Automatic Speaker Recognition (ASR) software systems in forensic speaker comparison casework has grown significantly in recent years (Hughes, Harrison, Foulkes, French, Kavanagh, and San Segundo, 2018; Gold and French, 2019). Its uptake is dependent upon the availability of appropriate reference databases, since the output is contingent upon, *inter alia*, the system performing an assessment of the typicality of features shared between the test and suspect samples in the 'relevant population' (Aitken and Taroni, 2004). There are both theoretical and practical considerations when assessing typicality in forensic speaker comparison casework. On a theoretical level, the relevant population is best defined as a 'suspect population' (Smith and Charrow, 1975) consisting of people who could conceivably have committed the crime. Accordingly, the relevant population is defined by properties of the offender, and in the context of speaker comparison, this means speakers who sound like the offender. For discussion of who should make this decision, see Hughes and Rhodes (2018). The ideal scenario, it can be argued, is one in which we have corpora of recordings at our disposal that are closely matched to the accent(s) to be heard in the samples being compared, and also as similar as possible in terms of other factors such as recording channel and speaking style (Morrison, 2012; Enzinger and Morrison, 2017; Enzinger, Morrison and Ochoa, 2016). However, in practice these things are often not possible. In many cases the available reference databases are small, dated, fragmentary, or composed of material that is otherwise inappropriate (e.g. technically, or in terms of speaking style, etc.), thereby potentially compromising the quality of our ASR-based analysis.

The extent to which the validity and reliability of a system and the resulting strength of evidence in a case are affected by the characteristics of matched and mismatched reference databases is currently the focus of investigation by a number of research groups (e.g. Enzinger and Morrison, 2017; Hughes and Foulkes, 2015; van der Vloed, Jessen and Gfroerer, 2017; Hughes et al., 2018). It is clear that the results reported by ASR systems will be sensitive to the nature of the reference data used including sample duration, database size, and channel characteristics. The degree to which ASR systems are sensitive to accent variation is an empirical question. However, given the holistic way in which abstract features are extracted from the speech signal and subsequently modelled, there is reason to predict that ASR systems could be robust to accent and even language variation. A small number of studies have addressed the issue of language variation (Van Leeuwen and Bouten, 2004; Künzel, 2013), although much less research has been conducted to assess the effect of accent variation (Caballero, Mariño and Moreno, 2009; Hughes, 2014). The general picture to emerge from last two of these studies is that, while accent-mismatched reference populations have little effect on system performance, the strength of evidence may be affected quite substantially.

Given the paucity of up-to-date accent corpora, we might wish for practical reasons to use corpora that are not matched in terms of accent to the test sample, or, indeed, to combine two or more existing corpora representing different regional varieties, rather than continually devoting resources to the collection of new reference material, even if the latter task might from time to time be necessary to counteract the constantly changing nature of spoken language.

But how appropriate is it to use reference populations that are mismatched in terms of accent, or to use heterogeneous populations? How much difference does it make to the output of the ASR system? If the difference transpires to be negligible, we might decide to use accent-mismatched corpora or to combine corpora as a matter of routine. Alternatively, if the accent mismatch or heterogeneity of the reference population appears to be linked to a sufficiently marked deterioration in system performance, we ought to advocate the collection of bespoke corpora, perhaps even at the level of individual cases, as has been argued for by Morrison (2018). Obviously, the latter course of action will have major implications in terms of the timeframe that is necessary to complete a case, and in respect of the costs to be borne by the instructing party.

Thus far, the potential consequences of these courses of action have been underexplored empirically. In this paper we report on a study that puts them to the test using two reference databases: the *Dynamic Variability in Speech* (DyViS) database (Nolan, McDougall, de Jong and Hudson, 2009) of recordings of young male speakers of Southern Standard British English (SSBE), and a newly-collected corpus of recordings of speakers from three urban communities in North-East England (Newcastle, Sunderland, Middlesbrough) obtained for the project *The*

Use and Utility of Localised Speech Forms in Determining Identity: Forensic and Sociophonetic Perspectives (TUULS) project (Braun, Llamas, Watt, French and Robertson, 2018). Further details of the two databases are provided in the following sections.

In presenting this research, we maintain an awareness of a divide in *IJSLL* readership between, on the one hand, those principally working in signal processing and related technical areas, and, on the other, those mainly engaged with phonetics. In an attempt to bridge the gap and make the research accessible to both, we explain rather more of the phonetic aspects and considerations of our work and also rather more of the principles of ASR systems and analysis than we would if we were writing exclusively for one or the other readership.

2. The corpora

2.1 DyViS

The DyViS database was created for the ESRC-funded project *Dynamic Variability in Speech: A Forensic Phonetic Study of British English* at the University of Cambridge.[2] It is comprised of audio recordings of the speech of 100 male students aged between 18 and 25 who were attending the University of Cambridge at the time of recording. The recruited speakers were screened prior to the recording sessions to ensure that they spoke with the same accent (Standard Southern British English, SSBE) without any apparent regional influences. The screening was also carried out as a way of avoiding the inclusion of speakers with atypical voice qualities, or unusually high- or low- average pitch (see further Hudson, de Jong, McDougall, Harrison and Nolan, 2007). All speakers were recorded in a range of contexts so as to elicit different speech styles. To achieve this aim, each speaker undertook a number of tasks. In Task 1 – the task from which the recordings in the present study were drawn – speakers adopted the role of 'suspect' in a simulated police interview about an alleged drug offence. Written and graphical prompts displayed on PowerPoint slides visible to the participant were used to obtain partially phonetically-controlled speech under simulated forensic conditions. Speakers were instructed that there were some pieces of information (the names of certain acquaintances, local businesses, street names, a variety of target nouns, etc.) that they were allowed to talk freely about in answer to interview questions, but that other information of relevance to the investigation should not be mentioned to the 'interviewing officer'. After completing the mock interview, participants moved on to a series of further tasks, the recordings from which are not relevant here as they were not used in the present study (but for further detail see Nolan et al., 2009).

2.2 TUULS

As mentioned above, the data from North-East England that we discuss in this paper were collected for a research project entitled *The Use and Utility of Localised Speech Forms in Determining Identity: Forensic and Sociophonetic Perspectives* (TUULS).[3] Recordings of the speech of a total of 120 participants from the three principal urban centres of the North-East of England - Newcastle upon Tyne, Sunderland, and Middlesbrough (see Figure 1) - were made replicating the DyViS Tasks. The TUULS recordings corresponding to DyViS Task 1 were those used here.

[FIGURE 1 NEAR HERE]

Figure 1. Map of North-East England, showing the locations of the Newcastle upon Tyne, Sunderland and Middlesbrough conurbations.

To the greatest possible extent we used the same types of recording equipment as had been used for DyViS, in spite of superior products having become available in the intervening period. Recordings were made on a Marantz PMD670 solid-state recorder (settings 44.1kHz, 16-bit) using a Sennheiser ME64-K6 floor-standing microphone. We simultaneously made backup recordings on a Zoom H4n digital recorder via a DPA4088 directional headset microphone with the capsule placed around 2cm somewhat to the side of the speaker's mouth.

Although all 100 DyViS speakers had been recorded using the same equipment in the same recording booth at the University of Cambridge's Department of Linguistics, for practical reasons it was necessary to record the TUULS informants in four different recording facilities. All were commercial studios constructed to meet industry-standard criteria in terms of isolation from external noise sources, electrical mains hum, etc., and internally sound-treated to reduce echoes. Owing to the fact that each recording studio booth had different dimensions and furnishings, the acoustic response characteristics of each room would inevitably be different at some level. However, we found the differences to be perceptually imperceptible, and, on testing (see Section 7.5 below), they were found not to have any significant impact on ASR system output.

3. Speakers and preparation of samples

The TUULS corpus contains speech from equal numbers of men and women, but because only male SSBE speakers are represented in the DyViS database, for our present study we made use only of the recordings of the male TUULS speakers. This brought the total number of source recordings to 160 (60 TUULS speakers, i.e. 10 younger (18-25) and 10 older (40-65) males from each of our three fieldwork sites, plus the full set of 100 DyViS speakers).

Segments 120 seconds in duration were extracted from each of the 160 TUULS and DyViS recordings, and then split into two sound files each 60 seconds in length, yielding a total of 320 individual samples. In this way, the comparisons are not entirely forensically realistic, since casework samples are typically recorded non-contemporaneously with a time gap between the questioned sample (QS) and known sample (KS) that could amount to weeks, months, or even years. In casework, there may also be other sources of variability across the two samples, such as style and interlocutor. While this means that the LRs produced in the present study are likely to be inflated relative to the LRs that might be generated in a real case, there is no reason to suppose that the general trends and directions of results would differ, and in any case following this approach was the only feasible means of testing our research questions.

The DyViS and TUULS sound files were downsampled according to the requirements of the Nuance Forensics system, the version of which we used for the current experiment demanding that the audio files for ingest be at 8kHz sampling rate, 16-bit resolution and in PCM .wav format. Despite the downsampling, the audio samples were otherwise of very good quality. Again, while this is likely to result in over-inflation of LRs and over-optimistic system performance relative to real casework data, the above comments apply as before.

4. DyViS versus TUULS: How the varieties differ linguistically

SSBE and North-Eastern English (NEE) generally are widely divergent phonologically. The differences mainly, though not exclusively, are in respect of vowels rather than consonants.[4] Some major disparities are summarised in Table 1. The features noted here have been cited in previous studies (cf. Beal, Burbano-Elizondo and Llamas, 2012) and are also apparent in the patterns observable in the DyViS and TUULS corpora.

Table 1. Example phonological variables distinguishing SSBE from a generalised form of North-Eastern English (NEE).

Variable	**Example words**	**SSBE**	**NEE**
BATH broadening	*bath, glass*	[bɑːθ], [glɑːs]	[baθ], [glas]
FOOT ~ STRUT split	*book, buck*	[bʊk], [bʌk]	[bʊk], [bʊk]
FACE and GOAT	*say, so*	[seɪ], [səʊ]	[sɛ̞ː], [sɔ̞ː]
horsEs	*washes, watches*	[wɒʃɪz], [wɒtʃɪz]	[wɒʃəz], [wɒtʃəz]
non-prevocalic /l/	*feel, milk*	[fiːl^{ɣ}], [mɪl^{ɣ}k]	[fiːl], [mɪlk]

However, while North-Eastern varieties of English share a range of close similarities, within the North-East region we can nevertheless identify features that differentiate Newcastle, Sunderland and Middlesbrough accents from one another. Some salient distinguishing characteristics of each, again noted both in Beal, Burbano-Elizondo and Llamas (2012) and in the TUULS recordings, are shown in Table 2. It will be seen that in some respects Newcastle and Sunderland are more similar to one another than either is to Middlesbrough, and that in other respects Sunderland and Middlesbrough resemble each other more closely than either resembles Newcastle.

Table 2. Example phonological variables with pronunciations differing between Newcastle, Sunderland and Middlesbrough.

Variable	**Newcastle**	**Sunderland**	**Middlesbrough**
FACE	[fɪəs], [feːs]	[fɪəs], [feːs]	[fɛːs]
GOAT	[gʊət], [goːt]	[gʊət], [goːt]	[gɔːt], [gɔ̈ːt]
MOUTH	[mɛʊθ], [muθ]	[mɛʊθ]	[mauθ]
START	[stɒːt], [stɑːt]	[stɑːt]	[stɑːt], [staːt]
NURSE	[nɔːs], [nɜːs], [nøːs]	[nɜːs]	[nɛːs], [nɜːs]
commA, letter	[kɒmɐ(ː)], [lɛʔ͡tɐ(ː)]	[kɒmə], [lɛʔ͡tə]	[kɒmə] / [kɒmɛ], [lɛʔə] / [lɛʔɛ]

5. Why might accent differences be relevant to ASR system analyses?

Having outlined some of the differences between the accents being tested, we now turn to the question of why these might have a bearing on ASR system performance and conclusions. In order to address this question, it is helpful first to consider the analytic principles of ASR systems and the procedures deployed in arriving at their conclusions.

5.1 Working principles of ASR systems

Once inputted to an ASR system, the recorded speech signal is sampled at short intervals (overlapping windows of e.g. 20 milliseconds duration), and the acoustic spectrum of each sample is modelled using a set of numerical values (predominantly mel-frequency cepstral coefficients, MFCCs) that capture the broad characteristics of the shape of the spectral envelope at that point in the signal, these being reflective of vocal tract dimensions and configuration (see Davis and Mermelstein, 1980 or Hansen and Hasan, 2015 for a more detailed and comprehensive account of MFCC derivation).

There are a number of commercially-available ASR systems, each of which functions in a slightly different way from the others. The system used for the current study was Nuance Forensics v.11.1, the output of which provides a number of different metrics relating to the similarity of speaker samples and their typicality vis-à-vis the pooled characteristics of a chosen reference population. Key among these is the *score*, a measure of the raw similarity of two samples that does not take account of their typicality. We can think of the score essentially as a measure of the similarity of the two speaker models (from the QS and KS), such that similar voices will produce a higher score while dissimilar voices will result in a smaller or negative score. The score, therefore, does not triangulate the similarity score against speaker models abstracted from a reference population. That task is done via the computation of the Likelihood Ratio (LR), an index of the similarity and typicality of the samples expressed as a single value. The process of converting a score to a LR is also known as calibration.

More formally, the LR is the ratio between the estimates of the similarity and typicality of the samples, which are respectively the numerator and the denominator of the equation shown in (1). The value of the denominator is derived from the reference population.

(1)

$$LR = \frac{p(E|H_{SS})}{p(E|H_{DS})}$$

According to (1), we divide the probability *p* of obtaining the evidence E under the assumption that the same-speaker (prosecution) hypothesis H_{ss} is correct by the probability of obtaining E given the different-speaker (defence) hypothesis H_{ds}. The LR is a measure of the strength of the evidence.

In the context of ASR, the score is the evidence (E). The LR then answers a specific question: are we more likely to get the score for the QS and KS comparison assuming it is a same-speaker score, or a different-speaker score? In order to do this, it is necessary to have distributions of same-speaker scores and different-speaker scores against which to evaluate. In order to estimate the numerator, it may be possible to generate a case-specific distribution consisting of scores for the KS against other samples of the known speaker (e.g. other bugged telephone calls or police interviews where the identity of the speaker is known). Otherwise, we can use a more generic same-speaker distribution based on scores generated from lots of same-speaker comparisons from a larger database (Solewicz, Jessen and van der Vloed, 2017). The different-speaker distribution relies on the reference population. It is generated from scores computed by comparing each speaker in the reference population against the questioned sample (Jessen, Meir and Solewicz 2019).

[FIGURE 2 NEAR HERE]

Figure 2: Hypothetical same- (blue) and different-speaker (red) score distributions produced by an automatic system, along with a score (2.5, denoted by the green vertical line) for the specific case comparison generating the two probabilities required to compute a Likelihood Ratio.

If the two probabilities are the same, LR will equal 1, and thus it will lend no support to either H_{ss} or H_{ds}. Values greater than 1 support the prosecution hypothesis H_{ss}, while values of less than 1 support the defence hypothesis H_{ds}. Because raw LR values can on occasion be very large or very small, the LR is generally expressed as its base-10 logarithmic equivalent, $\log_{10}$LR or just LLR. An LR of 1 is then equivalent to an LLR of 0, an LR of 10,000 = LLR 4, and so on. Figure 2 shows a hypothetical example of score-to-LR conversion within an ASR system. In this case, the score for the specific case comparison is 2.5. The probability of a score of 2.5 assuming it is a same-speaker score is 0.1295, while the probability of the score assuming it is a different-speaker score is 0.0175. Therefore, the LR in this case is 0.1295/0.0175, which is 7.389 (or 0.869, expressed on a $\log_{10}$ scale).

Tippett plots (Tippett, Emerson, Fereday, Lawton and Lampert, 1968; Evett, Lambert and Buckleton, 1995) represent the cumulative proportions of LLR values for multiple same- and different-speaker pairs on a single graph centred on LLR = 0 (for examples, see Figures 4 - 6 in Section 7 below). If the ASR system makes no ‘errors’ (i.e. produces no contrary-to-fact evidence), all the same-speaker pairs will have positive LLRs and all the different-speaker pairs negative LLRs, and the curves representing the cumulative distributions of each will not overlap. In reality, it is probable in the majority of cases that the system will not perform flawlessly, and that it will classify pairs of different-speaker samples as same-speaker samples (Type I errors, i.e. false positives, false alarms), and vice versa (Type II errors, i.e. false negatives, misses). Where the same- and different-speaker curves cross - that is, where the percentage of misses is the same as that for false alarms - we talk of the Equal Error Rate (EER). Low EERs therefore signal that the system is functioning well but not perfectly, while a zero EER shows that it is making no errors of one of the two kinds, or of both kinds. EER is a threshold-independent validity metric which deals simply with the number of ‘errors’; it assesses the separation of the same- and different-speaker LLRs irrespective of where the threshold is. This means that a system can produce an EER of 0%, even when relative to the LLR = 0 threshold the system *does* produce ‘errors’. A gap between the crossover in LLRs and the 0 threshold indicates a badly-calibrated system, perhaps due to the use of a statistical model that is inappropriate for the data (for instance, if using a normal distribution for non-normal data) or where the data used to train the system and the evidential data are mismatched with regard to certain factors (e.g. channel, quality, duration).

However, the EER does not provide any information relating to the magnitude of the ‘errors’ made; high-magnitude ‘errors’ are much more problematic for a system in the context of forensic evidence that may lead to a miscarriage of justice, compared with low-magnitude ‘errors’ that will not have much bearing on the overall decision made by the court. Therefore, the C_{llr} (‘log-Likelihood Ratio cost’; Brümmer and du Preez 2006) function is used. Ideally, the system should report no LLRs that support the counterfactual proposition, but if these do occur a penalty is applied to each in proportion to the magnitude of the counterfactual LLR, with larger positive or negative LLRs being weighted more heavily. C_{llr} can be decomposed into two types of ‘error’. The first is *discrimination error* ($C_{\text{llr min}}$). This is the lowest C_{llr} that the system could produce with a given set of data assuming that the scores were perfectly calibrated. It provides insight into the extent to which the system can successfully discriminate between pairs of same- and different-speaker samples. The second is calibration error ($C_{\mathrm{llr\ cal}}$). This is the error introduced into the system due to poor calibration (displayed on a Tippett plot, the further the intersection of same- and different-speaker scores is from the central line (the point of no support for either side, i.e. where the LLR is equal to 0), the poorer the calibration of the system; see below). The C_{llr} is the sum of the $C_{\mathrm{llr\ min}}$ and $C_{\mathrm{llr\ cal}}$, and as with the EER, the larger the C_{llr} the poorer the system performance.

5.2 Revisiting the accent question

Having considered the working principles of ASR systems, we now return to the question of whether and how the accent make-up of the reference population might affect an ASR's performance and output. It was mentioned at the outset of the previous section that the MFCCs that the systems produce by processing the signal are derived from the spectral envelope and are therefore related to the geometry of the speaker's vocal tract. This being so, there is one argument - the 'biological' view - that the analysis might be unaffected by the accent, or indeed the language, of the reference population relative to that found in the speech samples being analysed, and that the systems build their models of individual talkers on the basis of abstracted features which may correspond only indirectly to the phonetic features. That is, the algorithms are targeted at identifying the frequencies and amplitudes of the resonances of the vocal tract that produced the speech signal, and are either unaffected by, or only marginally affected by, temporally localised events such as the pronunciations of individual vowels and consonants. The models are more closely dependent upon the shapes and sizes of the resonating cavities within the speaker's vocal tract. Since we currently know of no consistent anatomical differences between NEE speakers and SSBE speakers, we have no reason to suppose that an ASR system would distinguish one set of speakers from the other purely on these grounds. On this biological view, the system would not classify speakers by virtue of the learned segmental, or indeed other, characteristics of their regional or local accents, but would instead group them according to their anatomical and/or physiological similarities.

On a second, contrasting view - the 'linguistic' view - however, it might be held that the long-term shape and configuration of a speaker's vocal tract is not only a function of his/her individual anatomy, it is also determined, or at least affected, by the accumulation of localised lingual gestures, lip positions and orientations of other vocal tract organs associated with the production of vowels and consonants. If the vowels and consonants are distributed and realised differently across accents, as we have seen with SSBE versus North-Eastern English, and (albeit to much a much lesser extent) within the different sub-regional varieties of the latter, then these accent-based differences in ASR system input will be reflected in system output.

Further, leaving aside the individual segmental features associated with accents and focussing on the long-term vocal tract settings that give rise to supralaryngeal voice quality, it would appear that biological differences between speakers can be 'neutralised', or exacerbated and over-ridden, by the results of linguistic socialisation (French and Stevens 2013). It is widely known that different languages are characterised by different vocal tract settings (e.g. French has a high degree of nasalisation and lip rounding; Russian has pervasive palatalisation). However, *within* languages different accents may have their own distinctive settings (see French and Stevens 2013; Dellwo, French and He, 2019; Wormald, 2016 for some general summaries in respect of English accents). Thus, among speakers there may be strong within-accent convergence and strong between-accent divergence, these being the products of linguistic socialisation. While we are not presently able to make any substantiated claims about the accents of the North-East having their own distinctive supralaryngeal voice qualities, there is nevertheless evidence of SSBE being characterised by distinctive clusters of vocal tract features (Stevens and French, 2013; San Segundo, Foulkes, French, Harrison, Hughes and Kavanagh, 2019), and we have not been able to observe these with any great frequency of occurrence in the TUULS data.

If one takes the linguistic as opposed to the biological view, that there is likely to be convergence within and divergence across accent communities in respect of features relevant to ASR system performance, it follows therefore that by varying the accent complexion of the reference population there will be consequences for the LR values the technology produces. For example, if one were to use a reference population that was mismatched with that in the recordings under examination, one might expect to find a high number of Type I errors (false alarms) in circumstances where the suspect and questioned samples came from different speakers just because they shared the same accent which was different from, and atypical of, that represented in the reference population. Also, where the suspect and questioned samples do happen to be from the same speaker, artificially inflated LR scores might be expected just because the features found in them would set them apart from the reference population.

