Odyssey 2012
The Speaker and Language Recognition Workshop
25-28 June 2012, Singapore

Database selection for forensic voice comparison

Geoffrey Stewart Morrison, Felipe Ochoa, Tharmarajah Thiruvaran

Forensic Voice Comparison Laboratory, School of Electrical Engineering & Telecommunications, University of New South Wales
geoff-morrison@forensic-voice-comparison.net

Abstract

Defining the relevant population to sample is an important issue in data-based implementation of the likelihood-ratio framework for forensic voice comparison. We present a logical argument that because an investigator or prosecutor only submits suspect and offender recordings for forensic analysis if they sound sufficiently similar to each other, the appropriate defense hypothesis for the forensic scientist to adopt will usually be that the suspect is not the speaker on the offender recording but is a member of a population of speakers who sound sufficiently similar that an investigator or prosecutor would submit recordings of these speakers for forensic analysis. We propose a procedure for selecting background, development, and test databases using a panel of human listeners, and empirically test an automatic procedure inspired by the above. Although the automatic procedure is not entirely consistent with the logical arguments and human-listener procedure, it serves as a proof of concept for the importance of database selection. A forensic-voice-comparison system using the automatic database-selection procedure outperformed systems with random database selection.

Different portions of this paper have benefitted from two principal sources of financial support:

Sections 2–4: The Australian Research Council, Australian Federal Police, New South Wales Police, Queensland Police, National Institute of Forensic Science, Australasian Speech Science and Technology Association, and the Guardia Civil through Linkage Project LP100200142. Unless otherwise explicitly attributed, the opinions expressed are those of the authors and do not necessarily represent the policies or opinions of any of the above mentioned organizations.

Sections 5–7: The Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), through the Army Research Laboratory (ARL). All statements of fact, opinion or conclusions contained herein are those of the authors and should not be construed as representing the official views or policies of IARPA, the ODNI, or the U.S. Government.

Thanks to Alex Biedemann, Jason Pelecanos, anonymous reviewers and others for comments on earlier drafts of parts of this paper. The opinions expressed in this paper are those of the authors, and any shortcomings in the paper are the responsibility of the authors.

Thanks to the organizers of Odyssey 2012, COLIPS, and Temasek Laboratories@NTU for the invitation to submit the written version of this paper for publication in the Odyssey 2012 proceedings and to present the oral version in Singapore in conjunction with Odyssey 2012, including financial support for the latter from Temasek Laboratories@NTU.

1. Introduction

Difficulty in defining the appropriate population to specify in the defense hypothesis has been cited as a reason for not adopting data-based implementations of the likelihood-ratio framework for forensic voice comparison (French & Harrison [1]; French et al. [2]; but see responses in Rose & Morrison [3]; Morrison [4], Morrison [5] §99.400). In the current paper, we present a logical argument as to the appropriate population for the forensic scientist to sample for background, development, and test databases, and propose a human-listener procedure for selecting recordings to include in the sample. This is followed by a discussion of how the procedure would have been applied in three casework examples. We also discuss some potential objections to our proposed procedure. We leave empirical testing of the human-listener procedure for future research, but in the mean time describe and empirically test an automatic procedure inspired by the logical arguments and human-listener procedure. Although the automatic procedure is not fully consistent with the logical arguments, the results of tests of this procedure indicate that database selection does lead to better system performance.

2. Logical arguments

2.1. A likelihood ratio is the answer to a specific question and this question specifies the relevant population

The aim of forensic voice comparison is to produce a likelihood ratio which is an expression of the strength of the evidence with respect to two competing hypotheses (Champod & Meuwly [6]; Rose [7]; Morrison [5]). The first hypothesis, the prosecution hypothesis, is usually that a voice of questioned identity on one audio recording (the questioned-speaker /offender recording) belongs to the same speaker as the voice on one or more other audio recordings for which the identity of the speaker is not disputed (the known-speaker / suspect recording). The alternative hypothesis, the defense hypothesis, is usually that the questioned voice does not belong to the suspect, but rather belongs to some other speaker. An appropriate defense hypothesis will, however, always be more specific than “some other speaker”, and the details of the defense hypothesis are part of the definition of the question which is answered by a likelihood ratio.

A likelihood ratio cannot be interpreted without

understanding the question which it answers. Imagine that we have two likelihood ratios, one with a value of 10 million and the other with the value of 1 million. The first would appear to indicate a greater strength of evidence than the latter, but they are only comparable if they are both answers to the same question. Imagine that the first likelihood ratio was an answer to a question in which the alternative hypothesis was "any other human on the planet", and also for the sake of argument that the trier of fact (the judge, panel of judges, or jury, depending on the legal system) assigned an equal prior probability to each human, then the posterior odds would be the prior odds multiplied by the likelihood ratio: $1/(\sim 7\times 10^9) \times 10^7 = \sim 1/700$. Imagine that the second likelihood ratio was an answer to a question in which the alternative hypothesis was "any one of the other 100 humans on an island", and also for the sake of argument that the trier of fact assigned an equal prior probability to each human on the island, then the posterior odds would be the prior odds multiplied by the likelihood ratio: $1/100 \times 10^6 = 10\,000$. Although the first likelihood ratio has a higher absolute value, relative to the questions asked the second likelihood represents a much greater strength of evidence in the sense that it leads to much higher posterior odds. Numerically, the posterior odds are simply the product of the prior odds and the likelihood ratio, but both the value of the prior odds assigned by the trier of fact and the value of the likelihood ratio calculated by the forensic scientist are dependent on the particular question asked, and the particular question defines the particular relevant population that both the trier of fact and the forensic scientist consider.

2.2. The background, development, and test databases represent the relevant population

The details of the alternative hypothesis specify the relevant population from which to sample the background, development, and test databases:

2.2.1. Background database

A background database is a sample of recordings from a number of speakers in the relevant population, and is used to build a model of the distribution of measured acoustic properties in those recordings. The background model, based on a sample, is an estimate of the distribution of those acoustic properties in the whole population. The background model is used to estimate the denominator of the likelihood ratio, i.e., the probability of obtaining the measured acoustic properties in the offender recording had it been produced by some speaker from the relevant population other than the suspect. The background sample must therefore be representative of the relevant population.

The numerator of the likelihood ratio is estimated using a model trained on the suspect recording. There are often speaking-style and channel mismatches between suspect and offender recordings, e.g., the former could be a recording made of a subdued conversation on a landline telephone from a remand center and the latter could be a recording of a lively conversation intercepted from a mobile telephone. Different speaking styles may result in differences in the acoustic properties of the voice, and different channels have different effects on the acoustic properties of the recording of the voice. So that the effect of any such mismatch is balanced in both the numerator and denominator of the likelihood ratio, the recordings in the background sample should be made using the same speaking style and on the same channel as that of the suspect recording. We are not the first to make this point nor is the principle unique to forensic voice comparison, see for example Alexander & Drygajlo [8] and González-Rodríguez et al. [9] in the context of forensic voice comparison and Neumann, Evett, & Skerrett [10] for parallels in the context of fingerprint comparison.

In forensic casework it is usually not difficult to determine the broad category of the transmission channel, e.g., landline versus mobile telephone, of the suspect recording. In some cases it may even be possible to utilize the same recording equipment, e.g., when the suspect recording is of a police interview. Offender recordings are often recorded under a warrant and it would be known whether the telephone number being intercepted was associated with a landline or a mobile phone.

2.2.2. Development database

A development database is used to run preliminary tests in order to optimize parameters in the suspect and background models. Another use is to calculate weights for logistic-regression calibration / fusion (Brümmer & du Preez [11]; van Leeuwen & Brümmer [12]; Pigeon et al. [13]; Brümmer et al. [14]). Optimization and calibration / fusion involve running pairs of recordings through the early stages of the forensic-voice-comparison system, where one member of each pair mimics a suspect recording and the other mimics an offender recording. The values of the output from the initial stages of the system (known as *scores*) and knowledge about whether each pair is a same-speaker or different-speaker pair are then used to calculate parameter values for suspect and background models, and for calibration / fusion. To achieve the best optimization and calibration / fusion results, the pairs of voice recordings in the development set should be representative of the relevant population and have the same channel and speaking-style conditions as the suspect and offender recordings, including any mismatches, e.g., landline versus mobile telephone (González-Rodríguez et al. [15]). Once all the parameter values are set, the system is frozen, i.e., no further changes to the system are allowed, and the system can then be tested.

2.2.3. Test database

A test database is a set of recordings which have not been used in the training and development of the system, and thus provide a fair test of how the system will perform on previously unseen data, such as the actual suspect and offender recordings from the case at trial. Pairs of recordings from the test set are run through the system, one member of the pair standing in for the offender recording and the other standing in for the suspect recording, and the validity and reliability (accuracy and precision) of the system are estimated using the values of the likelihood-ratio output and knowledge about whether each test pair is a same-speaker or different-speaker pair (see Morrison [16]). In order for these validity and reliability measures to be representative of the expected performance of the system in the case at trial, the pairs of voice recordings in the development set should be representative of the relevant population and have the same channel and speaking-style conditions as the suspect and offender recordings including any mismatches between those two recordings, e.g., landline versus mobile telephone. If the test database were sampled from some population other than the relevant population, or under different recording conditions, then the test results would not be informative as to the expected performance of the system on the actual suspect and offender samples.

2.3. What is the relevant population?

The relevant population is the population from which the true perpetrator of the crime could conceivably have come. See Aitken & Taroni ([17] pp. 274–271), Champod, Evett, & Jackson [18], Lucy ([19] pp. 129–133), and Robertson & Vignaux ([20] ch. 3) for discussion of the relevant population, primarily in the context of DNA.

In some branches of forensic science, such as DNA, the crime-scene investigator who collects a sample at the crime scene has no idea what the properties of that sample are until it is analyzed in the laboratory. The crime-scene investigator sends samples to the laboratory as a matter of routine, not because they think the properties of the sample are similar to those of a particular suspect.

In contrast, in forensic voice comparison someone (usually a police officer who is a layperson with respect to forensic voice comparison) has listened to an audio recording of an offender and has decided that the voice on that recording sounds sufficiently similar to the voice of a particular suspect that they will send it to a forensic scientist for evaluation (typically both the offender recording and a recording of the suspect are submitted together for comparison). If the police officer thinks that the voice on the offender recording does not sound like the voice of a suspect then they do not generate the same-speaker hypothesis and do not send the recordings for comparison. This filtering of which samples to send for comparison restricts the alternative hypothesis. We argue that unless the defense proposes something more restrictive, the default defense hypothesis adopted by the forensic scientist in forensic voice comparison should therefore be the following:

> The suspect is not the speaker on the offender recording, but is someone who sounds sufficiently similar to the voice on the offender recording that a police officer (or other appropriate individual) would submit the offender and suspect recordings for forensic comparison.

That is, the voice recordings sound sufficiently similar that a layperson, such as a police officer, could generate the same-speaker hypothesis. The relevant population to sample for building the background, development, and test databases is therefore speakers who sound sufficiently similar to the voice on the offender recording that a layperson could generate the same speaker hypothesis.

Given that the suspect does sound sufficiently similar to the voice on the offender recording that the prosecution has submitted them for forensic comparison, it would be illogical to argue that the appropriate defense hypothesis should be that the suspect is a member of a population who do not (necessarily) sound sufficiently similar to the voice on the offender recording that the prosecution would submit them for forensic comparison, and that the population to be sampled for the background database should include speakers who do not sound like the offender recording.

Also, note that if the test database contained different-speaker test pairs which sounded quite different, so different that a layperson such as a police officer would not submit these pairs for forensic comparison, then the results from these pairs would likely give an overly optimistic impression of how the system would work on the actual suspect and offender pair, which do sound similar to each other.

2.4. Full consideration and full disclosure

Although we think it is by far the most common scenario, it is not necessarily the case that all pairs of suspect and offender recordings submitted for forensic analysis have been pre-filtered in the way we describe above. In each case, the forensic scientist should carefully consider the circumstances before selecting what they consider to be an appropriate defense hypothesis to adopt and an appropriate database-selection procedure. There may be circumstances in which the forensic scientist anticipates several different defense hypotheses and conducts several different analyses.

The defense hypothesis adopted by the forensic scientist and the reasoning behind it should, in all cases, be fully explained to the trier of fact so as to allow them to make appropriate decisions. Appropriate decisions include deciding on priors, or even deciding whether to accept or reject the hypothesis adopted

by the forensic scientist.

3. Database selection by human listeners

An approach to determining whether speakers are sufficiently similar sounding to the voice on the offender recording could be to have a panel of laypersons listen to the offender recording and a series of recordings of other speakers, and give judgements as to whether they are sufficiently similar sounding that they would submit them for forensic comparison.

If there is a speaking-style or channel mismatch between the offender recording and the suspect recording in the case under investigation, then the recordings for potential inclusion in the background, development, and test databases should be presented using the same speaking style and channel as the suspect recording. The listeners should then compare these with the offender recording and decide whether they are sufficiently similar sounding that they would submit them for forensic comparison. Channel mismatches etc. could cause recordings of the same speaker to sound more different than they would otherwise sound, and could cause recordings of different speakers to sound more similar than they would otherwise sound. The suspect was one speaker who sounded sufficiently similar to the voice on the offender recording given these mismatches, either despite the mismatches or because of the mismatches. The appropriate defense hypothesis should therefore be:

> The suspect is not the offender but is one member of a population of speakers who *under the same recording conditions as the suspect recording* sound sufficiently similar to *the voice on the offender recording, given its recording conditions*, that a police officer (or other appropriate listener) would submit them for forensic comparison.

Sufficiently similar sounding speakers will typically at least speak the same language and dialect as the voice on the questioned-speaker recording and be of the same gender (Rose [7] pp. 64–65), but this is not necessarily the case. Similar sounding to a layperson may not be particularly similar in terms of objective measurements of acoustic properties, and it could even be that the suspect speaks with a different accent from that on the questioned-speaker recording, or that the gender of the questioned-voice is unclear. Although gender, accent spoken, etc. may be thought of as categories, in terms of any properties measurable from voice recordings there may be gradual transitions without obvious boundaries provided by discontinuities. An intermediate stage of categorizing speakers by gender, accent spoken, etc. is not necessary to determine whether the voice samples sound sufficiently similar to the voice on the questioned-speaker recording that a layperson could generate the same-speaker hypothesis. Database selection could therefore include both male and female speakers and speakers who ostensibly speak different dialects.

We explore these issues by considering three casework examples and asking what the appropriate database selection procedure would have been given our arguments above. Two of the examples are from cases in which the authors of the present paper were involved, and the third is a case previously discussed in the forensic-voice-comparison literature.

3.1. Casework example 1: Speaking-style, and recording and transmission channel mismatch

In 2009 in Western Australia v Mansell [21], a police officer testified that she listened to a series of telephone-intercept recordings, listening to each day's recordings on a daily basis. The telephone intercepted was a mobile telephone. The police officer was subsequently part of a team conducting a search of a suspect's office. In her testimony she stated that while conducting the search she heard the voice of someone who was out of sight talking with one of her colleagues and immediately recognized it as the same as the voice on the telephone intercepts (she also stated that prior to conducting the search she believed that she would be searching the premises of the person she had been listening to on the telephone intercepts). In addition to the mobile-telephone-intercept recordings, audio recordings of the suspect talking during the search and on subsequent occasions were available, thus it would have been possible to conduct a forensic voice comparison; however, the prosecution did not submit the recordings to a forensic scientist for analysis. A couple of weeks' before the trial, defense counsel approached the first author of the present paper. There was not time to conduct a forensic voice comparison before the trial, but the author provided expert testimony consisting of a review of the literature on the degree of validity of speaker identification by lay listeners (Morrison [5] §99.1040-99.1110). On the basis of all the evidence presented at trial (of which the voice evidence was only one part), the jury found the accused not guilty.

How might background, development, and test databases have been selected in this case had a forensic voice comparison been conducted? There was no dispute between the prosecution and the defense that the speaker on the questioned-voice recording was an adult male speaking Australian English. The wider set of voice recordings from which the panel of listeners could select recordings for inclusion in the background database could therefore consist of recordings of adult male Australian-English speakers. Ideally, the panel of listeners should listen to the mobile-telephone intercept questioned-voice recordings over a number of days, reflecting the way the police officer listened to them. The panel of listeners should subsequently be presented with the recordings for potential inclusion in the background database. Since the police officer heard the suspect speaking in person, not via a transmission

channel, the recordings should be presented to the panel of listeners as high-quality audio, if possible including a simulation of the room reverberation and any background noise conditions under which the police officer made her same-speaker identification (much of this information could potentially be derived from the recording made of the suspect at the time of the search and a reconstruction of the exact location of the suspect and the police officer at the time). Since at the time the police officer made her same-speaker identification the suspect was answering questions asked by one of her colleagues, the speaking style in the recordings presented to the panel of listeners should be of speakers responding to questions. As far as we know, the police officer was from Western Australia with no training in forensic voice comparison or any other relevant training (the latter was stated in her testimony), therefore the panel of listeners should likewise be from Western Australia and be lay persons with respect to forensic voice comparison.

3.2. Casework example 2: Male or female?

In 2009 the Forensic Acoustics and Audiovisual Section of the Central Forensic Science Laboratory of the Chilean Investigative Police worked on a case in which the questioned-voice recordings came from telephone intercepts of several conversations between two people. The prosecutor asserted that the accused was one of the speakers on these recordings. The laboratory worked in the likelihood-ratio framework using commercially available software (BATVOX <http://agnitio.es/>), and the casework was typical of forensic-voice-comparison analyses conducted by the laboratory, except that the gender of the questioned speaker was not clear.

The suspect was an adult male who had, what was for a male, a high pitched voice. The pitch of the suspect's voice was superficially similar to that of the voice on the offender recording. The forensic team (which included the second author of the present paper) discussed several alternatives as to what would be the appropriate defense hypothesis and therefore what would be the relevant population from which to select a background sample. Should the background sample consist of (1) only males with high pitch voices, (2) males irrespective of the pitch of their voice, or (3) a mixture of males and females?

With respect to (1), the team arrived at the conclusion that the resulting likelihood ratio would underestimate the atypicality of the voice of the suspect. By only comparing the high-pitched offender recording against other males who share the same high-pitched voice characteristics as the suspect, the size of the likelihood ratio would be smaller. Option (3) was also discarded. Why should females be included in the background database if the suspect is male? How could the strength of evidence be calculated and expressed if the genders were mixed? An atypically high pitch for an adult male could be a relatively typical pitch for an adult female. The decision was made to go with option (2) and the background database used consisted of adult males with no pitch criterion involved in their selection. The corresponding defense hypothesis was thought to be the most appropriate hypothesis for this case because it allowed the atypicality of the suspect's voice to be reflected in the resulting likelihood ratio. If the suspect has a voice which is atypical in the same way as the voice on the offender recording, the background database should be such that the fact of this atypicality leads to a smaller denominator for the likelihood ratio and therefore a larger likelihood ratio and greater support for the same-speaker hypothesis.

There were at least two errors in this reasoning which led to an inappropriate background database being used in this case. The first error was selecting the background sample on the basis of the voice characteristics of the suspect. The speaker on the offender recording is unknown, therefore except where gender can be gleaned from the offender recording itself or from a stipulation by the defense, the gender of the speaker on the offender recording is unknown (the argument extends to class information in general). In most forensic-voice-comparison cases the gender of the offender is not in dispute, it is obviously a male or obviously a female and the defense stipulates to the gender of the offender, hence the defense hypothesis is some variant of "the voice on the questioned-voice recording is not that of the suspect but of some other speaker of the same gender". In the case under consideration, however, there was no reason to suppose that the offender was a male just because the accused was a male. In fact, one would be more likely to get the fundamental frequency measured for the offender recording if the speaker were a female than if the speaker were a male.

The second error was the focus on gender as a category. What matters for the selection of voice recordings to include in the background database is whether the voices on those recordings sound sufficiently similar to the voice on the offender recording that the panel of listeners (in stead of the original listener) would deem it appropriate to submit them for forensic analysis. The category of the speakers' gender is irrelevant, the gender of the offender is unknown and the original listener could have generated the same-speaker hypothesis whether the offender was male or female. An appropriate background database selected by a panel of listeners could potentially have included both adult females and adult males with high pitched voices.

3.3. Casework example 3: Accent

Labov & Harris [22] describe a case which went to trial in 1985. Executives of Pan American Airlines based in Los Angeles received a number of telephone calls in which bomb threats were made. The executives thought they recognized the voice as that of a Pan Am cargo handler, Paul Prinzivalli. Recordings of the bomb threats were available and recordings of Prinzivalli were

made.

A spectrographic analysis of the recordings was made by Sandra Disner and Peter Ladefoged, phoneticians based at the University of California Los Angeles. See Gruber & Poza [23], Meuwly ([24] pp. 85–112), and Morrison ([5] §99.680–99.690), and references cited therein, for descriptions of the spectrographic approach and the controversy surrounding its use. There is no indication in Labov & Harris [22] that Disner and Ladefoged made use of anything akin to background, development, and test databases.

Prinzivalli was from New York City, and William Labov of the University of Pennsylvania, an expert in dialectology including New York City accents, was also asked to analyze the recordings. Labov very quickly came to the conclusion that whereas in the recordings of the suspect the speaker was speaking with a New York City accent, in the recordings of the offender the speaker was speaking with a Boston area accent. He proceeded to document the pronunciation features which differed between the suspect and offender recordings and how the pronunciation on the offender recordings had features typical of Boston but atypical of New York City and vice versa for the suspect recording. The trial was heard by a judge alone. Labov presented his evidence. The judge then asked the prosecutor whether in light of this evidence he wanted to withdraw the charges, which he did not. Disner and Ladefoged's spectrographic evidence was then presented, followed by the prosecutor's summation. Without hearing the summation prepared by the defense attorney, the judge found the defendant not guilty.

How would we select background, development, and test databases in such a case? The databases should be selected as voice recordings which to the panel of listeners (in stead of the listener who generated the original same-speaker hypothesis) sound sufficiently similar to the voice on the questioned-voice recording that they would submit them for forensic comparison. Importantly, the panel of listeners should be selected to be similar to the listener who generated the original same-speaker hypothesis in the following respect: They should be from Southern California and not familiar with North East US accents such that it may be that they do not correctly identify or distinguish Boston and New York City accents. The panel of listeners should be presented with recordings of speakers some of whom speak with Boston accents and others who speak with New York City accents and potentially other North East US accents and accents from elsewhere. The category of accent spoken per se is not relevant, the original listeners did not perceive the differences between Boston and New York City accents, and the panel of listeners may select a mixture of recordings with Boston accents and recordings with New York City accents. The defense hypothesis adopted is:

> The speaker on the questioned-voice recording is not the suspect but is a member of a population of speakers *whom to Southern California listeners unfamiliar with North East US accents* sound sufficiently similar to the voice on the offender recording that they would submit recordings of these speakers for forensic comparison.

Why should the type of accent mismatch described in this casework example be treated differently from the speaking-style or channel mismatch discussed in casework example 1? Why should the background database not consist of recordings of speakers with New York City accents? First, selecting New York City accented voices for the background database would be the error of conditioning on the suspect rather than the offender (see discussion in casework example 2). Second, whereas in a channel mismatch the listener who generates the same-speaker hypothesis is aware of the mismatch (we are all aware that people sound different on the telephone than how they sound face to face) and still generates the same-speaker hypothesis, in the accent mismatch considered here the listener who generated the same-speaker hypothesis was not aware of the fact that the speaker(s) on the suspect and offender recordings were speaking with different accents. Had the listener said that they could hear that the voices on the two recordings had different accents but that they still thought that it was the same speaker (presumably a bidialectal speaker), then it would be appropriate to treat the accent mismatch in the same way as a channel mismatch and build the background database from recordings of speakers speaking with New York City accents, and build the development and test databases from recordings of bidialectal speakers who produce at least one New York City accented recording and at least one Boston accented recording. As described in Labov & Harris [22], however, there is no indication that the Southern California listeners who generated the same-speaker hypothesis were aware of the accent mismatch.

4. Some potential objections

There are some potential objections which could be raised regarding the human-listener selection procedure we propose.

One objection could be that the procedure is subjective and database selection based on objective categories such as gender and accent would be better. Our whole procedure is based, however, on the fact that the decision as to whether or not to submit the suspect and offender recordings for forensic comparison was subjective. Also, as discussed in the examples above, the subjective impression of the submitter with respect to the similarity of the suspect and offender recordings may differ substantially from objective class information related to those recordings, and the true class information for the offender recording may be unknown.

Another potential objection is that the fact that the suspect sounds similar to the voice on the offender is itself evidence

whose strength should be assessed. One could simply use databases representative of a general population and let the degree of shared atypicality of the suspect and offender recordings with respect to the general population be reflected in the resulting likelihood ratio. Of course the circumstances of the case would still have to be considered and the relevant general population would still have to be defined, for example, we think it unlikely that anyone would seriously propose that the databases include speakers who do not speak the language(s) spoken on the suspect and offender recordings. Alternatively, one could potentially use the human-listener selection procedure to assess the strength of this evidence with respect to the fact that the suspect sounds similar to the voice on the offender recording. If the initial database of recordings from which recordings are selected for inclusion in the background, development, and test databases were representative of a general population of speakers, then the proportion of speakers in the initial database selected for inclusion in the background, development, and test databases could form the basis of a likelihood ratio calculation. We are not convinced of the absolute necessity of the forensic scientist calculating a strength-of-evidence with respect to the fact that the suspect sounds similar to the voice on the offender recording (as opposed to letting the trier of fact ponder this), but we think that the latter solution would be preferable to the former solution which would effectively imply random partitions of the initial larger database into the background, development, and test databases. In the case of the test set the this would lead to the inclusion of recordings which do not sound sufficiently similar that they would be submitted for forensic comparison, likely leading to overly optimistic test results. Also, as we demonstrate below, there is empirical evidence to indicate that database selection for background and development sets improves the performance of a forensic-voice-comparison system.

Another potential objection, related to the previous potential objection, is that it would be difficult to combine different likelihood-ratio strength-of-evidence statements from different pieces of forensic evidence if each assumes a different relevant population. It could be for example that the relevant population adopted for the DNA analysis is any person in a specified geographical area, while for the voice analysis it is any person within that geographical area who sounds sufficiently similar to the offender. Assuming that these different source-level likelihood-ratios can in fact be combined, to make them commensurate we would also have to alter the more general defense hypothesis from the DNA analysis to the more restrictive defense hypothesis from the forensic voice comparison. This would allow the trier of fact to adopt prior odds consistent with both the likelihood ratios presented. If there were no correlation between the different data types, then this would not alter the likelihood ratio calculated for the DNA evidence, but if there were correlation then a different background database would need to be used for the DNA analysis, e.g., a DNA database sampled from the same speakers as used for the forensic voice comparison. We have probably now entered a high impractical realm. In some cases it may be appropriate to apply our database-selection procedure jointly to intrinsically linked evidence, e.g., if the task is to compare an audio-video of the offender with a suspect, we may need to select audio-video recordings of people who both look and sound sufficiently similar to the person on the offender video that they would be sent for forensic analysis, but again the practicality is questionable. We must confess that our concern in this paper is with calculating an appropriate likelihood ratio for a single piece of evidence, voice evidence, given the circumstances of the case, and we do not offer a practical solution as to how to combine this with likelihood ratios from other pieces of evidence, although the point may be moot given that at present triers of fact seldom apply mathematical implementation of Bayes' theorem.

There may be some confusion between our database-selection proposal for forensic voice comparison and the database-search problem in DNA (see Biedermann et al. [25]). We think, however, that the two are quite different. In the database-search problem a suspect may be initially targeted solely because they are in the database and have a DNA profile matching the offender DNA profile. Our database-selection procedure for forensic voice comparison is not a database search for a potential suspect, but a procedure for selecting an appropriate sample of the relevant population given logical arguments with respect to defining the relevant population. We do not propose selecting a person as a suspect simply because they are in the database and have a similar sounding voice. Rather we would expect that the suspect has come to the attention of the investigators for some other reason, and the voice evidence is essentially independent of that reason. Also, in searching for DNA profiles, the basis of the search is the same measurements as will be used to ultimately calculate a likelihood ratio, whereas this is not the case in our database-selection procedure – the acoustic properties of the recordings could be quite dissimilar compared to their subjective perceived similarity. Our procedure would not be applicable to DNA profiles because the decision to send them for forensic comparison is not based on a subjective impression as to whether they are similar, the person submitting them for analysis has no idea what their properties are. Also, if DNA profiles are treated as discrete with no variability at the source then persons in the database who do not have matching DNA profiles can be definitively excluded as being the offender, thus reducing the size of the population of people who could conceivably have committed the crime, i.e., reducing the size of the relevant population. This is not the case in our database-selection procedure for forensic voice comparison. First we assume that the offender is not in fact included in our larger initial database,

and second even if they were, by chance, included in the initial database and not selected for inclusion in the background, development, and test databases this would not exclude them as being the offender (we work with anonymized databases and do not know the identity of individuals in the database in any case).

5. Automatic database selection

5.1. Selection of offender-similar background, development, and test databases

The human-listener procedure suggested above would be potentially expensive to investigate in research because of the need to test many mock offender recordings and find appropriate databases for each mock offender recording. It would be cheaper to implement in casework, where the number of offender recordings is limited. We plan to work on the human-listener procedure in future research, but in the present paper, as a preliminary, we consider a less ideal but cheaper automatic substitute. In place of human listeners, we substitute whether an automatic forensic-voice-comparison system "thinks" that voice recordings sound similar to the voice on the offender recording. As described in greater detail below, we use an initial forensic-voice-comparison system trained on a large diverse background database to rank recordings in a second diverse database in terms of their similarity with the offender sample. We then use the top-ranked recordings, those which this procedure determined to be most similar to the offender recording, as the sample of the relevant population. These are distributed to the background, development, and test databases.

Our procedure for selecting the relevant population is almost identical to that in the German Federal Police Office's speaker recognition system (Bundeskriminalamt's, BKA's, SPrecher-Erkennungs-System, SPES, Becker et al., 2010). We differ from Becker et al. [26], however, in that whereas they select recordings which are similar to the suspect recording, we select recordings which are similar to the offender recording, and whereas they select a background database only, we select background, development, and test databases.

We think that our reasons for selecting development and test databases in addition to a background database are sufficiently explained above (note that Becker et al. [26] preselected model parameters and did not perform calibration / fusion and hence did not need to make use of a development database). We condition on the offender rather than the suspect because this is the standard approach within forensic science (Aitken & Taroni [17] pp. 274–271; Champod, Evett, & Jackson [18]; Lucy [19] pp. 129–133; Robertson & Vignaux [20] ch. 3). Since we do not know the identity of the offender and in particular we cannot assume that the suspect is the offender, we cannot assume that anything we know about the suspect is also true for the offender. Some recordings which are similar to the suspect recording may be relatively dissimilar to the offender recording and could be sufficiently dissimilar that they would not be submitted for forensic evaluation. Including such recordings in the test database may result in an overly optimistic assessment of the performance of the system on the actual suspect and offender recordings.

5.2. Contrast with cohort selection in automatic speaker recognition

Database selection in automatic speaker recognition usually goes under the name *cohort selection*, and tends to have different motivations than those which we present for forensic voice comparison. Motivations for the former are primarily telic (better system performance) rather than philosophical.

In the *fast scoring method* applied to Gaussian-mixture models only the *n* Gaussians in the target and background models which are closest to the known-speaker or questioned-speaker recording are used for calculating a score (Reynolds [27]; Auckenthaler, Carey, & Lloyd-Thomas [28]; Ramos-vcastro et al. [29]; Reynolds, Quatieri, & Dunn [30]; Kinnunen & Li [31]; there appears to be some variability or ambiguity as to whether closeness is with respect to the known- or questioned-speaker recording). The fast scoring method does not select speakers to include in the training of the background model, but the selection of Gaussians does alter the background model used to calculate the denominator of the likelihood ratio. As the word "fast" in the name of the method suggests, the primary goal is computational efficiency, and it is based on the assumption that the more distant Gaussians would have contributed little to the probability density in the vicinity of the claimant recording so the answer would be approximately the same, at least in terms of rank order which is sufficient for making a binary accept or reject decision. The motivations are not based on calculating a likelihood ratio as an interpretable strength-of-evidence statement in response to a particular question defined by particular hypotheses.

A general class of *score normalization* procedures (including test-normalization, T-norm, and zero-normalization, Z-norm) is dependent on cohort selection and ameliorates problems due to potential mismatches in channel etc. between test data and training data used for background and known-speaker models. The score obtained for the questioned-speaker recording is adjusted on the basis of scores obtained for recordings from a cohort where the members of the cohort are selected on the basis of their similarity to either the known- or questioned-speaker recording (Rosenberg et al. [32]; Reynolds [27]; Auckenthaler, Carey, & Lloyd-Thomas [28]; Ramos Castro [33]; Kinnunen & Li, [31]). Similarity can be measured using a number of metrics including the Bhattacharyya distance (Campbell [34]) and the Kullback-Leibler divergence (Hasan & Hansen [35]; Ramos Castro [33]). The ultimate motivation is to obtain better system performance, typically on a hard-thresholded posterior-

probability metric such as equal error rate (EER) or the detection error threshold cost (C_{DET}), rather than to produce a likelihood ratio as an interpretable strength-of-evidence statement in response to a particular question defined by particular hypotheses. The composition of databases in automatic speaker recognition is typically much less controlled than what we recommend for forensic voice comparison.

Reynolds [27] employed a cohort-selection procedure for selecting recordings for inclusion in a background database. His procedure was generally similar to our automatic procedure (described in greater detail below), but his metric of similarity was a cross-metric combining the probability of data A given model A versus given model B and the probability of data B given model B versus given model A. He found that the speaker-dependent background model outperformed a speaker-independent background model when the known-speaker model was built from scratch, but that the speaker-independent background model outperformed both when the known-speaker model was adapted from the speaker-independent background model. He did not test the combination of a speaker-dependent background model and adaptation of the known-speaker model from the speaker-dependent background model, which is the approach which we adopt in the present paper.

Similarity metrics applied to select cohorts in automatic speaker recognition could also potentially be applied to database selection in forensic voice comparison, although in the present paper we only test one automatic procedure.

Hasan & Hansen [35] made philosophical arguments backed up by empirical results with respect to appropriate Gaussian mixture model - universal background model (GMM-UBM) background database selection for automatic speaker verification leading to equal or better performance with very large improvements in computational efficiency. The appropriate aim in that context, to efficiently represent the whole range of variability in a large heterogeneous population, given that a potential questioned speaker could be almost anyone and that the system has to deal with many different questioned speakers, is almost the opposite of our aim which is to best represent the relatively small relatively homogeneous population of speakers who are subjectively similar-sounding to a single questioned-speaker recording.

With respect to automatic speaker verification using support-vector machines (SVC), a discriminative approach, only the support vectors from the training data rather than the distribution of the data are used for classification. A number of automatic-speaker-verification studies, including McLaren et al. [36], Suh et al. [37], and Zhang, Shan, & Liu [38], have looked at background database selection for SVC. A background database which includes speakers who are more similar to the questioned-speaker will likely result in a set of support vectors which result in better classification results. Zhang, Shan, & Liu [38] used a vocal-tract length estimate extracted from the acoustic signal to partition the background speakers into subgroups and build a series of background models, and subsequently applied procedures to select the best background model for each questioned-speaker recording. Although the Zhang et al. approach differs in many details from our approach, their approach being discriminative with a hard classification objective whereas ours is generative with a likelihood-ratio objective, it shares the technique of trial by trial selection of background data on the basis of similarity to the questioned-speaker recording.

5.3. Outline of empirical tests of the automatic database-selection procedure in the remainder of the paper

In the remainder of the paper we present two experiments. Detailed descriptions of our automatic database-selection procedure appear in the methodology sections for those experiments. