The relevant population in forensic voice comparison: effects of varying delimitations of social class and age

Abstract

In forensic voice comparison, the expert is typically instructed to compare the voices in a pair of offender and suspect samples. To appropriately evaluate the strength of such evidence, it is necessary to consider both the similarity between the samples and their typicality in the wider, relevant population. This paper considers the effects of different definitions of the relevant population when computing numerical likelihood ratios (LR), with specific regard to socio-economic class and age. Input data consist of cubic polynomial estimations of F1, F2 and F3 trajectories for /eɪ/ in New Zealand English. Calibrated LRs are computed for a sociolinguistically homogeneous sets of test data using three systems comprising of training and reference data which, with regard to the social class or age of the test speakers, are Matched, Mismatched or Mixed. The distributions of LRs were found to be relatively stable across systems, although LRs for individual comparisons may be substantially affected. As expected, the Mismatched systems produced the worst validity, while the Matched systems produced the best validity. The implications of these results for voice comparison casework are considered in light of the paradox that one cannot know for certain the sociolinguistic community to which the offender belongs.

1 Introduction

Forensic voice comparison (FVC) typically involves comparison between a recording of the voice of an unknown offender (e.g. in a covertly recorded drug deal) and a recording of the voice of a known suspect (from a police interview in the UK). The expert speech evidence is used by the trier-of-fact, along with other evidence in the case, in establishing whether the voices belong to the same or different individual(s). The likelihood ratio (LR) is now widely accepted across forensic disciplines as the logically and legally correct framework for evaluating the strength of such evidence (Robertson and Vignaux, 1995; Aitken and Taroni, 2004). The LR is the ratio of the probability (*p*) of the evidence (E) assuming the prosecution proposition ($\mathrm{H_p}$) and the probability of the evidence assuming the defence proposition ($\mathrm{H_p}$). The LR can be expressed as (1):

(1)

$$\frac{p(\mathrm{E}|\mathrm{H_p})}{p(\mathrm{E}|\mathrm{H_d})}$$

In FVC, the prosecution proposition may be expressed as: *the source of the offender recording is the suspect*. Therefore, the numerator of the LR is equivalent to the similarity between the offender and suspect samples. The defence proposition, in general terms, can be expressed as: *the source of the offender recording is not the suspect but some other speaker from the relevant population*. The denominator is equivalent to the typicality of the offender sample (i.e. the evidence) with respect to the relevant population. Therefore, the LR is the answer to a specific question (Morrison, 2009a), and the definition of the relevant population is an essential element in determining, specifically, what that question is.

In principle, the relevant population is defined by the defence proposition, which may vary in specificity depending on the case. For example, the defence claim might be that the offender was not their client but his brother. In such a circumstance, analysis need only be made of the speech of the two individuals concerned. Unfortunately, narrowly-defined defence propositions like this are rare in casework. It is far more common that the defence offer a non-specific alternative proposition (e.g. *the offender* *is not the defendant, it was someone else*) or no alternative at all. In most FVC cases it is therefore important that the analyst considers what the appropriate definition of the relevant population is (Morrison et al., 2012; Morrison and Stoel, 2014; Morrison, 2014). This is highlighted by Morrison et al. (2012) who provide an empirical demonstration of better system performance when selecting speakers to represent the relevant population based on an assumption about what the relevant population is compared with selecting speakers randomly from a larger database.

In the absence of a specific defence proposition, the concept of *logical relevance* (Kaye, 2004) has been used to define the default relevant population (Morrison et al., 2012 propose an alternative based on speaker similarity; for issues with this see Gold and Hughes, 2014). Logical relevance refers to the grouping variables which are known to affect the frequency of observations in the population at large (e.g. ethnicity in forensic DNA analysis). Since the relevant population is defined by the defence proposition, logical relevance must account for characteristics of the offender rather than the suspect. However, in forensic casework we face a paradox: the community of which the offender is a member cannot be established without knowing the offender's identity, yet this is the very issue at stake in the case. As a consequence, the logically relevant factors which define the relevant population cannot themselves be known for certain. It is therefore usually the case that pragmatic decisions, presumably based on some sort of linguistic analysis, are taken by the expert to permit analysis to be undertaken.

The offender sample can be analysed to make reasonable estimates of the speech community (or communities) to which the offender belongs. In effect this means that the analyst must produce a speaker profile of the offender (Ellis, 1994; French and Harrison, 2006; Jessen, 2008). In FVC the default assumption about the defence proposition has predominantly been that: *the voice in the offender sample does not belong to the suspect, but to another same-sex speaker of the language* (Rose, 2004). Thus, the offender profile rarely extends beyond identifying the speaker's sex and main language(s) or dialect(s) (Rose, 2004). This default definition of the relevant population, limited to sex and broadly-defined regional background (dialect or language), has been used extensively in LR-based research (e.g. Kinoshita, 2002; Rose et al., 2006; Rose, 2006; Morrison, 2009b) and casework (Rose, 2013).

However, speech is a complex form of forensic evidence. As is well known in phonetics and linguistics, particularly sociolinguistics and sociophonetics, speech is affected by a remarkably wide range of factors that generate both within- and between-speaker variation (Rose, 2002; Foulkes and Docherty, 2006; French et al., 2010). Systematic variation is found as a function of social factors such as the speaker's socio-economic class, age, and ethnicity, the social networks and communities of practice in which the speaker participates, and a very broad range of factors that can be collectively labelled 'speech style', which include variation related to topic, formality, self-consciousness, interlocutor, conversational function, and physical setting. Further sources of variation relate to short- and long-term health issues, and technical effects introduced by recording and transmission media. Since these factors can affect linguistic and phonetic variables, it is imperative that appropriate control is exercised over data used in any analysis: incorrect or inappropriate delimitation of extralinguistic factors could in principle yield misleading LRs.

The focus of the present discussion is variation related to sociolinguistic factors. Linguistic-phonetic variables are socially stratified both within and between regional communities (Labov, 1971). This has important implications for FVC. Different variables are often socially stratified in different ways. For example, /uː/-fronting in English (the vowel in *goose, boot,* etc.) is a widespread change in progress, and is generally correlated with speaker age. By contrast, another on-going change, /əʊ/-fronting (in *goat, boat* etc.), correlates with both age and speaker sex, being led by young females (Haddican et al., 2013). Thus, based on observed sociolinguistic patterns, analysis of /u:/ in a forensic case would need to control for speaker age, while analysis of /əʊ/ would need to take account of both age and sex. Further, the social stratification of a variable may differ according to regional variety. For instance, the vowels /əʊ/ (*goat, boat*) and /eɪ/ (*face, bait*) carry a great deal of social conditioning in the north-east of England, but much less so in the south-east of England (Watt, 2000).

Despite such complexity, the potential logical relevance of socio-indexical factors beyond sex and language is rarely considered in LR-based analysis (exceptions include Loakes, 2006; Zhang et al., 2011; Hughes and Foulkes, in press). Furthermore, an underlying assumption of the Rose (2004) default is that sex and language information is readily accessible from the offender sample. However, many cases present themselves where even speaker sex and the language being spoken are not trivial issues (examples are cited by French et al., 2010: 145 and Foulkes and French, 2012: 569). Conversely, it will often be possible for the sociolinguistically-informed expert to determine considerably more demographic information about the offender, beyond sex and language. This paper therefore explores the extent to which different delimitations of two important sociolingustic factors - socio-economic class and speaker age - affect LR output. Class and age were chosen as illustrative of the wide array of socio-indexical factors which may affect LR output, but which are typically overlooked.

While it is important that a FVC system is tested under conditions which reflect the case at trial (Morrison, 2014), given the complexity of speech evidence there will inevitably be some degree of mismatch for any set of data used for training and testing. For this reason, it is important that the analyst makes pragmatic decisions to control those factors which are known to substantially affect the strength of evidence and system performance. In the case of class and age variation, by far the best outcome, in terms of the implications for casework, would be to find very little variation in system performance using different sets of reference data. However, the extent to which such factors are important is an empirical question which merits testing, rather than basing the relevant population on assumptions that sources of variation other than sex and language are not relevant for FVC.

The data we analyse to address these issues are cubic polynomial coefficients of the first three formants (four coefficients per formant) of the diphthong /eɪ/ in New Zealand English (NZE). We discuss two experiments which consider variability in (1) class and (2) age. In each experiment, LRs are computed for a homogeneous set of test data in three conditions, each comprising training and reference data based on different definitions of the logically relevant population. The **Matched** condition represents a narrow definition of the class or age of the relevant population, which is matches the class or age of the offender. The **Mismatched** condition is illustrative of the paradox that the analyst cannot know for certain the population of which the offender is a member. In this condition, the relevant population is defined narrowly, but incorrectly with regard to the class or age of the offender. Finally, the **Mixed** condition reflects the current approach to applying logical relevance in FVC whereby neither class nor age are controlled. In this condition, the training and reference data consist of speakers of the same sex and regional background, and are balanced for class and age.

In a FVC case, the relevant population defines the properties of the speakers used in system testing, prior to the application of the system to compute a numerical LR for the suspect and offender samples. Therefore, in these experiments, Matched, Mismatched and Mixed data are used throughout both the feature-to-score conversion and score-to-LR mapping (i.e. calibration) stages (Morrison, 2013). The output from the different systems is compared in terms of the distributions of parallel sets of calibrated same-speaker (SS) and different-speaker (DS) log LRs (LLRs) from the same test data. The variability in LLRs from individual comparisons across systems is considered, along with overall system validity (C_{llr}). The results are also compared across the two experiments.

2 Method

2.1 /eɪ/ in New Zealand English (NZE)

The choice of /eɪ/ is motivated by known patterns of variation and change in NZE and the availability of a large amount of data. There has been considerable change in the quality of NZE diphthongs over the last century. With /eɪ/ there is clear evidence of *diphthong shift,* attested as far back as 1887, in which the onset element has lowered and backed towards [a] or [ɐ] (Gordon et al., 2004). Gordon et al. (2004: 149) claim that a second phase of change involved *glide weakening*, reducing the amount of articulatory movement between onset and offset. Despite broad processes of change over time there remains considerable variation in the phonetic realisation of /eɪ/ in present-day NZE.

There is also correlation between phonetic variation and speakers’ socio-economic background, and thus these vowels act as social markers in NZE (Hay et al., 2008).

Hay et al. (2008) distinguish between ‘cultivated’ and ‘broad’ accents of NZE. For /eɪ/, phonetic variation relates primarily to the onset, which is more open and back in broader NZE. In terms of the acoustic output, this predicts that broad speakers will display higher F1 and lower F2 values at the onset than cultivated speakers. Variation in the position of the onset element also causes differences in the amount of phonetic movement across the duration of the vowel, since the offset position for both groups is located in a similar position (close-mid [e]).

2.2 Dealing with auto-generated formant data

Data were extracted from the Canterbury Corpus (CC) of the Origins of New Zealand English (ONZE) database (see §2.5 and Gordon et al., 2007). The CC contains spontaneous speech recordings from 418 people born between 1926 and 1987. For the purposes of this study only male speakers were included. The data form a regionally homogeneous population, containing almost equal numbers of younger (20-30 years) and older (45-60 years) speakers, and professionals and non-professionals. All participants were born in New Zealand (NZ), with the majority from Canterbury. The sound files in the corpus have been collected since 1994 and have been digitised at a sampling rate of 44.1kHz and a 16-bit depth. The ONZE sound files along with orthographic transcriptions, coding at different levels of representation, and metadata about speakers, are embedded within the LaBB-CAT software[1] (Fromont and Hay, 2008; Fromont and Hay, 2012). LaBB-CAT is an online platform for storing and sharing large corpora. It is optimised for searching for specific linguistic variables according to different speaker groups, and can be used to extract large amounts of data automatically.

Phoneme-level forced-alignment was performed as part of the ONZE project, using the Hidden Markov Model Toolkit (HTK; Young et al., 2006; see further Fromont and Hay, 2012). Using the LaBB-CAT platform it was possible to inspect the forced-aligned TextGrids in order to manually hand-correct erroneous segmental boundaries for target /eɪ/ tokens using the waveform and spectrogram. Segment boundaries marked during hand-correction were determined by a variety of criteria depending on

[1] http://labbcat.sourceforge.net (accessed 5th September 2014)

the adjacent phonological context, including changes in amplitude and the onset/offset of periodicity (Turk et al., 2006).

The first three formants of /eɪ/ were measured using an automatic Praat (Boersma and Weenink, 2011). extraction script[2] embedded in LaBB-CAT. The script creates a Formant object for the entire sound file using the To formant (burg)… function which performs short-term spectral analysis of 5ms wide Gaussian windowed frames shifted at 2.5ms steps (i.e. with 50% overlap between adjacent window shifts). For each window, the function estimates formant frequencies based on linear predictive coding (LPC) coefficients using the Burg algorithm. Pre-emphasis was also applied, which was set to amplify frequency components above 50Hz to account for spectral tilt. The script identifies maximally five formants (i.e. an LPC order of 10) within a range of 0 to 5 kHz, based on an expectation for roughly one formant per 1kHz for adult male speakers (Keller, 2005). The script then uses the regions from the TextGrids to extract the dynamic data for F1, F2 and F3. This produced nine time-normalised frequency measurements taken at each +10% step of each formant, which capture the dynamic properties of the full formant trajectory (following McDougall 2004).

This semi-automatic approach was used to generate a large amount of data in a short space of time, since manual formant extraction is labour intensive. As highlighted in Zhang et al. (2012), the reliability of auto-generated formant data is expected to be poorer than human-supervised formant extraction even with high quality recordings. However, Zhang et al. (2013) conclude that human supervised formant extraction is not warranted for FVC casework “give the high-cost … and the relatively small levels of meaningful improvement it provides” (p. 808) relative to the performance of a much cheaper, generic MFCC-based FVC system. Nonetheless, a series of heuristic procedures were implemented to correct or remove errors.

Given the number of speakers and the relatively small number of tokens for most speakers it was not possible to ensure that the same number of tokens in every phonological context were included for each speaker. Rather, all tokens with adjacent /l/ and /r/ were removed due to coarticulatory and long-domain resonance effects of

[2] Hudson, T. and Williams, C. ‘IntervalFormants_use_me3.praat’ and ‘common.praat’

liquids on vowel formants (West, 1999). The data also contained multiple tokens of the indirect object ‘a’, all of which were removed since in spontaneous speech it is predicted that almost all instances will be reduced to schwa [ə]. Broad accept-reject thresholds were then applied to all of the data to remove obvious measurement errors (such as F2 measured as F1). A wide pass-band for F1 values of between 200 and 900Hz was chosen based on expectations for considerable movement on the open-close dimension between onset and offset. For F2 a range of 1100 to 2200 Hz was implemented, to capture the maximal F2 movement assuming the onset of /eɪ/ can be central [ɐ] and offset front [ɪ]. For F3, a range of 2000 to 3000 Hz was used. Tokens with values outside of these ranges were removed.

2.3 Dividing the data

Within ONZE speakers are classified for social class, and labelled as either ‘professional’ or ‘non-professional’ based on occupation and education level (Gordon *et al.* 2007: 91). A six-point version of the Elley-Irving scale was used as a metric of occupation level (Elley and Irving 1985). A similar six-point scale adapted from Gregersen and Pedersen (1991) was used to code for education level. Scores were added together, with low values representing higher social class. In the ONZE data, those classed as ‘professional’ scored on average between 4 and 4.5, whilst ‘non-professionals’ scored between 8.5 and 9.5. The labels assigned to speakers in the ONZE coding were used to divide the current data.

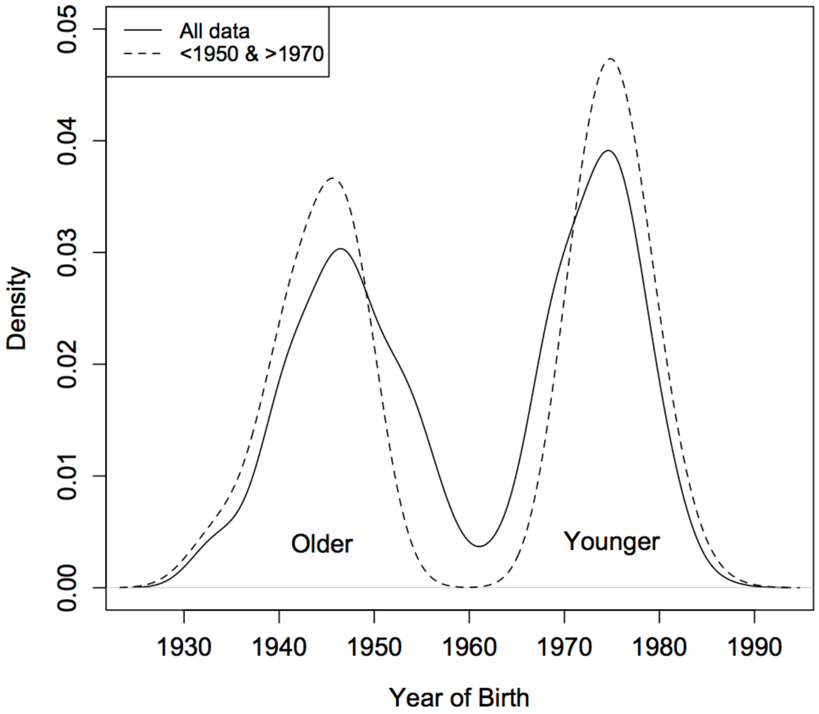

Figure 1 – Density plot of bimodal distribution of year of birth from the entire dataset (solid) and from the subdivided dataset consisting of speakers born before 1950 and after 1970 (dashed)

Information relating to age in the ONZE data was limited to year of birth, although it is possible to deduce a range for age at the time of recording based on when the recordings were made. The continuous 'year of birth' variable was converted into a discrete variable with two levels, 'older' and 'younger'. Across the entire data set there is a wide age range with speakers born between 1932 and 1982. The distribution of year of birth is also bimodal, with a dip around the mean (ca. 1960 – see Figure 1). For our purposes speakers born after 1970 were classed as *younger*, and those born before 1950 were classed as *older* (see Figure 1). Speakers born between 1951 and 1969 were removed (ca. 50 speakers from the entire dataset). The decision to divide the sample in this way ensured we avoided a cliff-edge turning point at 1960, and also took into account the fact that age at the time of recording is not known precisely.

The class-by-age classification generated four sub-groups: younger professionals, younger non-professionals, older professionals, and older non-professionals. After speakers had been separated into subgroups, z-scores for each +10% formant measurement were calculated relative to the pooled mean across all speakers within each sub-group in order to remove univariate outliers. Tokens containing a value greater than ± 3.29 standard deviations (SDs) (Tabachnick and Fidell, 2007) from the sub-group mean were removed. The sub-groups were analysed separately in order to preserve patterns of sociolinguistic variation across groups whilst also removing measurement errors.

2.4 Parametric representations

The procedures outlined above removed the most obvious measurement errors. However, such procedures are reductive in that tokens were removed if any single measurement value did not fit the criteria. Therefore, a final procedure was implemented to correct more localised errors without removing the entire token from the analysis. Each formant trajectory from each token was fitted with a cubic polynomial curve of the form $y = f(x) = ax^3 + bx^2 + cx + d$. Cubic polynomials were chosen on *a priori* assumptions about the complexity of the formant trajectories of /eɪ/ in NZE and based on Morrison (2009b). Individual raw frequency values with residuals of greater than 50 Hz for F1 and F2 or 100 Hz for F3 (relative to the fitted value) were then removed (Figure 2). These heuristics were determined based on expectations for the maximal extent of potential acoustic variation between adjacent points in the formant trajectory (separated typically by less than 10ms). A cubic polynomial curve was then re-fitted to the remaining data.

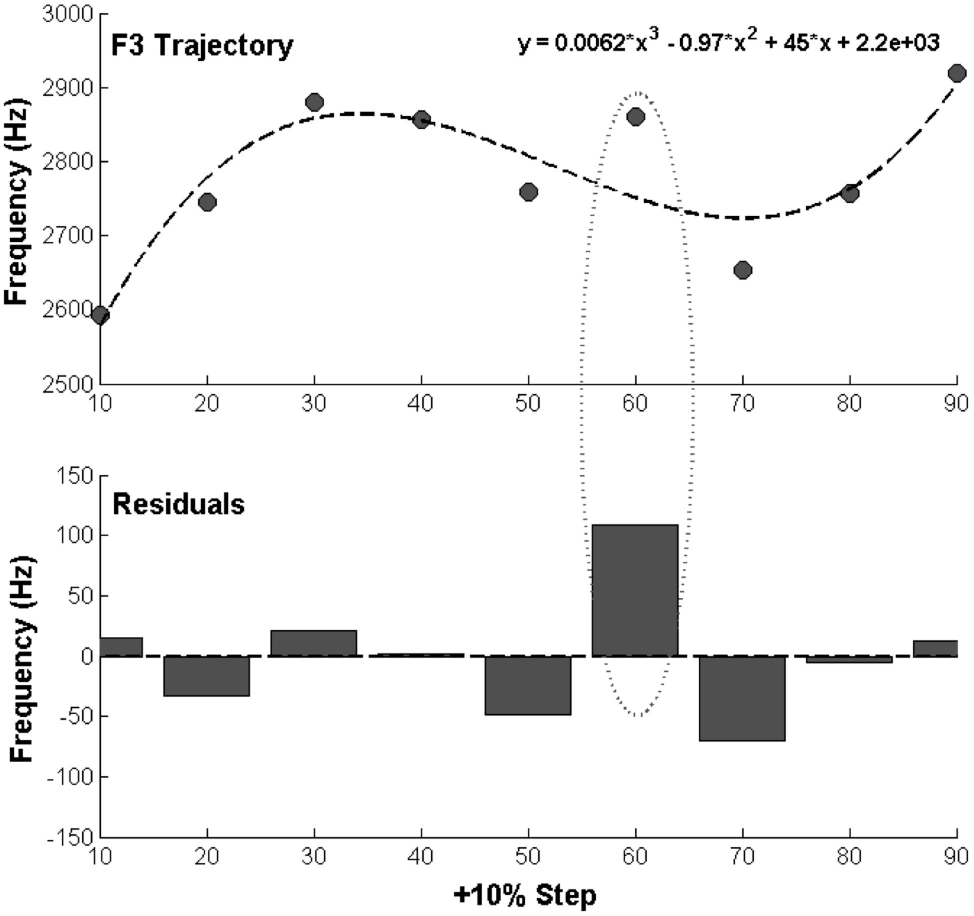

Figure 2 – Raw F3 values for a single token (y) fitted with a cubic polynomial (yfit) (black dashed curve) (above) and values with a residual greater than ±100 Hz identified (ellipsis) (below)

Finally, between-speaker z-scores within each class-by-age group were calculated for each cubic polynomial coefficient from the refitted curve. Tokens with any outlying values of greater than ±3.29 SDs from the group mean were again removed. Speakers with fewer than eight available tokens were then removed. A minimum of eight tokens per speaker was chosen after trial and error procedures comparing the trade-off between number of tokens and maximal number of speakers. The final dataset consisted of 120 speakers with eight tokens per speaker: 33 younger professionals, 31 younger non-professionals, 32 older professionals and 24 older non-professionals.

2.5 Variability in the data

The raw data were analysed to assess the extent to which class and age effects were present. Figure 3 displays mean F1~F2 trajectories within the vowel plane plotted for each class-by-age group. The data are interpreted with regard to the acoustic-articulatory correlations between F1 and the open-close dimension and F2 and the front-back dimension (Ladefoged and Johnson, 2014). As is common in phonetics, the axes are reversed in order to interpret the acoustic data in traditional articulatory terms.

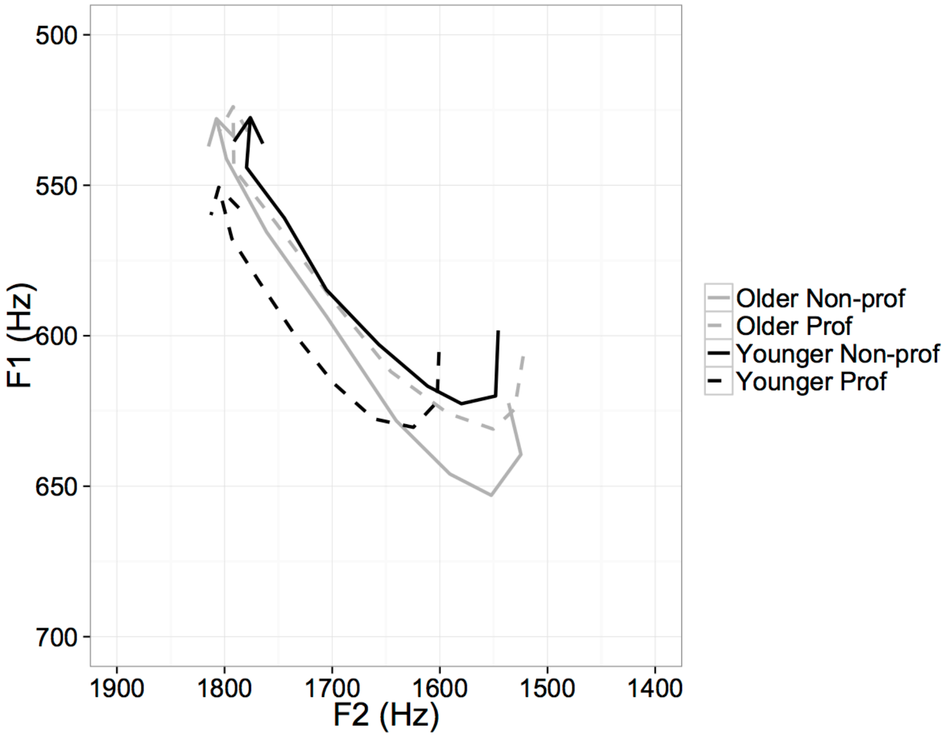

Figure 3 – F1~F2 plot of mean /eɪ/ trajectories according to class and age groupings for all 120 speakers based on 8 tokens per speaker

Figure 3 provides evidence of differences between the groups. The older non-professionals display the most open onset (i.e. highest F1), with the older professionals, younger professionals and younger non-professionals differing primarily on the F2 dimension. This is consistent with a more [a]-like first target for the older non-professionals. Generally /eɪ/ is fronter (i.e. higher F2) for the younger professionals, who also display a more open offset position, resulting in less F1~F2 movement across the vowel compared with the other groups. For the other three groups the offset position is very similar. These patterns are broadly consistent with the predictions in Gordon et al. (2004), although the magnitude of the differences between groups is somewhat less than anticipated. Due to the divergence of the older non-professionals from the other groups, age-based differences are expected to have a greater effect on LR output than class-based differences for this dataset.

3 Experiments

We now turn to the two experiments conducted to investigate the role of class and age in the definition of the relevant population, and their effect on LRs.

3.1 System testing

Within each experiment, the output from different systems is compared, where each system represents a different definition of the relevant population. In these experiments, systems are made up of training and reference data, both of which contain speakers from the relevant population. The training data are treated as suspect and offender samples, and the feature vectors for each SS and DS comparison converted to LR-like scores using the reference data to assess typicality. Feature-to-score conversion is performed by taking the ratio of the probability of the offender value at the intersection of the multivariate suspect model and the probability of the offender value at the intersection of the multivariate reference model.

Throughout the experiments in this paper, a single set of homogeneous (with regard to class and age) test data is used across different relevant population systems. The test data are treated as mock suspect and offender recordings. A single set of homogeneous test data was used to recreate typical FVC conditions in which the suspect and offender are members of the same sociolinguistic speech community. As with the training data, SS and DS scores are computed for the test data. The training scores are then used to build a calibration model which is applied to the test scores to convert them to calibrated log LRs (LLRs) (Morrison 2013).

3.2 Relevant population conditions

20 speakers were identified at random from the 33-speaker younger professional group. This group was used because it consisted of the largest number of speakers allowing for separate sets of test and training/reference speakers. In each experiment, calibrated LLRs for the 20-speaker test set were computed using three conditions based on different definitions of the relevant population (Table 1).

Table 1 – Number of training, test and reference speakers used in each condition within each experiment

	System	**Test**	**Training and reference**
Experiment (1): class	Matched	20 Younger Profs	24 Profs
	Mismatched	20 Younger Profs	24 Non-profs
	Mixed	20 Younger Profs	12 Profs +12 Non-profs
Experiment (2): age	Matched	20 Younger Profs	24 Younger
	Mismatched	20 Younger Profs	24 Older
	Mixed	20 Younger Profs	12 Younger + 12 Older

The **Matched** condition involved training and reference data consisting of 24 speakers who matched the test data for class (professional) or age (younger). The Matched system reflects an appropriate, narrowly-defined relevant population according to the demographic background of the offender. In this case, the defence proposition may be formulated as: *the voice on the offender sample is not that of the suspect, but of another professional/younger male speaker of New Zealand English.*

The **Mismatched** condition used 24 non-professional or older speakers as training and reference data. This represents a narrowly-defined but inappropriate relevant population, based on an incorrect judgement about the class or age of the offender. In this case the defence proposition may be formulated as: *the voice on the offender sample is not that of the suspect, but of another non-professional/older male speaker of New Zealand English.*

Finally, the **Mixed** condition contained 24 speakers consisting of equal numbers of professionals and non-professionals, and younger and older speakers. By balancing the class and age profile, the Mixed system reflects the current application of logical relevance to FVC in which neither class nor age are controlled. The use of Matched, Mismatched and Mixed training and reference data ensures that the different definitions of the relevant population are applied during the feature-to-score stage as well as during score-to-LR mapping.

In experiment (1) development and reference data in all systems were balanced for age. That is, the Matched and Mismatched sets consisted of equal numbers of younger (12) and older (12) speakers. Similarly, in experiment (2) development and reference data were balanced for class, with equal numbers of professionals (12) and non-professionals (12) used in the Matched and Mismatched sets. The same Mixed data, consisting of six speakers from each class-by-age combination, was used in both experiments. In all cases, the 24 Matched, Mismatched and Mixed speakers were identified at random from the appropriate class-by-age sub-group.

3.3 Feature-to-score conversion and score-to-LR mapping

Due to there only being recordings available from a single session, data for each speaker in the training and test sets was divided in half to act as suspect and offender material. This is a limitation of the current data since, in FVC casework, it is common for there to be some time delay (of weeks or months) between the recording of the two samples. (Cases do arise where contemporaneous data may be appropriate, however, for example where the question concerns speaker identity at different points in the same recording.) As shown in Enzinger and Morrison (2012), the use of contemporaneous samples may result in an underestimation of the within-speaker variation typically found in casework, leading to overly optimistic strength of evidence and system validity.

A further limitation of the data is the relatively small number of tokens available for analysis. By dividing speaker data in half, the suspect and offender data consisted of only four tokens each. As highlighted by Brümmer and Swart (2014), small amounts of data lead to poor estimates of density functions which may in turn lead to larger magnitude LLRs. However, such small numbers of tokens are consistent with typical forensic casework conditions in which speech samples are often very short and data are sparse. Small numbers of tokens are also common in LR-based research (e.g. Rose, 2011). Despite the limitations of the database, for the purposes of this study it is considered useful since it was not possible to find a forensically realistic database with adequate controls over the socio-indexical factors of interest.

In each experiment, cross-validated scores were initially computed for the 24-speaker Matched, Mismatched and Mixed training sets based on the four-token suspect and offender data using Aitken and Lucy's Multivariate Kernel Density (MVKD) formula implemented in R using the *Comparison* package.[3] The input data consisted of four polynomial coefficients per formant, generating a 12 dimensional density function for the suspect and reference data. The MVKD procedure models between-speaker variation with a speaker-dependent Gaussian kernel density and within-speaker variation with an assumption of normality. The two halves of each sample were used to compute 24 SS scores. DS comparisons (276 DS scores) used the first half of the suspect data and the second half of the offender data. In this way all DS scores were independent of each other.

Using cross-validation, typicality was assessed using a reference set of 22 speakers, although the speakers included in this set changed for each comparison. For DS comparisons, the reference data consisted of all of the available speakers in the Matched, Mismatched or Mixed set with the exception of the suspect and the offender. For SS comparisons, the speaker being compared was excluded from the reference data along with another speaker identified at random. This ensured that the same number of speakers was used as reference data across all comparisons. Cross-validation was used since philosophically $p(\mathrm{E}|\mathrm{H_d})$ is estimated based on a population which is defined by, but which does not include, the offender. The suspect was also excluded from the reference data since in casework the suspect will not be used to build the reference model for computing $p(\mathrm{E}|\mathrm{H_d})$.

[3] Lucy, D. 2013. Comparison (version 1.0-4) (R package). http://cran.r-project.org/web/packages/comparison/index.html (accessed 8th August 2014).

SS (20) and DS (190) MVKD scores for the 20 test speakers were then computed using the 24-speaker Matched, Mismatched and Mixed reference sets to generate three sets of parallel scores per experiment. The distributions of SS and DS scores for each of the three sets of training data per experiment were used to train a logistic regression calibration model for each system (Morrison, 2013). The calibration coefficients were calculated using a robust version of Brümmer's (2007) procedure.[4] The logistic regression calibration procedure minimises the log LR cost function (C_{llr}; see §3.4) and also typically minimises the magnitudes of LLRs which offer contrary-to-fact support for prosecution or defence propositions as a result. The calibration coefficients for each relevant population system (per experiment) were then applied to the appropriate set of test scores (for each system) to convert the scores to LLRs.

3.4 Evaluation of performance

Within each experiment, systems were compared in terms of the distributions of calibrated LLRs and the central tendencies evaluated using the median LLR, since the distributions of LLRs are typically skewed. Validity was assessed for each system using C_{llr} (Brümmer and du Preez, 2006). C_{llr} is a gradient measure of the validity of a forensic comparison system, which penalises the system for high contrary-to-fact LRs. It is preferred over a threshold-based, accept-reject metric of validity such as equal error rate (EER), since it is philosophically consistent with the LR framework.

Given that the calibrated LLRs for each relevant population system were generated from the same 20 test speakers, it was possible to assess the variation in the LLRs for each individual comparison across systems. Within each experiment, the deviation of the SS and DS LLRs produced by the Mismatched and Mixed systems was assessed relative to the LLRs from the Matched set using the Root Mean Square (RMS) difference. The mean RMS difference for each of the Mismatched and Mixed systems, analysing SS and DS LLRs separately, is defined as:

[4] Morrison, G. S. 2009. Robust version of train_llr_fusion.m from Niko Brümmer's FoCal toolbox. http://geoff-morrison.net/#TrainFus (accessed 25th March 2013).

(2)

$$\text{RMS difference} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2},$$

where:

n = number of comparisons (20 SS or 190 DS)

x_i = i^{th} Matched LLR

y_i = i^{th} Mismatched/ Mixed LLR.

The RMS difference is an absolute value (i.e. non-negative) which represents the average deviation of individual LLRs produced by the Mismatched or Mixed systems from those in the Matched condition, where large RMS differences indicate greater divergence. The value itself can be interpreted on a $\log_{10}$ LR scale. The Matched system is used as the point of comparison based on the narrowest, appropriate definition of the relevant population.

4 Results

The results are assessed within experiments in terms of the distributions of calibrated SS and DS LLRs, C_{llr} and the RMS differences. The systematicity of the patterns in the two experiments is considered in §4.3.

4.1 Experiment 1: class

Figure 4 displays the Tippett plot of LLRs according to class-based definitions of the relevant population. The distributions of SS LLRs are similar across the three systems. In all cases, the median SS LLR is between zero (i.e. neutral evidence) and +1. There are small differences in terms of contrary-to-fact SS LLRs, with the Mixed system producing the strongest support for the defence (up to -0.725). The Mismatched system produced the weakest contrary-to-fact SS LLRs, as well as the lowest proportion of contrary-to-fact SS LLRs (5%).

Slightly larger differences between systems are revealed for DS LLRs. The Matched median (-1.042) is one order of magnitude stronger than the Mismatched median (-0.210), indicating that the Matched system generates the strongest DS LLRs. The Mixed median is marginally weaker (-0.931) than that of the Matched system, although the absolute difference is small. The highest magnitude contrary-to-fact values are generated by the Matched system (up to +1.548), while the weakest contrary-to-fact LLRs are produced by the Mismatched system (up to +0.766). However, this system also produces the highest proportion of positive DS LLRs (37.4%).

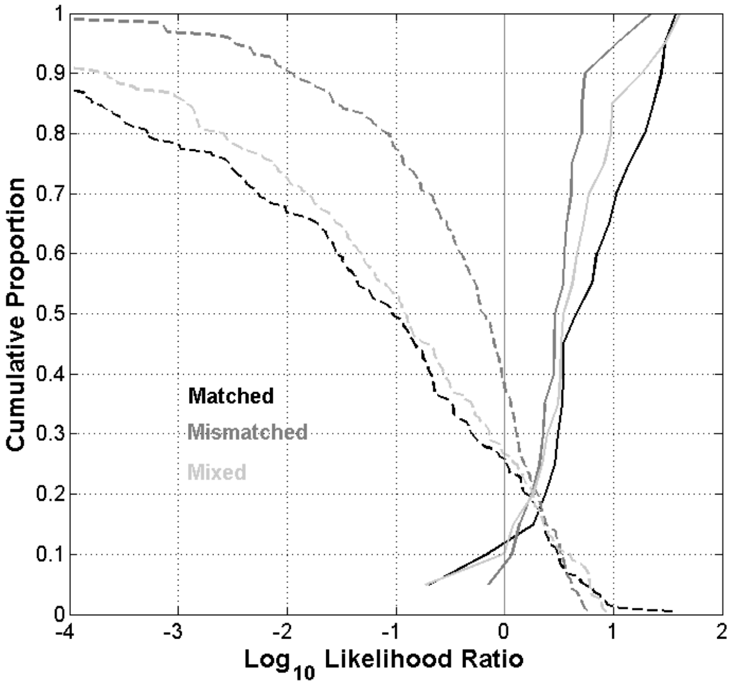

Figure 4 – Tippett plots of calibrated SS and DS LLRs using the three systems based on different definitions of the relevant population according to class

The effects of using Mismatched and Mixed systems, relative to the Matched system, on the LLRs from individual SS and DS comparisons were evaluated using the RMS difference. In terms of SS LLRs, the RMS difference is relatively small and only marginally greater for the Mismatched system (0.485) than for the Mixed system (0.410). The RMS difference between the Matched and Mismatched/Mixed LLRs is larger for SDS comparisons than for SS comparisons. However, the RMS difference is again greater for the Mismatched system (1.118) than for the Mixed (0.951) system.

In the case of the Mismatched system, the direction of the difference from the Matched values is almost always towards weaker LLRs.

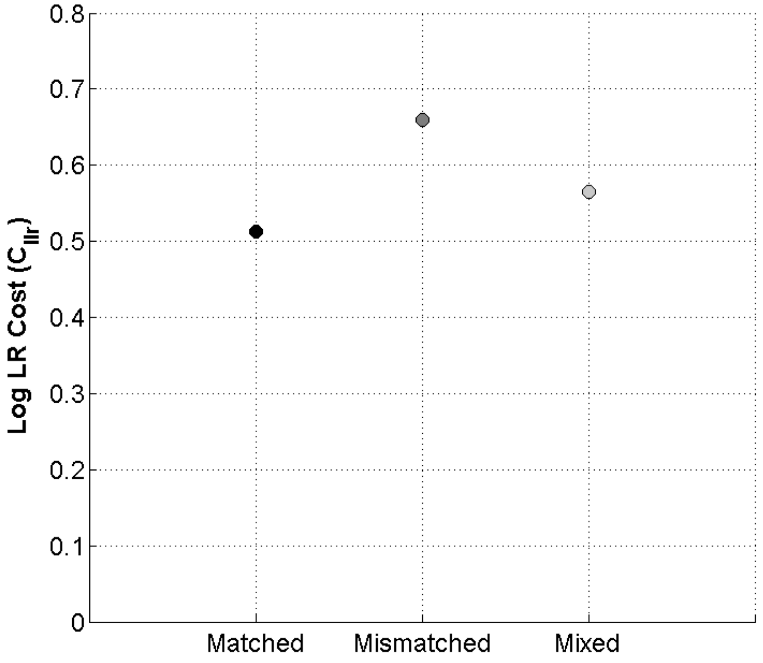

Figure 5 – Log LR Cost (C_{llr}) for each of the three class-based systems

Figure 5 displays C_{llr} values for each of the three class-based systems. The best performing system is the Matched system which produces a C_{llr} of 0.513. The most divergent performance from the Matched system is found for the Mismatched system, which produces the poorest validity (0.659). Consistent with the distributions of SS and DS LLRs in Figure 4, the C_{llr} for the Mixed system (0.565) is much closer to that using the Matched data. Despite this, validity is still poorer for the Mixed system. This reflects, primarily, the marginally higher magnitude of contrary-to-fact SS LLRs, and the higher proportion of higher magnitude contrary-to-fact DS LLRs in the Mixed results.

4.2 Experiment 2: age

Figure 6 displays the Tippett plot of LLRs for the three age-based systems. The general patterns are similar to those in §4.1, although the absolute differences between systems are smaller. SS medians across all three systems are within the same order of magnitude, between zero and +1. The strongest overall ranges of SS LLRs are also comparable, with values maximally extending to marginally above +1. The Mismatched system produces no contrary-to-fact SS LLRs, while the magnitude of the contrary-to-fact LLRs in the Matched and Mixed set are all within the range of zero and -1.

Similar patterns to those in §4.1 are revealed in the distributions of DS LLRs. Median DS LLRs are within the same order of magnitude, between zero and -1, although the absolute numerical differences are greater than for the SS LLRs. The median is weakest using the Mismatched data (-0.106), with values from the Mismatched system generally indicating weaker support for the defence compared with the Matched and Mixed systems. The distribution of Mixed DS LLRs is much closer to that from the Matched system. The magnitudes of contrary-to-fact LLRs are, however, similar across the three conditions, producing values consistently lower than +1. As in §4.1 the proportion of contrary-to-fact DS LLRs is greatest using the Mismatched system.

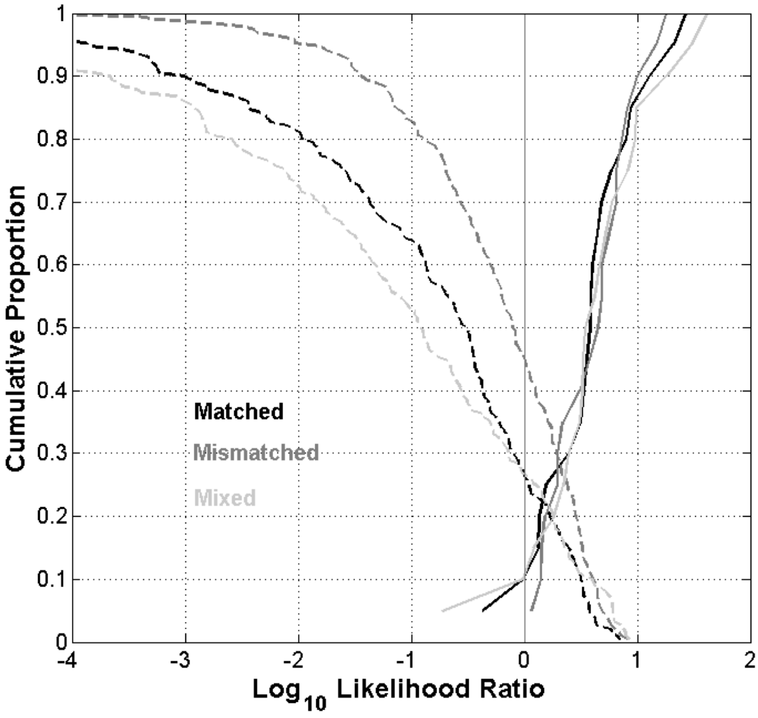

Figure 6 – Tippett plots of calibrated SS and DS LLRs using the three systems based on different definitions of the relevant population according to class

Consistent with Figure 6, the RMS differences indicate relatively little fluctuation of individual SS LLRs across systems. The RMS difference is marginally higher for the Mismatched condition (0.424) than for the Mixed condition (0.356), although both values are lower than in Experiment (1). The RMS difference is again greater for DS LLRs than for SS LLRs. However, unlike §4.1, the largest RMS difference from the Matched LLRs is produced by the Mixed condition (1.049). Further, unlike the SS LLRs, the RMS difference between the Matched and Mixed DS LLRs is greater in Experiment (2) than in Experiment (1). Conversely, the RMS difference for the Mismatched DS LLRs is lower than in experiment (1), indicating greater stability in individual LLRs using the Mismatched data based on age variation compared with that based on class variation.

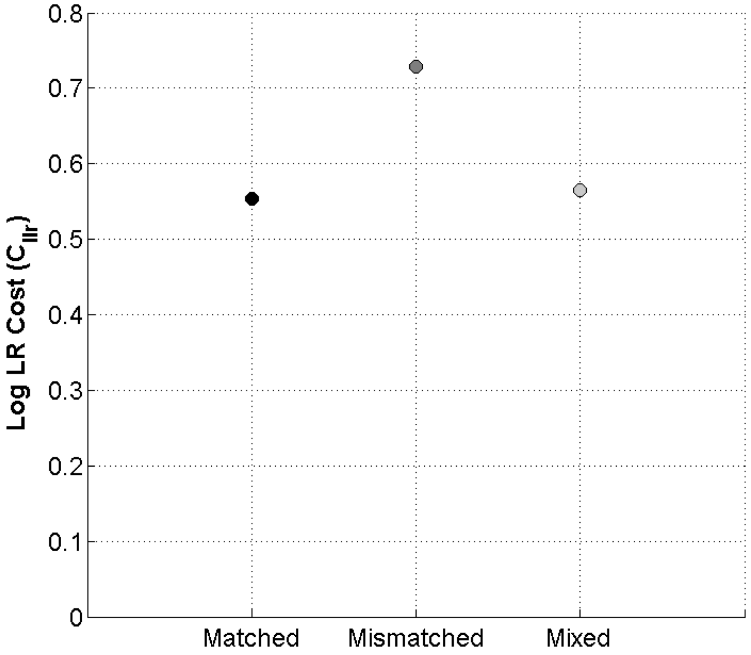

Figure 7 – Log LR Cost (C_{llr}) for each of the three age-based systems

Finally, Figure 7 displays the C_{llr} values produced by the age-based systems. The best performing system is again the Matched system (0.561), although the Mixed system produces a C_{llr} of almost equal magnitude (0.565). This indicates that the overall performance of the system using Mixed data is considerably closer to that using Matched data, compared with using Mismatched data. The poorest performing condition is again the Mismatched condition (0.712).

4.3 Systematic patterns or random variation?

The results in §4.1 and §4.2 revealed some differences in the distributions of LLRs and system validity as a function of the class- or age-based definition of the relevant population. However, it is not clear whether the these patterns are an inherent property of using Matched, Mismatched and Mixed data or random variation as a function of sample size. To test this issue, Experiments (1) and (2) were re-run 20 times using speakers sampled randomly from the appropriate class-by-age subsets. In each replication, the speakers used as test, Matched, Mismatched and Mixed data changed, although sample size remained constant (20 test speakers; 24 training/reference speakers). The test data again contained young professionals and was the same for the class- and age-based experiments in each replication. The results of these replications are evaluated against the main patterns in §4.1 and §4.2.

In §4.1, the distributions of SS LLRs were found to be broadly similar across the three conditions. Consistency in the distributions of SS medians across systems was also found in the replications with values consistently within the same order of magnitude (between zero and +1). In §4.1, DS LLRs were found to be weakest in the Mismatched condition. However, the results of the replications indicate slightly different patterns. The DS median appears more stable, with values of consistently between zero and -1 across all systems across all replications. Despite this, similar RMS differences were found in §4.1 and the replications, with the DS LLRs displaying greater divergence than SS LLRs, and Mismatched LLRs displaying greater divergence than Mixed LLRs.

In §4.1, C_{llr} was found to be best using the Matched data and poorest using the Mismatched data. In the replications, the Mismatched condition also generates the highest median C_{llr}, producing the poorest validity across conditions in 16 of the 20 replications (Figure 8). However, the absolute difference between the Matched and Mismatched median C_{llr} values is relatively small, suggesting that the extent of the validity differences between the Matched and Mismatched conditions may not be as great as predicted by the results of Experiment (1). Although the distribution of C_{llr} values in the Matched and Mixed conditions are similar, in 13 of the 20 replications the Mixed condition produced poorer validity.

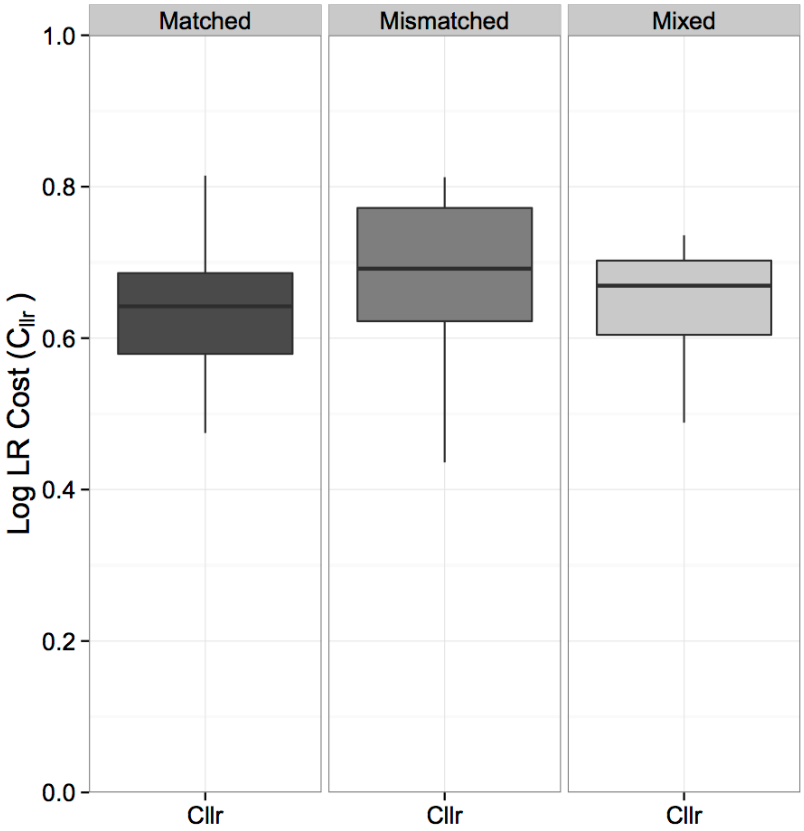

Figure 8 – Boxplots (with median, interquartile range and overall range) of the distributions C_{llr} values for the 20 replications of experiment (1) according to the three class-based conditions

In §4.2, the distributions of LLRs were found to be much more stable across systems than in §4.1, although somewhat weaker DS LLRs were again found in the Mismatched condition. Consistent with these predictions, in the replications SS medians were found to be remarkably stable across conditions with values typically fluctuating between just +0.4 and +0.6. The patterns in §4.2 were also found for DS LLRs, with the interquartile range of medians for the Mismatched system much closer to zero than those of the Matched and Mixed systems. In terms of C_{llr}, §4.2 predicted poorer validity for the Mismatched set and similar performance of the Matched and Mixed sets. This pattern was also found across the replications (Figure 9), with the median and interquartile range of C_{llr} values highest in the Mismatched condition. However, as in experiment (1), the absolute numerical difference between conditions is typically smaller than in §4.2.

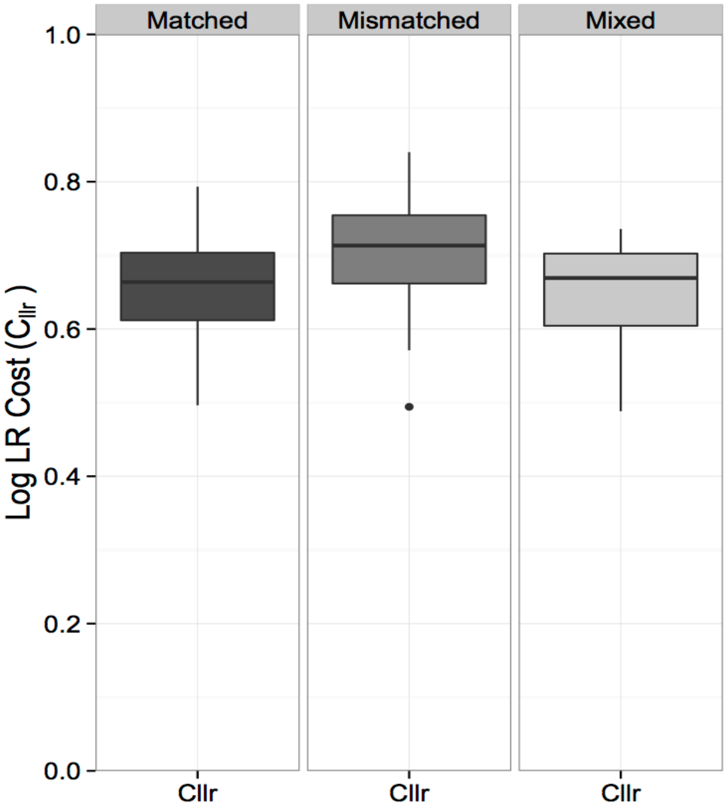

Figure 9 – Boxplots (with median, interquartile range and overall range) of the distributions C_{llr} values for the 20 replications of experiment (2) according to the three age-based conditions

5 Discussion

The results of Experiments (1) and (2), supported by the multiple replications in §4.3, reveal a number of effects of using different definitions of the class- and age-based relevant population on LR output. For class, the results of §4.1 and the replications suggest that the distribution of SS LLRs is relatively stable across different relevant population systems. §4.1 suggests that the distribution of DS LLRs is shifted closer to zero (i.e. weaker evidence) using Mismatched data and that the Mismatched system also produces highest proportion of contrary-to-fact DS LLRs. In the replications, little difference was found in the distributions of DS LLRs, although consistently the Mismatched system produced the highest proportion of false hits. For age, the overall distributions of SS and DS LLRs were relatively stable across the three systems both in the results in §4.2 and over the 20 replications.

However, somewhat bigger differences across systems were revealed in terms of LLRs from individual comparisons and validity. Despite the relative stability of the overall distributions of SS and DS LLRs, the RMS differences suggest that individual comparisons may be substantially affected by using different systems. The RMS differences in §4.1 and §4.2 revealed that individual LLRs are most divergent from the Matched value using the Mismatched set. For DS LLRs the Mismatched mean RMS difference was higher in §4.1, while the Mixed mean RMS difference was higher in §4.2. Across both experiments, individual DS LLRs were found to be much more unstable across systems than SS LLRs. Comparable patterns were found across §4.1, §4.2 and the replications in terms of validity. The Matched system consistently achieved the lowest C_{llr}, followed by the Mixed system and then by the Mismatched system. However, the absolute numerical C_{llr} differences across the systems were in some replications relatively small.

While these results do indicate patterns of divergence across the systems, the effects on LR output are somewhat smaller than expected based on the predictions in §2.1. There are a number of potential reasons for this. Firstly, the range of acoustic/phonetic variation in the data was rather less marked that the descriptive literature had suggested. Secondly, the relatively small number of tokens per speaker means that the magnitude of the LRs will necessarily be smaller, thus offering a narrower range of potential variation across the different systems. Thirdly, the results suggest that /eɪ/ offers relatively weak strength of evidence (with C_{llr} optimally ca. 0.5 in the Matched condition) meaning that SS and DS LLRs are inherently closer to zero (neutral evidence). Again this reduces the range of potential variation across the three systems.

However, it is not clear why there are not bigger differences between systems in Experiment (2), despite more marked age-based variation in the raw data. To understand the results of Experiment (2), it is necessary to explore in more detail the separate effects of using Matched, Mismatched and Mixed data in feature-to-score conversion and score-to-LR mapping. The uncalibrated test scores from §4.2 and the replications in §4.3 were therefore analysed. The magnitudes of SS LLRs were found to be weakest for the Matched system, followed by the Mixed system, while the

Mismatched system generally produced the strongest SS LLRs. This is exemplified by the fact that 40% of the SS scores in §4.2 are over one order of magnitude stronger in the Mismatched condition than the Matched condition.

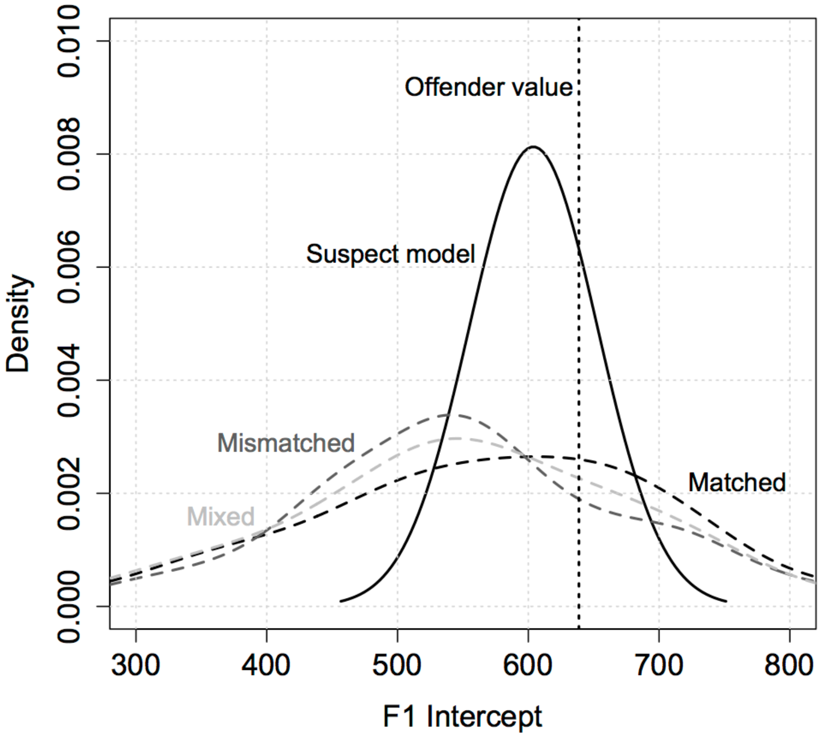

Figure 10 – Univariate example of a SS comparison (test speaker 11) from §4.2 assessing the probability of the offender value (639) at the intersection of the normal suspect model and kernel density Matched, Mismatched and Mixed models

A potential explanation for this finding is that while the suspect and offender data remain stable (i.e. $p(E|H_p)$ is the same), the distribution of the inappropriate reference data (i.e. Mismatched) is shifted relative to that of the appropriate reference data (i.e. Matched). Therefore, certain offender data will be located further into the tails of the inappropriate reference distribution, meaning that $p(E|H_d)$ is lower than it would be using appropriate reference data. The result is scores which are higher in magnitude for the inappropriate data than for the appropriate data. In the case of the Mixed data, $p(E|H_d)$ is typically between that of the Matched and Mismatched data, producing intermediate scores. A univariate kernel density example of a SS comparison from §4.2 using the F1 intercept of /eɪ/ is displayed in Figure 10. In this case, the Matched (raw) score is approximately 2.41, compared with 2.82 in the Mixed condition and 3.42 in the Mismatched condition.

Examining the uncalibrated scores based on age variation in each condition reveals a dominant effect of overinflated SS scores when using inappropriate reference data, which is not manifested in the resulting calibrated LLRs. This may be explained by the use of Matched, Mismatched and Mixed data throughout training as well as testing. If, for instance, the reference data were inappropriate with regard to age in the training data, there would be inflation of the SS scores (as in Figure 10) used to train the calibration model (the SS and DS categories are equally weighted when training the logistic regression model, meaning that individual SS scores have a bigger influence than individual DS scores). This could potentially lead to a shallower mapping function which would keep the magnitude of SS LLRs in the Mismatched condition larger than those in the Matched condition after calibration.

However, given that the training and reference sets in these experiments contained speakers of the same demographic background (whether younger or older speakers, or both), the logistic regression models across the three conditions are predicted to be roughly equivalent. This is because the output from homogeneous training and reference data (irrespective of the test data) should perform in broadly the same way when using the same input data. When the calibration model is applied to the test scores, the magnitude of the overinflated scores produced by using inappropriate reference data is reduced. Calibration therefore appears to have the effect of scaling the distribution of SS scores in the Mismatched condition towards the distribution of scores in the Matched condition, meaning that the resulting calibrated LLRs display a similar distribution.

6 Conclusion

This study has investigated the effects of applying varying definitions of the relevant population with regard to the sociolinguistic factors of class and age on the outcome of LRs. The results suggest that individual LLRs and system validity may be affected if the relevant population is defined narrowly, but incorrectly, with regard to the class or age of the offender. This is particularly important in light of the paradox we outlined earlier: in most forensic cases it cannot be known for certain whether the correct delimitation of the population has been drawn. Across both experiments, LR output based on Mixed data was considerably more similar to that produced using Matched data, although poorer validity was nonetheless still found for the Mixed systems. Therefore, a more general definition of the relevant population should be preferred unless there is no dispute as to the class and/or age of the offender. Alternatively, in casework, it may also be necessary to offer multiple LRs based on different assumptions about the relevant population, as is typical in other branches of forensic sciences (e.g. forensic DNA analysis; Kaye, 2004; Gill and Clayton, 2009).

Comparison of the uncalibrated scores and calibrated LLRs reveals that differences between the Matched, Mismatched and Mixed conditions are considerably reduced when applying a score-to-LR mapping model based on Matched, Mismatched or Mixed data. SS scores were found to be overinflated when using inappropriate reference data due to shifting of the distribution relative to the distribution of the appropriate reference data. The results suggest that calibration ameliorates the potentially detrimental effects of using inappropriate reference data to compute scores, particularly for SS comparisons.

As outlined in Morrison et al. (2012), these results highlight that it is important that analysts consider the appropriate definition of the relevant population and understand the potential effects of different controls on the outcome of the numerical LR. While this study has considered the effects of individual sources of variation using a single variable, more research is required to assess how different definitions of the relevant population would affect more sociolinguistically diverse variables. Further research should also consider the effects of different definitions of the relevant population on the overall LR in more realistic forensic conditions, where multiple variables are typically analysed and combined.

Acknowledgements

This research is funded by the UK Economic and Social Research Council and an ESRC Overseas Fieldwork bursary. We are indebted to the New Zealand Institute for Language, Brain and Behaviour (NZILBB) at the University of Canterbury for granting access to the ONZE database and to Robert Fromont for generating the formant data. Specific thanks go to Jen Hay, Pat LaShell, Richard Rhodes, Philip Harrison and the reviewers and subject editor of *Speech Communication* for helpful discussion and comments. Thanks also to Anil Alexander, Esam Alzqhoul, Niko Brümmer, Philip Harrison, Geoffrey Morrison and Balu Nair for scripts which facilitated data analysis.

References

Aitken, C. G. G. & Lucy, D. 2004. Evaluation of trace evidence in the form of multivariate data. *Appl. Stat.* 53: 109-122.

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

Boersma, P. & Weenink, D. 2011. Praat: doing phonetics by computer [Computer Program] Version 5.2.32. http://www.praat.org (accessed 22nd July 2011).

Brümmer, N. (2007) FoCal multi-class: toolkit for evaluation, fusion and calibration of multi-class recognition scores. http://sites.google.com/site/nikobrummer/focal (accessed 12th April 2012).

Brümmer, N. & du Preez, J. 2006. Application-independent evaluation of speaker detection. *Comput. Speech Lang.* 20(2/3): 230–275.

Brümmer, N. & Swart, A. 2014 Bayesian calibration for forensic evidence reporting. Paper presented at International Conference on Forensic Inference and Statistics, Leiden University, Netherlands. 19-22 Aug 2014.

Elley, W. B. & Irving, J.C. 1985. The Elley-Irving Socio-Economic Index: 1981 census revision. *New Zealand Journal of Educational Studies,* 20: 115-128.

Ellis, S. 1994. The Yorkshire Ripper enquiry: part 1. *Int. J. Speech. Lang. Law.* 1(2): 197-206.

Enzinger, E. and Morrison, G. S. 2012. The importance of using between-session test data in evaluating the performance of forensic-voice-comparison systems. In *Proceedings of the 14th Australasian International Conference on Speech Science and Technology*, Sydney, Australia. pp. 137-140.

Foulkes, P. & Docherty, G.J. 2006. The social life of phonetics and phonology. *J. Phon.* 34: 409-438.

Foulkes, P. & French, J. P. 2012. Forensic speaker comparison: a linguistic-acoustic perspective. In P. Tiersma & L. Solan (eds.) *Oxford Handbook of Language and the Law*. Oxford: OUP. pp. 557-572.

French, J. P. and Harrison, P. 2006. Investigative and Evidential Application of Forensic Speech Science. In A. Heaton-Armstrong, E. Shepherd, G. Gudjonsson & D. Wolchover (eds.) *Witness Testimony: Psychological, Investigative and Evidential Perspectives*. Oxford: Oxford University Press. pp. 247-262.

French, J. P. et al. 2010. The UK position statement on forensic speaker comparison: a rejoinder to Rose and Morrison. *Int. J. Speech. Lang. Law.* 17(1): 143-152.

Fromont, R. & Hay, J. 2008. ONZE Miner: the development of a browser-based research tool. *Corpora* 3(2): 173-193.

Fromont, R. & Hay, J. 2012. LaBB-CAT: an annotation store. *Proceedings of the 14th Australasian International Conference on Speech Science and Technology*, Sydney, Australia. pp. 113-117.

Gill, P. & Clayton, T. 2009. The current status of DNA profiling in the UK. In J. Fraser & R. Williams (eds.) *The Handbook of Forensic Science*. Cullompton: Willan Publishing. pp. 29-56.

Gold, E. & Hughes, V. 2014. Issues and opportunities: the application of the numerical likelihood ratio framework to forensic speaker comparison. *Science and Justice* 54(4): 292-299.

Gordon, E. et al. 2004. *New Zealand English: Its Origins and Evolution*. Cambridge University Press: Cambridge.

Gordon, E. et al. 2007. The ONZE corpus. In J. C. Beal, K. P. Corrigan & H. Moisl (eds.) *Models and Methods in the Handling of Unconventional Digital Corpora.* Volume 2: Diachronic Corpora. Palgrave: London. Pp. 82-104.

Gregersen, F. & Pedersen, I. 1991. *The Copenhagen study in urban sociolinguistics.* Copenhagen: C.A. Reitzels Forlag.

Haddican, B. et al. 2013. Social correlates of two vowel changes in Northern England. *Lang. Var. Change.* 25(3): 371-403.

Hay, J. et al. 2008. *New Zealand English*. Edinburgh University Press: Edinburgh.

Hughes, V. and Foulkes, P. in press. Variability in analyst decisions during the computation of numerical likelihood ratios. *Int. J. Speech. Lang. Law.*

Jessen M. 2008. Forensic phonetics. *Lang. Ling. Compass* 2(4): 671-711.

Johnson, K. 2011. *Acoustic and auditory phonetics* (3rd edition). Oxford: Blackwell.

Kaye, D. H. 2004. Logical relevance: problems with the reference population and DNA mixtures in *People v. Pizarro*. *Law, Probability and Risk* 3: 211-220.

Keller, E. 2005. The analysis of voice quality in speech processing. In G. Chollet, A. Esposito, M. Faundez-Zanuy, and M. Marinaro (eds.) *Nonlinear Speech Modeling and Applications*. Berlin: Springer Verlag. pp. 54-73.

Kinoshita, Y. 2002. Use of likelihood ratio and Bayesian approach in forensic speaker identification. *Proceedings of the 9th Australian International conference on Speech Science and Technology.* 2-5 December 2002, Melbourne, Australia. pp. 297-302.

Labov, W. 1971. The study of language in its social context. In J. A. Fishman (ed.) *Advances in the Sociology of Language* (vol. 1). The Hague: Mouton. 152-216.

Ladefoged, P. & Johnson, K. 2014. *A Course in Phonetics* (7th ed). Stamford, CT: Cengage Learning.

Loakes, D. 2006. A forensic phonetic investigation into the speech patterns of identical and non-identical twins. Unpublished PhD Dissertation, University of Melbourne, Australia.

McDougall, K. 2004. Speaker-specific formant dynamics: an experiment on Australian English /aɪ/. *Int. J. Speech. Lang. Law.* 11(1): 103-130.

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

Morrison, G. S. 2009b. Likelihood-ratio voice comparison using parametric representations of the formant trajectories of diphthongs. *J. Acoust. Soc. Amer.* 125: 2387-2397.

Morrison, G. S. 2010. Forensic voice comparison. In I. Freckleton and H. Selby (eds.) *Expert Evidence* (Ch. 99). Sydney: Thomson Reuters.

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. 2014. Distinguishing between forensic science and forensic pseudoscience: Testing of validity and reliability, and approaches to forensic voice comparison. *Science & Justice*. 54(3): 245-256.

Morrison, G. S. et al. 2012. Database selection for forensic voice comparison. *Proceedings of Odyssey 2012: The Language and Speaker Recognition Workshop, Singapore, International Speech Communication Association*.

Morrison, G. S. & Stoel, R. D. 2014. Forensic strength of evidence statements should preferably be likelihood ratios calculated using relevant data, quantitative measurements, and statistical models - a response to Lennard (2013) Fingerprint identification: How far have we come? *Australian Journal of Forensic Sciences*. 46: 282-292.

Rose, P. 2002. *Forensic Speaker Identification*. London: Taylor and Francis.

Rose, P. 2004. Technical Forensic Speaker Identification from a Bayesian Linguist's Perspective. Keynote paper, Forensic Speaker Recognition Workshop, Speaker Odyssey '04. pp. 3-10.

Rose, P. 2006. The intrinsic forensic discriminatory power of diphthongs. *Proceedings of the 11th Australasian International Conference on Speech Science and Technology*. 6-8 December 2006, University of Auckland, New Zealand. 64-69

Rose, P. 2011. Forensic voice comparison with Japanese vowel acoustics – a likelihood ratio-based approach using segmental cepstra. *Proceedings of the 17th International Congress of Phonetic Sciences.* 17-21 August 2011, Hong Kong. 1718-1721.

Rose, P. 2013. Where the science ends and the law begins: likelihood ratio-based forensic voice comparison in a $150 million telephone fraud. *Int. J. Speech. Lang. Law.* 20(2): 277-324.

Rose, P. et al. 2006. Realistic extrinsic forensic speaker discrimination with the diphthong /aɪ/. *Proceedings of the 11th Australasian International Conference on Speech Science and Technology*. University of Auckland, New Zealand, 6-8 December 2006. pp. 329-334.

Rose, P. & Morrison, G. S. 2009. A response to the UK position statement on forensic speaker comparison. *Int. J. Speech. Lang.* 16: 139-163.

Tabachnick, B. G. & Fidell, L. S. 2007. *Using Multivariate Statistics* (5th Edition). New York: Harper Collins.

Turk, A., Nakai, S. & Sugahara, M. 2006. Acoustic segment durations in prosodic research: a practical guide. In S. Sudhoff, D. Lenertová, R. Meyer, S. Pappert, P. Augurzky, I. Mleinek, N. Richter & J. Schließer (eds.) *Methods in Empirical Prosody Research*. Berlin: De Gruyter. pp. 1-28.

Watt, D. 2000. Phonetic parallels between the close-mid vowels of Tyneside English: are they internally or externally motivated? *Lang. Var. Change* 12(1): 69-101

West, P. 1999. The extent of coarticulation of English liquids: an acoustic and articulatory study. *Proceedings of the 14th International Congress of Phonetic Sciences*. San Francisco, US. 1-7 August 1999. pp. 1901-1904.

Wilder, C. 1978. Vocal aging. In B. Weinberg (ed.) *Transcripts of the seventh symposium: care of the professional voice. Part II: life span changes in the human voice.* New York: Voice Foundation.

Young, S. et al. 2006. *The HTK Book (for HTK Version 3.4).* http://htk.eng.cam.ac.uk/prot-docs/htkbook.pdf (accessed 11th September 2013)

Zhang, C. et al. 2011. Forensic voice comparison using Chinese /iau/. *Proceedings of the 17th International Congress of Phonetic Sciences.* 17-21 August 2011, Hong Kong. pp. 2280-2283.

Zhang, C. et al. 2012. Reliability of human-supervised formant-trajectory measurement for forensic voice comparison. *Journal of the Acoustical Society of America* 133: EL54-EL60.

Zhang, C. et al. 2013. Effects of telephone transmission on the performance of formant-trajectory-based forensic voice comparison – female voices. *Speech Communication* 55: 796-813.