In the following sections we put these views to the test by applying ASR analysis to assess the effects of changing the nature of the reference population when running comparisons on speaker samples drawn from the TUULS corpus. The questions we address are as follows:

1. How large are the changes in our two measures of ASR system performance - viz., the Equal Error Rate (EER) and LLR cost function (C_{llr}) - when the reference population is modified according to the regional accent(s) represented in the database used for this purpose?
2. Are better results achieved where the regional accent represented in the reference population is matched closely to that of the speakers in the samples being compared?
3. Where reference populations that are unmatched for accent are used, is system performance better if the reference accents are phonologically similar to the accent of the speakers under comparison?
4. If only small numbers of recordings of the accent in question are available, is it valid to pool dissimilar reference population databases in the interests of increasing the size of the overall reference population?
5. If system performance improves under any of the above conditions, how much improvement should we deem satisfactory?

6. Method

6.1 ASR analysis

While the version of Nuance Forensics used in this study is an iVector system and cannot strictly be classed as an exemplar of the emerging generation of xVector ASR systems, it nevertheless overlaps with these newer systems in many significant respects, and shares with them the use of deep neural nets (DNNs) for MFCC modelling (see Snyder, Garcia-Romero, McCree, Sell, Povey and Khudanpur (2018) for an explanation of the distinctions).

The comparison of speakers was undertaken using the batch processing mode of Nuance Forensics with a default configuration. The batch processing mode requires that a single reference population is used to calculate the LRs for all the comparisons in a batch. Rather than run a very large number of batch processes with different reference populations, a set of scores, which are not influenced by the choice of reference population, were obtained in a single execution of the batch processing mode. These values are reported in Section 7.1 below. The scores were then converted to LLRs in MATLAB using the reference populations described in Sections 7.2 to 7.4, following exactly the same procedures as those performed by Nuance Forensics. This process allowed us much greater control over the composition of the reference populations which were generated on a per comparison basis due to the need to use a 'leave-one-out' approach. Same-speaker (SS) and different-speaker (DS) score distributions were modelled with normal distributions. Since only one recording per speaker was available, the SS distribution was defined based on the default parameters ($N(\mu, \sigma) = N(4,1)$) in Nuance. In this way, the SS score distribution was the same for all comparisons. The parameters for the SS score distribution in Nuance are based on a large number of SS comparisons performed using recordings from the NIST evaluations (specifically, evaluations SRE08 and SRE10; see www.nist.gov/itl/iad/mig/speaker-recognition). These recordings are often of considerably lower quality than those in the DyViS and TUULS databases and are generally recorded via telephone transmission. This mismatch is likely to produce poorly calibrated LLRs in our study. However, in the absence of multiple suspect recordings in real casework, users may have to rely on the default parameters in the system. The DS distribution changed for each comparison, and was based on the scores for comparison between every speaker in the custom reference population and the questioned test speaker.

7. Results

7.1 Score data

As mentioned in Section 3, the 120-second recordings for the TUULS and DyViS speakers (N=160) were divided into two 60-second halves, half A and half B. Our comparisons were based on comparing each speaker's A sample against all of the B samples, including that speaker's own B sample. 25,600 comparisons in total, i.e. 160×160, were therefore carried out The resulting scores for all 25,600 comparisons, represented as a heatmap, are shown in Figure 3.

[FIGURE 3 NEAR HERE]

Figure 3. Heatmap representing scores for same- and different-speaker comparisons between all combinations of TUULS and DyViS speakers (N = 25,600).

The first 100 rows of the heatmap in Figure 3 represent the A samples for the DyViS speakers, the remaining 60 those for the TUULS speakers. The columns represent the 160 B samples. Colours towards the white/yellow end of the spectrum indicate greater similarity between speaker sample pairs. The diagonal line represents the 160 same-speaker comparisons. It is important to note that the cells on the diagonal are all shaded white in the plot. This is because they represent the comparisons between two 60-second halves (A and B) of the same recording from the same speaker. Thus, they produce exceptionally high scores. While there is of course some variation in these high scores, the colour spectrum in the plot is thresholded such that the same-speaker scores all appear white, so as to preserve the visible differences among the different-speaker scores. It can be seen that the within-DyViS comparisons (top left) generate higher scores (i.e. a greater proportion of white, yellow and orange cells) than the within-TUULS comparisons at bottom right. This is a result of the fact that as a group the DyViS speakers are more similar to one another than the TUULS speakers are to other TUULS speakers. Such a finding is predictable, given that the DyViS speakers are all approximately the same age, and that they were chosen specifically to sound alike. Much greater heterogeneity is evident in the TUULS quadrants of the figure, which is again unsurprising in view of the fact that this half of the TUULS corpus is composed of men in two age groups, who come from three different urban areas. The TUULS speakers are not grouped here by age, but they are ordered by place of origin, the ordering being Middlesbrough, Newcastle and Sunderland, from left to right. There is, however, no clear patterning within these results based on the origin of the speakers. Furthermore, because we are primarily concerned with the effects of differences in regional and local accent on the results of our ASR-based trials, we will not consider speaker age any further in this paper. Age as a speaker variable was not incorporated into the design of the DyViS corpus, so we cannot run any inter-corpus comparisons that take speaker age into account. While we acknowledge the desirability of breaking the TUULS sample down further into two separate age groups (Hughes and Foulkes, 2015), this step would result in speaker groups of only 10 speakers, which would be too small to produce satisfactorily reliable results.

7.2 Log-Likelihood Ratio (LLR) data

We present the LLR data in the form of Tippett plots (Figures 4-6; see Meuwly, 2001). These allow us to visualise the effects of changing the reference population database when running same- and different-speaker comparisons on the 60 TUULS speakers. It is worth noting at this point that a well-calibrated system will produce SS and DS LLRs that intersect at the centre line (where LLR=0). The three Tippett plots below display LLRs that are shifted rightwards relative to the centre line (i.e. miscalibrated), with no contrary-to-fact SS LLRs, but a relatively high proportion of contrary-to-fact DS LLRs. This is likely due to the differences between the experimental materials we are using from DyViS and TUULS and the data used to train the system (as discussed in §4.3), although accent mismatch may introduce further calibration error. In this section we report both $C_{\mathrm{llr\ min}}$ and $C_{\mathrm{llr\ cal}}$ values. We also discuss the magnitude of the LLRs, but only in relative terms according to the different reference data used (i.e. ignoring the centre line threshold). This is because we know that the LLRs are overinflated due to the use of high-quality, contemporaneous samples that are not reflective of real casework conditions.

Figure 4 shows three pairs of curves representing three different reference population datasets. In each curve pair, the left-hand curve represents the cumulative LLRs for different-speaker comparisons (N = 3,540, i.e. 60×59), while the right-hand curve represents the same-speaker comparisons (N = 60). The comparisons were carried out using the 'leave-one-out' method, whereby the samples for the target speaker(s) are removed from the reference population during the comparison so as to avoid the effects of self-referentiality. TUULS is represented as a blue line in the plot, DyViS as a dashed orange line, and the two databases combined (TUULS&DyViS) as a dash-dot yellow line.

[FIGURE 4 NEAR HERE]

Figure 4. Tippett plot of same- and different-speaker comparisons of TUULS speaker samples using three reference populations: TUULS (solid blue line), DyViS (dashed orange line), and TUULS&DyViS combined (dash-dot yellow line).

It can immediately be seen that for all three pairs of curves there is a clear separation between the different-speaker comparisons at left, and the same-speaker comparisons at right. Since the curves in each pair do not overlap, the Equal Error Rate (EER) in each case is zero. This is possibly an outcome of the high-quality and contemporaneous recordings used, as suggested above. The ASR system is apparently performing well here, in that no speaker in the 60 same-speaker pairs is being mistaken for someone other than himself. As discussed in Section 3, the use of the C_{llr} function gives us a single measure of overall system performance by assigning a weighted penalty to the counterfactual LLRs according to their distance above and below zero. The C_{llr} values for the data displayed in Figure 4 are shown in Table 3.

As with EER, since the curves do not overlap, discrimination is perfect and so the $C_{\mathrm{llr\ min}}$ in all cases is zero - again, the comments concerning studio quality and contemporaneity apply. As such, with these data, changing the reference population has no effect on the discrimination performance of the system; using samples of this type makes the task of discrimination relatively straightforward for the system, irrespective of the choice of reference population. Differences were found across the different tests in terms of $C_{\mathrm{llr\ cal}}$, with DyViS producing the highest calibration error, followed by TUULS&DyViS combined, and finally the matched TUULS population. Since there are no counterfactual SS LLRs, the C_{llr} and $C_{\mathrm{llr\ cal}}$ values are only a result of the counterfactual DS LLRs.

Table 3. C_{llr} values for same- and different-speaker comparisons of TUULS speaker samples using three reference population databases: TUULS, DyViS, and TUULS&DyViS combined.

Reference population	$C_{\mathrm{llr\ min}}$	$C_{\mathrm{llr\ cal}}$	C_{llr}
TUULS	0	0.145	0.145
DyViS	0	0.427	0.427
TUULS&DyViS	0	0.218	0.218

We now turn to the magnitude of the LLRs produced using each reference population. The biggest differences are found in the SS LLRs. The strongest LLR values are yielded when the DyViS corpus is used as the reference population. The LLRs are substantially higher, by an average of two orders of magnitude comparing median SS LLRs, than the LLRs attained when the TUULS and TUULS&DyViS reference populations are used. There is very little difference between the magnitude of the SS LLRs for TUULS and TUULS&DyViS, such that the median SS LLR for both is within the same order of magnitude. Looking next at the LLRs for the different-speaker comparisons in Figure 4, there is relatively little difference in the curves for the different reference populations. The DS LLRs are marginally weaker on average when using the DyViS reference population. The effect of adding DyViS to TUULS so as to expand the reference population database is evidently not a large one. Relative to the centre line threshold, there is a higher proportion of contrary-to-fact DS LLRs when using DyViS. However, as outlined above, this is due to calibration error rather than an issue with discrimination.

The implications of these results are discussed further in Section 8. Before turning to them, we consider the effects (a) of varying the size of the DyViS database and (b) splitting the TUULS database into its three constituent subcorpora (Newcastle, Sunderland, Middlesbrough).

7.3 Effects of varying the DyViS database in size

The DyViS database is two-thirds larger again than the TUULS database, and so rolling the two corpora together creates an imbalance in terms of the representation of one accent (SSBE) versus the other (NEE). In the interests of achieving parity between the two accents, the DyViS database was reduced to the first 60 speakers, to match it with the size of TUULS. The results of the comparisons run using the reduced DyViS database combined with TUULS (TUULS&60DyViS) are shown in Figure 5. The TUULS, DyViS and TUULS&DyViS results are the same as those reported in §7.2.

[FIGURE 5 NEAR HERE]

Figure 5. Tippett plot of same- and different-speaker comparisons of TUULS speaker samples using four reference population databases: TUULS (solid blue line), DyViS (dashed orange line), TUULS&DyViS combined (dash-dot yellow line) and TUULS&60DyViS combined (dotted purple line).

Figure 5 illustrates very clearly that reducing the size of the DyViS database by 40 speakers has a marked effect upon the same-speaker LLRs, which are smaller, on average, by one order of magnitude compared with those for TUULS&DyViS. While combining TUULS with the full set of 100 DyViS speaker models makes little difference relative to the LLR values for TUULS on its own, we see a reduction in the magnitude of SS LLRs when the two databases are evenly matched in terms of the number of speaker models they contain. By contrast with the same-speaker comparisons, there is practically no effect on the LLRs for the DS comparisons when the databases are matched for size in this way.

The C_{llr} values for the combined TUULS&60DyViS reference population is shown in Table 4. Again, the differences here are all attributable to calibration error rather than discrimination error, since the TUULS&60DyViS population still produces completely separated SS and DS LLR distributions. Overall, C_{llr} drops to some extent from 0.218 to 0.198 after the DyViS database is reduced to 60 speakers from the full set of 100. This change is a consequence only of the counterfactual DS LLRs.

Table 4. C_{llr} values for same- and different-speaker comparisons of TUULS speaker samples using four reference population databases: TUULS, DyViS, TUULS&DyViS combined, TUULS&60DyViS combined.

Reference population	$C_{\mathrm{llr\ min}}$	$C_{\mathrm{llr\ cal}}$	C_{llr}
TUULS	0	0.145	0.145
DyViS	0	0.427	0.427
TUULS&DyViS	0	0.218	0.218
TUULS&60DyViS	0	0.198	0.198

7.4 *Effects of splitting TUULS database into subcorpora*

Figure 6 displays the differences between the outcomes of the SS and DS trials run on the TUULS corpus when the reference population is varied according to the place of origin of the TUULS speakers represented in the reference population database (Newcastle, Sunderland or Middlesbrough). The DyViS and TUULS results are the same as those presented above.

[FIGURE 6 NEAR HERE]

Figure 6. Tippett plot of same- and different-speaker comparisons of TUULS speaker samples using five reference population databases: the entire TUULS corpus (solid blue line), DyViS (dashed orange line), and the three TUULS subcorpora: Middlesbrough (dash-dot green line), Newcastle (dotted light blue line), and Sunderland (solid burgundy line).

The general pattern for SS comparisons in Figure 6 indicates that splitting the TUULS corpus into three has little overall effect on the LLR scores, though the slopes of the curves for each individual subcorpus are shallower than that for TUULS as a single corpus, indicating greater variability in LLRs. The LLRs are smallest for Newcastle and highest for Sunderland, with those for Middlesbrough falling in between. However, on average the SS and DS LLRs for each of the subcorpora are within the same order of magnitude. It must be remembered that each subcorpus is comprised of only 20 speaker models, however, and that the speakers in each are a mix of equal numbers of older and younger males. Hence, variability of the sort of we observe in Figure 6 is to be expected. While 20 speakers is below the minimum number of 30 required by Nuance Forensics to form a reference population, studies by Hughes (2014) and Kinoshita and Ishihara (2014) have shown that such small sets do not in fact lead to erratic results.

The C_{llr} data shown in Table 5 indicate that the division of the TUULS corpus into the three subcorpora has only a small negative effect by comparison with the value for the TUULS corpus as a whole, with the C_{llr} values for each subcorpus being fairly tightly clustered relative to one another. Those for Sunderland and Middlesbrough are especially close. All four TUULS corpora yield C_{llr} values which are some distance from that for the DyViS database. This might suggest that the key factor here is not the extent to which the individual subcorpora are a good match accent-wise to the TUULS corpus as a whole, but rather the size of the reference population. If the latter is reduced in size by two-thirds (i.e. 60 → 20) the C_{llr} value will increase, but not by a very large margin, and nothing like by as large a margin as is the case where DyViS is used, in spite of DyViS containing five times as many speaker models.

Table 5. C_{llr} statistics for same- and different-speaker comparisons of TUULS speaker samples using five reference population databases: TUULS, DyViS, Newcastle, Sunderland, Middlesbrough.

Reference population	$C_{\mathrm{llr\ min}}$	$C_{\mathrm{llr\ cal}}$	C_{llr}
TUULS	0	0.145	0.145
DyViS	0	0.427	0.427
Newcastle	0	0.159	0.159
Sunderland	0	0.153	0.153
Middlesbrough	0	0.154	0.154

[FIGURE 7 NEAR HERE]

Figure 7. Tippett plot of same- and different-speaker comparisons of Middlesbrough speaker samples using four reference population databases: the entire TUULS corpus (solid blue line), and the three TUULS subcorpora (Middlesbrough; dash-dot green line), Newcastle (dotted light blue line) and Sunderland (solid burgundy line).

As a final test of the effects of varying the reference population on the outcomes of the ASR comparisons, we restricted the speakers selected for the SS and DS pairs to just those from Middlesbrough, while keeping the (TUULS) reference populations as they were for the tests which yielded the results shown in Figure 6 and Table 5 (viz., TUULS, Newcastle, Sunderland, Middlesbrough, but not DyViS). We might predict that when comparing Middlesbrough speakers with one another the ASR would perform best if a corpus of recordings of

Middlesbrough English speakers were used as the reference population. As Figure 7 reveals, however, this is not necessarily the case. Changing the reference population results in no obvious difference to the DS comparison results, while for the SS comparisons the use of the Middlesbrough reference population does not produce LLRs which are a great improvement on those generated when using a reference population composed just of Newcastle English speakers. The use of the pooled TUULS reference population results in LLR values which are approximately the same as those for the Middlesbrough-only reference corpus; the most marked effect brought about by changing the composition of the reference population is that produced when the Sunderland-only speaker corpus is used.

Table 6 shows the corresponding C_{llr} data. It can be seen that while the Middlesbrough reference population results in lower scores than those when the Newcastle or Sunderland populations are used, the difference (at least between Middlesbrough and Sunderland) is not a large one, and overall the pooled TUULS corpus performs most strongly. It is worth remembering, of course, that the TUULS corpus is three times the size of the Middlesbrough corpus. Again, the C_{llr} results are derived from just the counterfactual DS LLRs.

Table 6. C_{llr} statistics for same- and different-speaker comparisons of Middlesbrough speaker samples using four reference population databases: TUULS, Newcastle, Sunderland, Middlesbrough.

Reference population	**$C_{\mathrm{llr\ min}}$**	**$C_{\mathrm{llr\ cal}}$**	**C_{llr}**
TUULS	0	0.188	0.188
Middlesbrough	0	0.190	0.190
Newcastle	0	0.218	0.218
Sunderland	0	0.197	0.197

Before we turn to a discussion of the implications of the data presented in this section in terms of our research questions, we must consider the potential effects that the different recording facilities used for the collection of the TUULS corpus may have had on the figures reported above.

7.5 Channel effects

To check whether any channel effects from our different studios might be sufficiently large that they could influence the outcome of the speaker comparisons – e.g. by biasing the ASR software towards determining speaker sample pairs to be more linguistically similar than they in fact were, just because they had been recorded in the same facility – we ran comparisons of non-speech excerpts from each (entire) original recording. Silent portions in each recording (≥ 100ms duration; silence threshold 0.3, rather than the less conservative default threshold of 0.7) were identified and concatenated using the *vadsohn* voice activity detection function in MATLAB (Sohn, Kim and Sung, 1999). The scores for the same- and different-origin pairs of silent-portion files were compared to those reported by Nuance Forensics for the speech-active portions of the corresponding recordings. No correlation between the scores from the non-speech samples and the Nuance Forensics speaker scores was apparent. That is, we have no evidence that high-scoring pairs according to the Nuance scores are also very similar in terms of their silent sections. In summary, the data suggest that any channel differences do not have any significant impact on the ASR output.

8. Discussion

Let us review the research questions (RQs) that were listed in Section 5.2.

1. How large are the changes in our two measures of ASR system performance - viz., the Equal Error Rate (EER) and LLR cost function (C_{llr}) - when the reference population is modified according to the regional accent(s) represented in the database used for this purpose?
2. Are better results achieved where the regional accent represented in the reference population is matched closely to that of the speakers in the samples being compared?
3. Where reference populations that are unmatched for accent are used, is system performance better if the reference accents are phonologically similar to the accent of the speakers under comparison?
4. If only small numbers of recordings of the accent in question are available, is it valid to pool dissimilar reference population databases in the interests of increasing the size of the overall reference population?
5. If system performance improves under any of the above conditions, how much improvement should we deem satisfactory?

In answer to RQ1, we can observe changes which pattern more or less in accordance with our expectations. The heatmap of scores in Figure 3 shows that on average, TUULS speakers are more similar to other TUULS speakers (more yellow cells in the lower right region) than TUULS speakers are to DyViS speakers (more orange cells in the lower left and upper right regions). So, when DyViS speakers are used as a reference population the distributions of different-speaker scores used to calculate the LR reflect the fact that TUULS and DyViS speakers are less similar. The consequence is that, on average, the distributions of different-speaker scores using the DyViS reference population have lower means than the distributions where the reference populations are other TUULS speakers (i.e. where the accent is matched, at least a general level). This means that, by and large, for a same-speaker score the probability of obtaining that score under the defence hypothesis will be lower for a DyViS reference population than for a TUULS reference population. This results in a larger LR when the DyViS reference population is used, since the denominator of the LR equation is smaller. This effect is clearly seen in Figure 4, in which the same-speaker curve for the DyViS reference population (dashed orange line) sits some distance to the right of the TUULS reference population curve. This effect has been termed *right shift*, and has previously been reported by van der Vloed, Jessen and Gfroerer (2017) when using reference populations with a language mismatch to the comparison samples. An effect of this type, albeit a smaller one, among the different-speaker LRs is also revealed in Figure 4 by the rightward displacement of the curve representing the different-speaker comparisons under the DyViS condition. These results are predictable on theoretical grounds, given that where DyViS is used as the reference population for comparisons involving TUULS speakers, each comparison involves a phonologically highly divergent mismatching accent (SSBE) as well as different speakers. We would therefore expect the probability of obtaining a score, whether it be for a same- or different-speaker comparison, to be lower when using a reference population that is as mismatched as DyViS is to the north-eastern English accents represented in the TUULS corpus.

In terms of assessing the performance of an ASR system the above situation is somewhat paradoxical, as one would normally consider higher same-speaker LRs to be an indicator of good performance. However, in the current context the larger LRs are obtained with an accent-mismatched reference population, and we can directly compare these results with those generated using what we know is an accent-matched reference population. The higher same-speaker LRs from the mismatched reference population should therefore be considered as overinflated, or an overestimate of an LR that would be obtained with an appropriately accent-matched reference population.

Given the explanation of the cause of the right shift effect, it might be expected that if a reference population contained a combination of accent-matched and -mismatched speakers then the SS curve on a Tippett plot would be placed between the curves from the reference populations containing just the matched and mismatched speakers. However, this is not the case. It can clearly be seen in Figure 4 that the TUULS&DyViS SS curve falls very close to the TUULS-only curve. In Figure 5, which includes the reference population with a balanced number of accent-matched and -mismatched speakers, there is even a slight leftward shift of the SS curve relative to the accent-matched TUULS reference population.

An examination of the distributions of all the different-speaker scores generated from the three reference populations shows that they pattern as expected. That is, the means of the distributions show that TUULS speakers are most similar to other TUULS speakers, and then to the mixed set of TUULS and DyViS speakers. They are the least similar to just the DyViS speakers. Also, the spread of the distributions varies as expected, with the DyViS reference population having the narrowest spread (reflecting the homogeneous nature of the speakers), followed by TUULS. Finally, and as expected, the greatest spread is found in the mixed set, due to the diversity introduced by combining the two sets of speakers.

It must, however, be remembered that the different-speaker score distribution is different for every questioned sample. This is because the different-speaker score distribution is derived from the scores obtained from comparing the questioned sample with each speaker in the reference population. The effect of changing the different-speaker score distributions can be seen by examining the change in the same-speaker LRs on a per-speaker basis for the different reference populations. These results show that the shift in LR for each speaker is not uniform across the reference populations. In general, the DyViS reference population produces the highest LRs, which leads to the right shift seen in the Tippett plots. But no clear pattern emerges within the LRs for the TUULS-only condition and the TUULS and DyViS reference populations, in terms of either in the size or the direction of the LR shift. The lack of systematicity, and the fact that the Tippett plots show a cumulative representation of the LRs, appears to account for the lack of right shift for the mixed reference population same-speaker results.

The right shift pattern is also present in Figure 7, where the test speakers are from Middlesbrough only. The same-speaker LRs obtained with the accent-mismatched reference populations of Newcastle and Sunderland are shifted to the right compared with the curve from the accent-matched Middlesbrough population. This shift is smaller than the one seen in Figures 4, 5 and 6, when the DyViS speakers were the accent-mismatched reference population. This smaller right shift may be accounted for by a combination of the greater level of phonological similarity among the accents of Middlesbrough, Newcastle and Sunderland relative to SSBE, and the more closely aligned channel characteristics of the TUULS recordings.

Despite differences in the magnitudes of LLRs, particularly for SS comparisons, $C_{\text{llr min}}$ and EER were consistently 0 when using DyViS, TUULS, or a mixture of the two (irrespective of the number of speakers from each corpus). This shows that discrimination error is unaffected by changes to the reference population. It provides some evidence to suggest that the overall performance of the ASR is insensitive to variation in the accent used for the reference population. However, it is also likely to be a result of the fact that high-quality, contemporaneous samples were used in the present study. The task is therefore a relatively straightforward onse for the system, such that we have a floor effect whereby it is not possible to further improve the discrimination error. If we made use of lower-quality materials that are more challenging for the system, it might be that discrimination error differences would emerge as a function of the reference population used. As expected, the differences between reference populations were manifested in $C_{\text{llr cal}}$, and it is here that we see some evidence of accent sensitivity. The poorest calibration was found when using DyViS (0.427). Calibration error is reduced when combining DyViS and TUULS, and reduced again when using TUULS by itself. In this sense, the ASR system performs better with reference material that is matched for accent, even if that match is somewhat approximate (cf. the differences between the three North-East varieties listed in Table 2). The above remarks can also be taken in answer to RQs 2 and 3, in that substantially lower C_{llr} scores are achieved where TUULS alone, TUULS in combination with DyViS (with the latter in either its full or reduced form), or any of the individual Newcastle, Sunderland or Middlesbrough subcorpora, are used as the reference population.

With regard to RQ4, the picture is a little more mixed. Enlarging the reference population by adding DyViS speaker models (whether 60 or 100) to TUULS results in greater C_{llr} values, and thus poorer performance. On the other hand, expanding DyViS by adding TUULS speaker data yields marked improvements in system performance. Pooling TUULS together with an equal number of DyViS speakers, i.e. TUULS&60DyViS, produces a C_{llr} value approaching that of TUULS alone (0.198 versus 0.145), the latter being by far the best-performing reference population in the current set of trials.

We may at this point wish to consider whether, in real forensic speaker comparison cases, differences of the kind found here would have any impact on understanding of the evidence by triers of fact, and therefore on the outcome of criminal trials. In order to demonstrate this, consider Table 7, which shows the mapping between the numerical (L)LR scale, a verbal Likelihood Ratio scale (marked 'Term on scale', and glossed in the adjacent column; see Champod and Evett, 2000; Association of Forensic Science Providers, 2009). Differences in the value of an LR can be quite large before a threshold between categories is crossed. It is more than 9,000 in the case of the transition from 'strong support' to 'very strong support'. Conversely, they could be very small indeed, but could nevertheless still suffice to tip the conclusion over a support scale boundary.

Table 7. Evaluation-of-evidence framework mapping verbal Likelihood Ratio terms ('Term on scale') to raw numerical LR and Log$_{10}$LR scales (JP French Associates, after Association of Forensic Science Providers, 2009; Champod and Evett, 2000).

Term on scale	**Gloss**	**LR**	**LLR (Log$_{10}$LR)**
Extremely strong support	*The possibility that these results could be found under a different-speaker hypothesis can effectively be ruled out*	> 1,000,000	> 6
Very strong support	*The probability of obtaining these results is very much greater under a same-speaker hypothesis than under a different-speaker hypothesis*	10,000 - 1,000,000	4 - 6
Strong support	*The probability of obtaining these results is much greater under a same-speaker hypothesis than under a different-speaker hypothesis*	1,000 - 10,000	3 - 4
Moderately strong support	*The probability of obtaining these results is greater under a same-speaker*	100 - 1,000	2 - 3

	hypothesis than under a different-speaker hypothesis		
Moderate support	*The probability of obtaining these results is somewhat greater under a same-speaker hypothesis than under a different-speaker hypothesis*	10 - 100	1 - 2
Limited support	*The probability of obtaining these results is only slightly greater under a same-speaker hypothesis than under a different-speaker hypothesis*	1 - 10	0 - 1
Inconclusive	*The results do not provide support for either hypothesis*	1	0
Limited support	*The probability of obtaining these results is only slightly greater under a different-speaker hypothesis than under a same-speaker hypothesis*	1 - 0.1	0 to -1
Moderate support	*The probability of obtaining these results is somewhat greater under a different-speaker hypothesis than under a same-speaker hypothesis*	0.1 - 0.01	-1 to -2
Moderately strong support	*The probability of obtaining these results is greater under a different-speaker hypothesis than under a same-speaker hypothesis*	0.01 - 0.001	-2 to -3
Strong support	*The probability of obtaining these results is much greater under a different-speaker hypothesis than under a same-speaker hypothesis*	0.001 - 0.0001	-3 to -4
Very strong support	*The probability of obtaining these results is very much greater under a different-speaker hypothesis than under a same-speaker hypothesis*	0.0001 - 0.000001	-4 to -6
Extremely strong support	*The possibility that these results could be found under a same-speaker hypothesis can effectively be ruled out*	< 0.000001	< -6

In practice, it may be only the support scale conclusion rather than the numerical LR that is presented to the court. Even if the numerical LR were to be presented alongside it, it is plausible to suppose that the support scale conclusion might have the greater psychological impact on triers of fact (for relevant discussion, see Rose, 2013; Thompson, Grady, Lai and Stern, 2018). If this is indeed the case, practitioners using ASR systems and deploying accent-mismatched or accent-mixed reference populations would be advised to be heedful of the effects reported here in relation to the numerical thresholds of the support scale categories. When interpreting and reporting results, practitioners should also consider the right shift effect discussed earlier. Since an accent-mismatched reference population can, paradoxically, lead to artificially high same-speaker LRs compared with those that would be obtained with a well-matched reference population, it may appear appropriate to apply a compensatory left shift to comparison results. However, the effect seen in our Tippett plots only demonstrates an overall shift in the results, and does not reveal the extent of the effect for individual speakers. A per-speaker examination of the data reveals that the extent of the shift of an LR under matched, mixed and mismatched reference population conditions is highly variable across speakers and does not always occur in the general expected direction, nor is it highly correlated with, for example, the comparison score. Therefore, the application of a compensatory left shift cannot be recommended. These results demonstrate the need for further research into the speaker-specific factors that influence the size of the shift in LRs.

9. Conclusion

In this paper we have assessed the effects on the results of an Automatic Speaker Recognition system-based simulated forensic speaker comparison brought about by modifying the nature and size of the reference population used. We found that while there was essentially no difference is discrimination error, the system produced more poorly-calibrated and over-inflated LLRs when we used an accent-mismatched reference population. It is possible, of course, that the differences found could, at least in part, be attributable to differences in the different acoustic spaces in which the various sets of recordings were made. However, as stated in Section 7.5, those differences appear to be slight and are generally imperceptible. If consideration of them is offset against the accent differences, then – other things being equal – we would suggest that bringing down the magnitudes of the C_{llr} values reported here is most likely to be achieved by avoiding the use of an accent-mismatched database, provided that a more closely-matched alternative is available (cf. the C_{llr} of 0.427 for DyViS alone, which is considerably larger than that for the next poorest-performing reference population we tested).

On the other hand, the LLR data suggest that using any of the chosen permutations of reference populations would be legitimate, in the sense that the results – at least those for the same-speaker comparisons – are in every case very good; in all trials the Equal Error Rate was zero. With improved calibration, there is no reason to think that the C_{llr} scores could not be reduced by an appreciable margin.

Nevertheless, it must be remembered that the recordings used in this experiment were of near-optimal quality. We might therefore ask how system performance would differ if we were to have made use of more forensically realistic samples. Suspects in genuine police interviews are frequently uncooperative and/or unwilling to give anything more than minimal and quiet responses. The target voices in questioned recordings may be speaking in an aggressive or animated way using raised voice. The participants in the DyViS and TUULS databases complied willingly and consistently with the researchers' instructions and spoke at normal voice levels, mainly on 'neutral' tone. They were recorded in near-ideal acoustic environments (sound-treated rooms, high-quality recording equipment, fixed microphone distance, etc.) performing highly controlled and closely comparable verbal tasks. The recordings were made contemporaneously without any significant time lag between sections of the data collection procedure. In real criminal investigations, appreciable periods of time can elapse between the offence taking place and a suspect being interviewed, and the offender's voice and that of the suspect can differ markedly between recording sessions, even if the two samples were indeed spoken by the same person. Also, the participants were not subject to the level of psychological stress that might be experienced by a suspect in a real inquiry. Further research using non-contemporaneous samples of shorter duration and of degraded quality, representing different speech styles, and so forth, would provide us with a firmer sense of whether pooling existing speech corpora for use as reference data in ASR-assisted forensic speaker comparison is something that we would advocate in practice.

References

Aitken, C. G. G. and Taroni, F. (2004) *Statistics and the Evaluation of Evidence for Forensic Scientists* (2nd ed.). Hoboken, NJ: John Wiley and Sons.

Association of Forensic Science Providers (2009) Standards for the formulation of evaluative forensic science expert opinion. *Science and Justice* 49: 161--164.

Beal, J., Burbano-Elizondo, L. and Llamas, C. (2012) *Urban North-Eastern English: Tyneside to Teesside*. Edinburgh: Edinburgh University Press.

Braun, A., Llamas, C., Watt, D., French, P. and Robertson, D. (2018) Sub-regional 'other-accent' effects on lay listeners' speaker identification abilities: a voice line-up study with speakers and listeners from the North East of England. *International Journal of Speech, Language and the Law* 25(2): 231--255.

Caballero, M., Mariño, J.B. and Moreno, A. (2002) Multidialectal Spanish modeling for ASR. *Proceedings of the 3rd International Conference on Language Resources and Evaluation* (LREC'02), Las Palmas, Spain, May 2002, 892--895.

Champod, C. and Evett, I. (2000) Commentary on Broeders 1999. *Forensic Linguistics* 7(2): 238--243.

Davis, S. B. and Mermelstein, P. (1980) Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 28(4): 357--366.

Dellwo, V., French, P. and He, L. (2018) Voice biometrics for speaker recognition applications. In S. Frühholz and P. Belin (eds.) *The Oxford Handbook of Voice Perception* 777--795. Oxford: Oxford University Press.

Enzinger, E., Morrison, G. S. and Ochoa, F. (2016) A demonstration of the application of the new paradigm for the evaluation of forensic evidence under conditions reflecting those of a real forensic-voice-comparison case. *Science and Justice* 56: 42--57.

Enzinger, E. and 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.

Evett, I., Lambert, J. and Buckleton, J. (1995) Further observations on glass evidence interpretation. *Science and Justice* 35(4): 283--289.

French, P. (2017) A developmental history of forensic speaker comparison in the UK. *English Phonetics* 21: 271--286.

French, P. and Stevens, L. (2013) Forensic speech science. In M. Jones and R. Knight (eds.) *The Bloomsbury Companion to Phonetics* 183--197. London: Continuum.

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

Hansen, J. H. L. and Hasan, T. (2015) Speaker recognition by machines and humans: A tutorial review. *IEEE Signal Processing Magazine* 32(6): 74--99.

Hudson, T., de Jong, G., McDougall, K., Harrison, P. and Nolan, F. (2007) F0 statistics for 100 young male speakers of Standard Southern British English. *Proceedings of the 16th International Congress of Phonetic Sciences, Saarbrücken, August 2007*: 1809--1812.

Hughes, V. (2014) *The Definition of the Relevant Population and the Collection of Data for Likelihood Ratio-Based Forensic Voice Comparison*. PhD Thesis. York: University of York. http://etheses.whiterose.ac.uk/8309/1/Hughes, V. (2014) PhD.pdf

Hughes, V. and Foulkes, P. (2015) The relevant population in forensic voice comparison: effects of varying delimitations of social class and age. *Speech Communication* 66: 218--230.

Hughes, V., Harrison, P., Foulkes, P., French, P., Kavanagh, C. and San Segundo, E. (2018) The individual and the system: assessing the stability of the output of a semi-automatic forensic voice comparison system. *Proceedings of Interspeech 2018, Hyderabad, India*: 227--231.

Hughes, V. and Rhodes, R. (2018) Questions, propositions and assessing different levels of evidence: forensic voice comparison in practice. *Science and Justice* 58(4): 250--257.

Jessen, M., Meir, G. and Solewicz, Y. A. (2019) Evaluation of Nuance Forensics 9.2 and 11.1 under conditions reflecting those of a real forensic voice comparison case (forensic_eval_01). *Speech Communication* 110: 101--107.

Kinoshita, Y. and Ishihara, S. (2014) Background population: how does it affect LR based forensic voice comparison? *International Journal of Speech, Language and the Law* 21(2): 191--224.

Künzel, H. J. (2013) Automatic speaker recognition with cross-language speech material. *International Journal of Speech, Language and the Law* 20(1): 21--44.

Meuwly, D. (2001) *Reconnaissance de locuteurs en sciences forensiques: l'apport d'une approche automatique*. PhD thesis, University of Lausanne. Retrieved on 25 March 2020 from https://serval.unil.ch/resource/serval:BIB_R_7892.P001/REF

Morrison, G. S. (2012) The likelihood-ratio framework and forensic evidence in court: a response to R v T. *International Journal of Evidence and Proof* 16: 1--29.

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.

Nolan, F., McDougall, K., de Jong, G. and 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(1): 31--57.

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(2): 277--324.

San Segundo, E., Foulkes, P., French, P., Harrison, P., Hughes, V. and Kavanagh, C. (2019) The use of vocal profile analysis for speaker characterization: methodological proposals. *Journal of the International Phonetic Association* 49(3): 353--380.

Smith, R. L. and Charrow, R. P. (1975) Upper and lower bounds for the probability of guilt based on circumstantial evidence. *Journal of the American Statistical Association* 70: 555--560.

Snyder, D., Garcia-Romero, D., McCree, A., Sell, G., Povey, D. and Khudanpur, S. (2018) Spoken language recognition using X-vectors. *Proceedings of Odyssey 2018: The Speech and Language Recognition Workshop*, Les Sables d'Olonne, France, June 2018: 105--111.

Sohn, J., Kim, N. S. and Sung, W. (1999) A statistical model-based voice activity detection. *IEEE Signal Processing Letters* 6(1): 1--3.

Solewicz, Y. A., Jessen, M. and van der Vloed, D. (2017) Null-hypothesis LLR: a proposal for forensic automatic speaker recognition. *Proceedings of Interspeech 2017*, Stockholm, August 2017, 2849--2853. doi: 10.21437/Interspeech.2017-1023

Thompson, W. C., Grady, R. H., Lai, E. and Stern, H. S. (2018) Perceived strength of forensic scientists' reporting statements about source conclusions. *Law, Probability and Risk* 17(2): 133--155.

Tippett, C., Emerson, V., Fereday, M., Lawton, F. and Lampert, S. (1968) The evidential value of the comparison of paint flakes from sources other than vehicles. *Journal of the Forensic Science Society* 8: 61--65.

van der Vloed, D., Jessen, M. and Gfroerer, S. (2017) Experiments with two forensic automatic speaker comparison systems using reference populations that (mis)match the test language. *Proceedings of the Audio Engineering Society International Conference on Audio Forensics*, Arlington, VA, June 2017. Retrieved on 25 March 2020 from http://www.aes.org/e-lib/browse.cfm?elib=18743

Van Leeuwen, D.A. and Bouten, J.S. (2004) Results of the 2003 NFI-TNO forensic speaker recognition evaluation. *Proceedings of the Odyssey 2004 Speaker and Language Recognition Workshop*, International Speech Communication Association: 75--82.

Wells, J. C. (1982) *Accents of English 1: An Introduction*. Cambridge: Cambridge University Press.

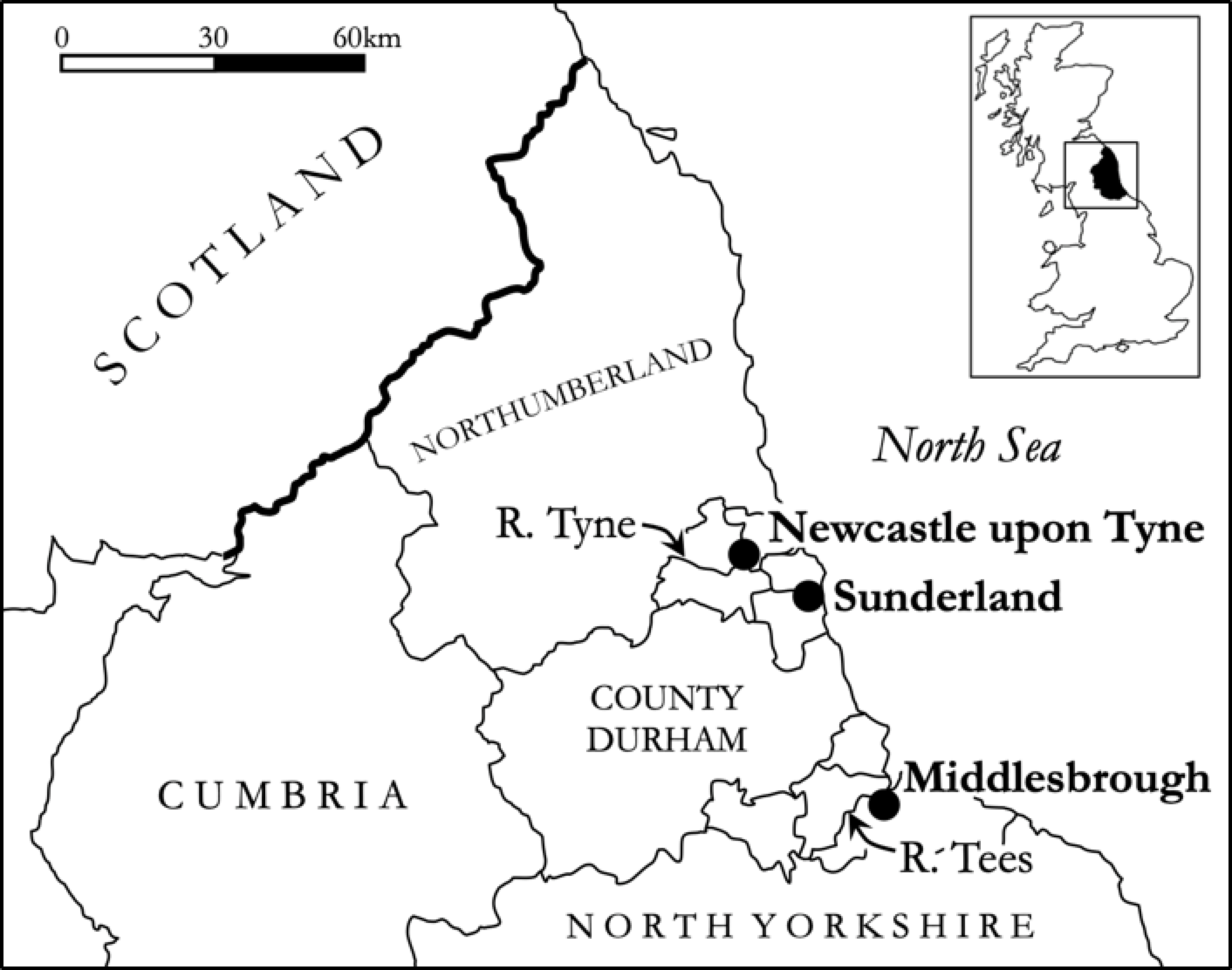
Wormald, J. (2016) *Regional Variation in Panjabi-English*. PhD Thesis, University of York. Retrieved on 25 March 2020 from http://etheses.whiterose.ac.uk/13188/1/Wormald_PhD_final.pdf

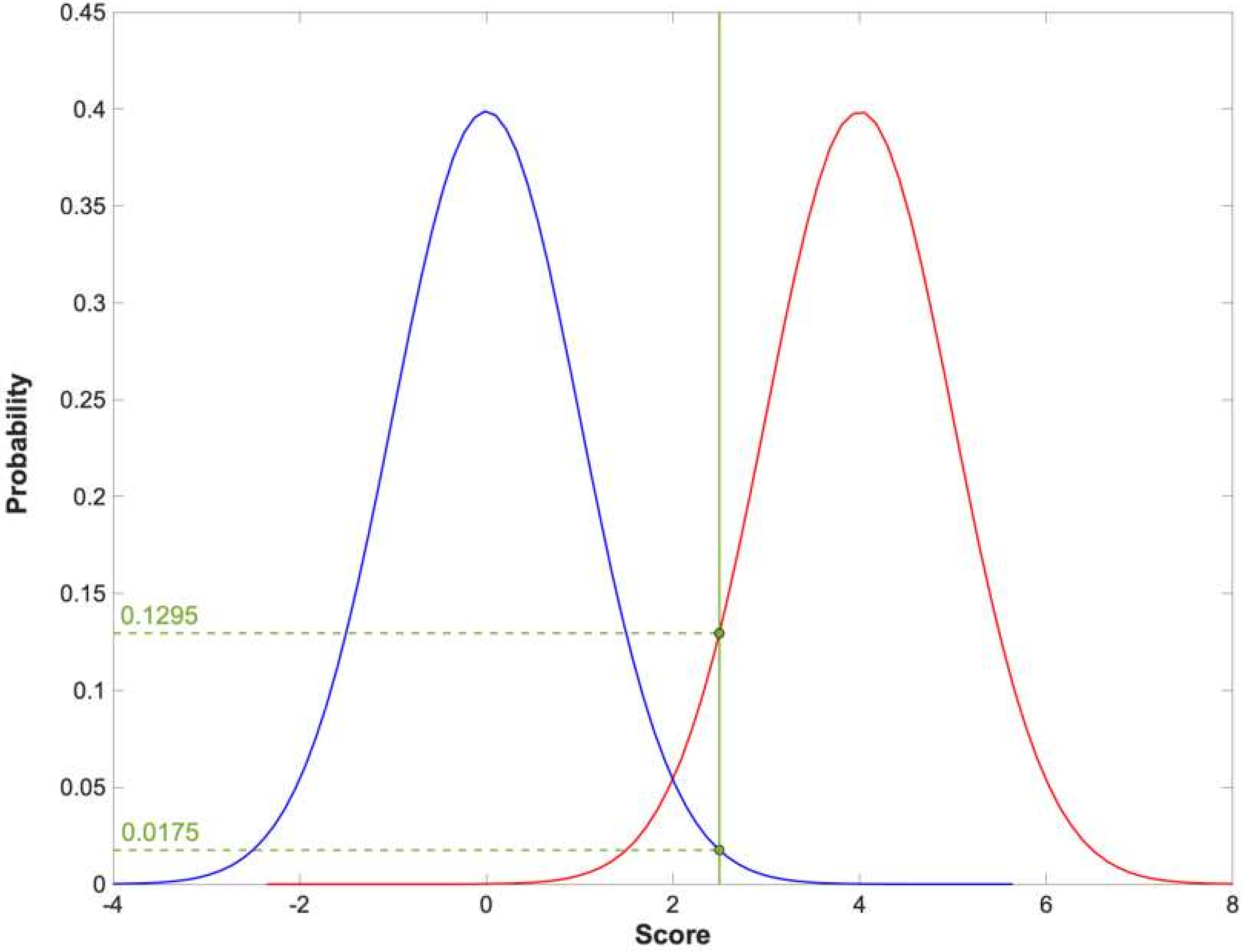
[1] We would like to express our thanks to our initial reviewer for his extensive and insightful commentary on the draft version of this article.

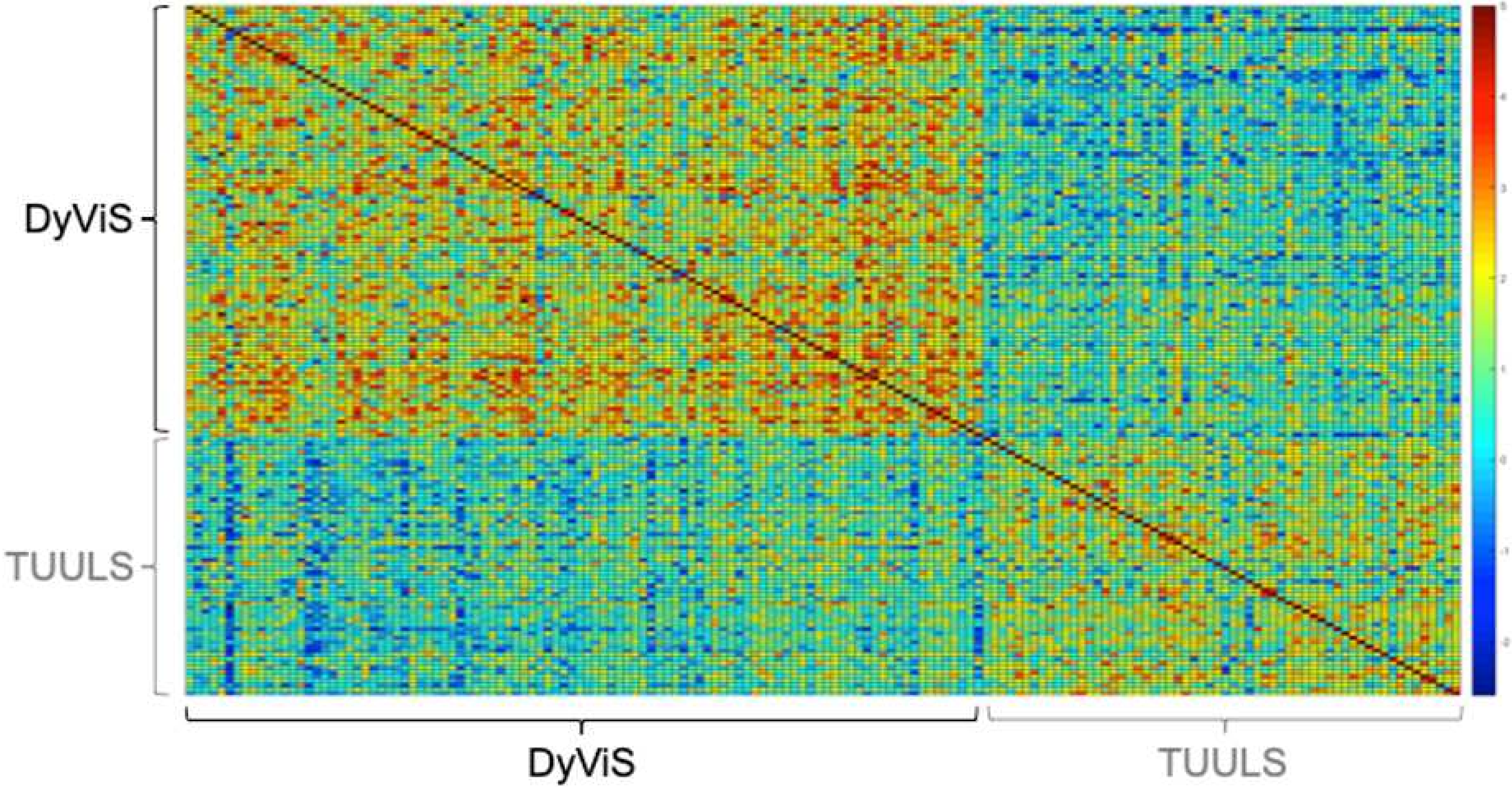
[2] Supported by UK Economic and Social Research Council grant no. RES-000-23-1248, 2005-2009.

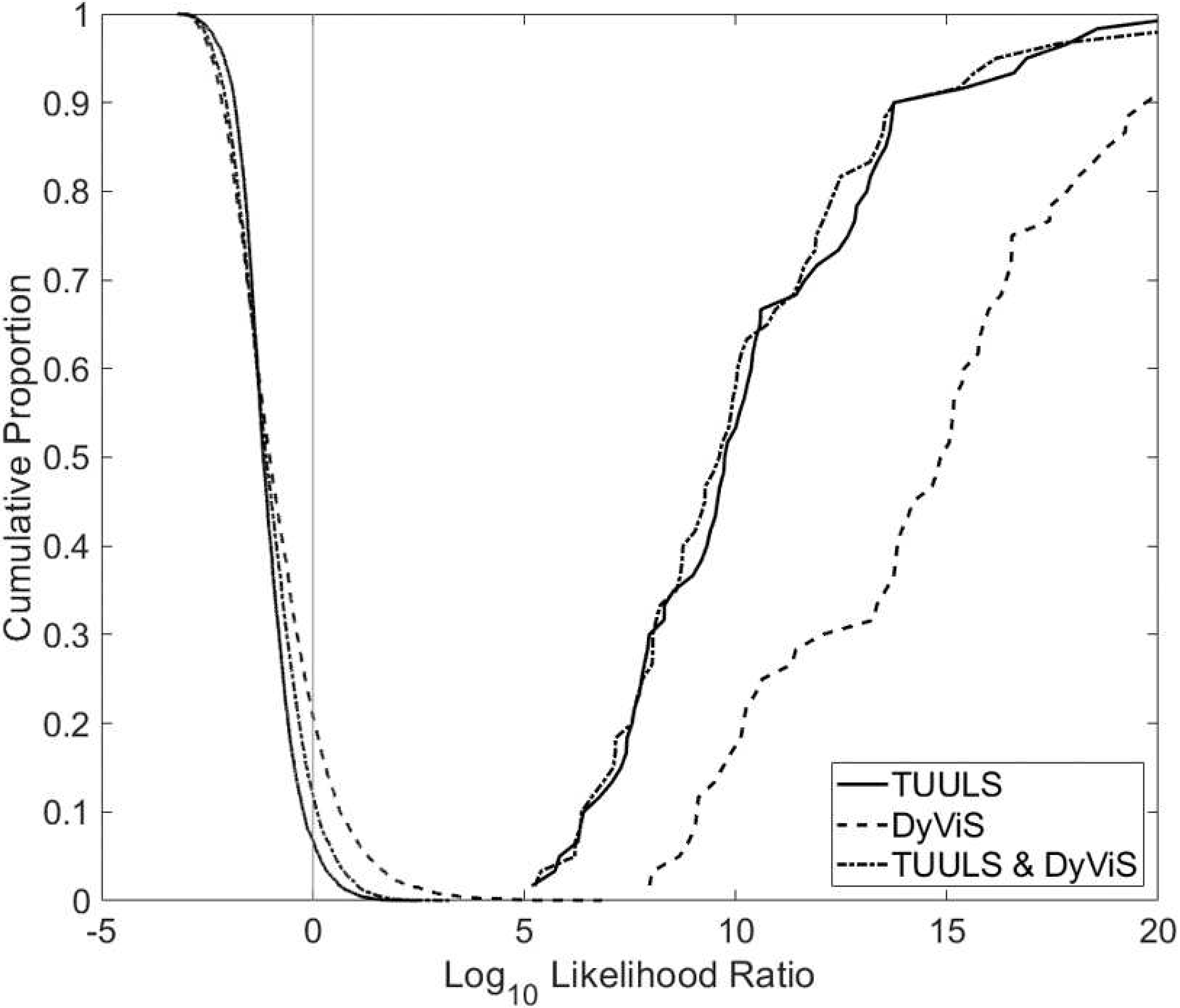
[3] Supported by UK Economic and Social Research Council grant no. ES/M010783/1, 2016-2019.

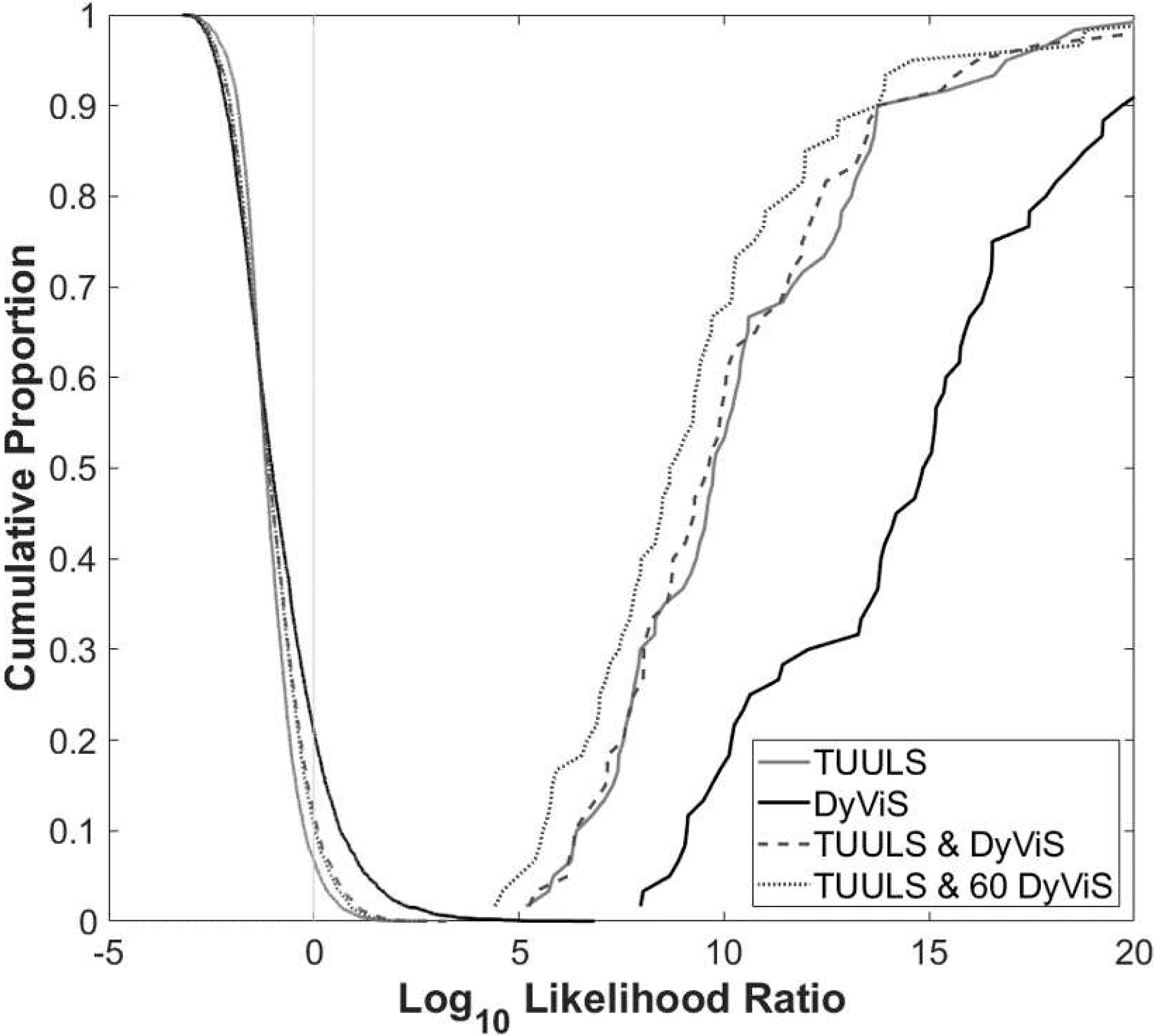
[4] In Wells' (1982) terms, these differences are *systemic*, *distributional* and *realisational*.

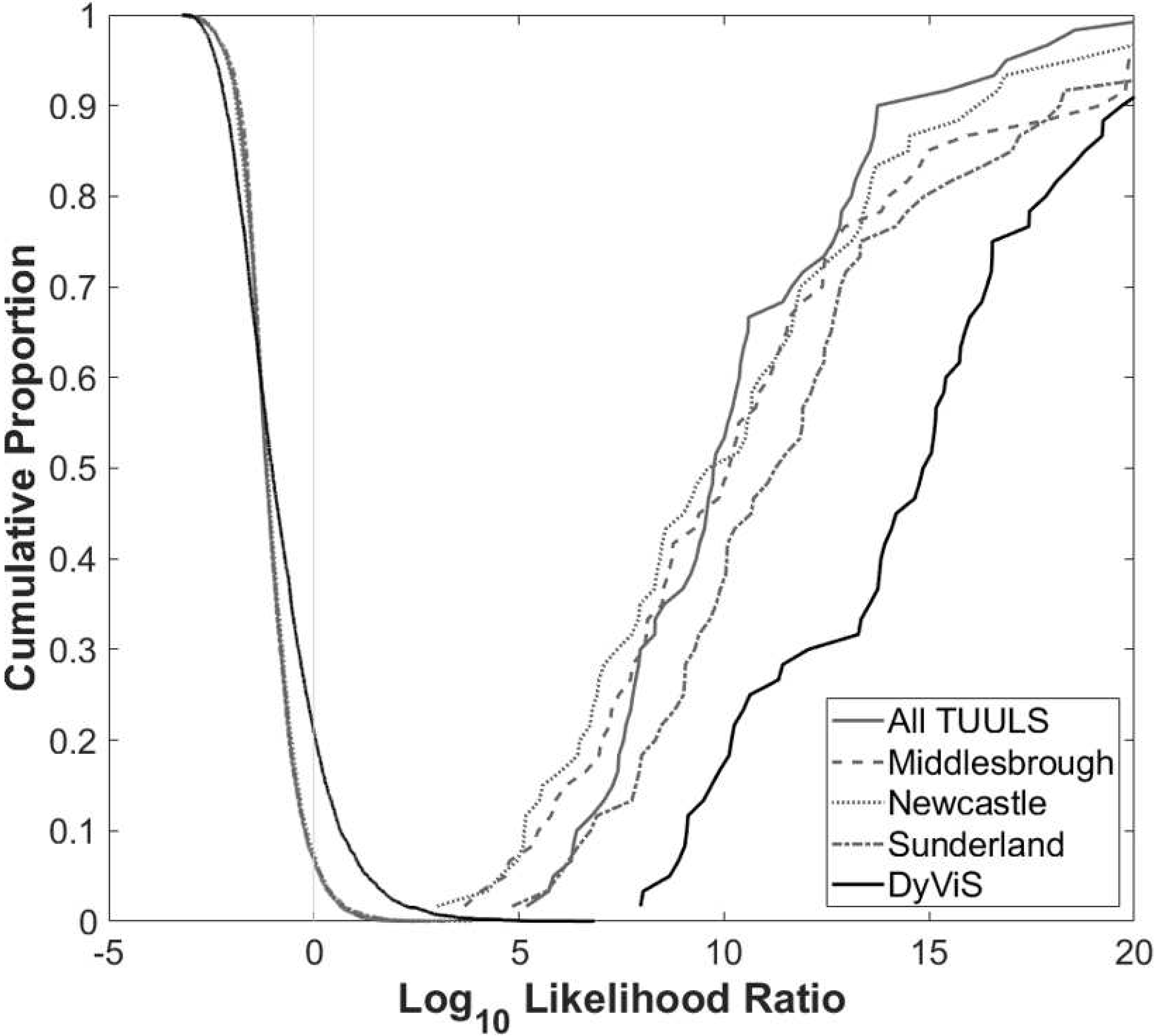

0
30
60km
SCOTLAND
NORTHUMBERLAND
North Sea
R. Tyne
Newcastle upon Tyne
Sunderland
COUNTY DURHAM
CUMBRIA
Middlesbrough
R. Tees
NORTH YORKSHIRE

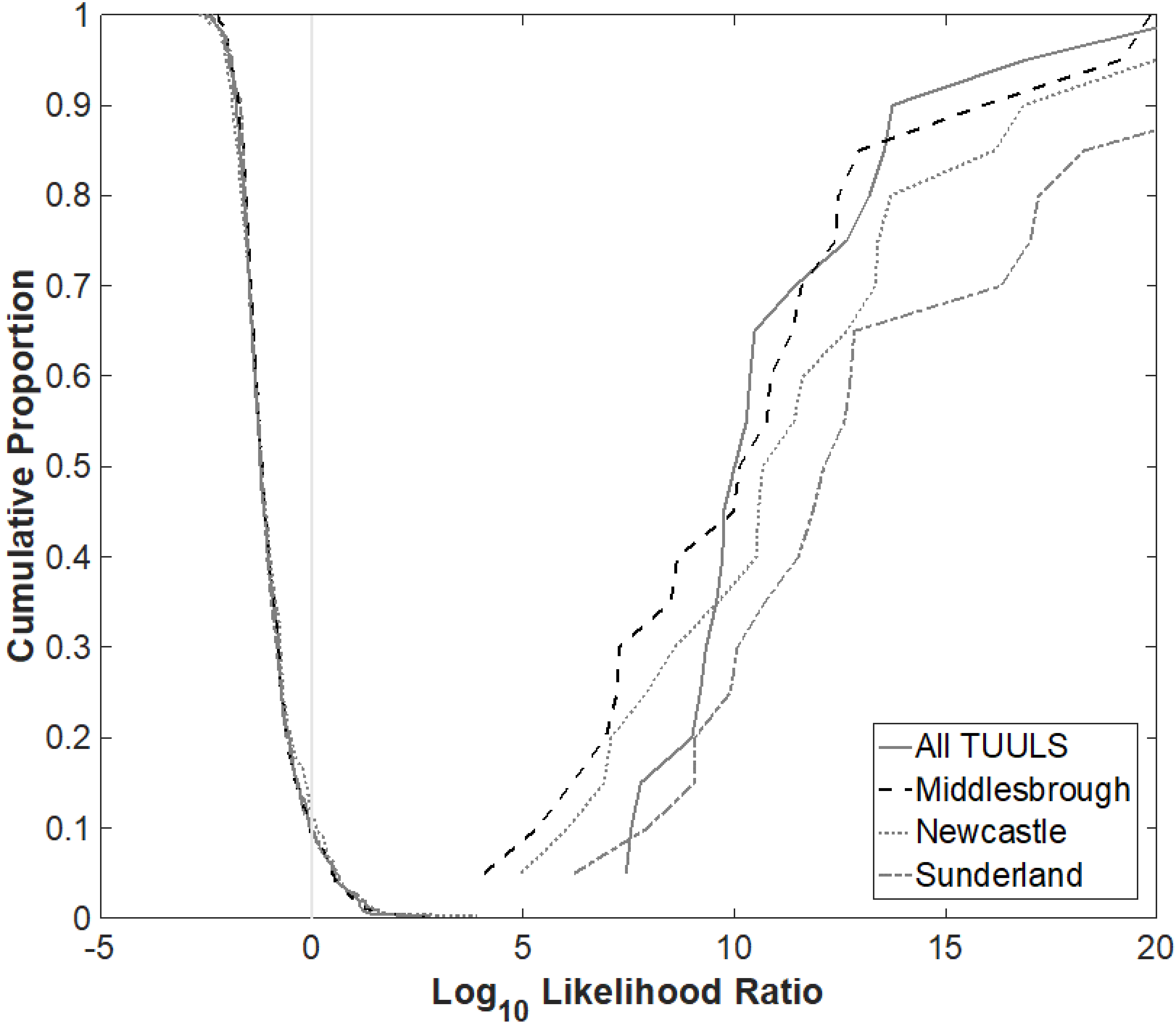

0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
Probability
0.1295
0.0175
-4
-2
0
2
4
6
8
Score


DyViS
TUULS
DyViS
TUULS
5
4
3
2
1
0
-1
-2


Cumulative Proportion
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-5
0
5
10
15
20
Log$_{10}$ Likelihood Ratio
TUULS
DyViS
TUULS & DyViS


Cumulative Proportion
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-5
0
5
10
15
20
Log$_{10}$ Likelihood Ratio
TUULS
DyViS
TUULS & DyViS
TUULS & 60 DyViS


Cumulative Proportion
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-5
0
5
10
15
20
Log$_{10}$ Likelihood Ratio
All TUULS
Middlesbrough
Newcastle
Sunderland
DyViS


Cumulative Proportion
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-5
0
5
10
15
20
Log$_{10}$ Likelihood Ratio
All TUULS
Middlesbrough
Newcastle
Sunderland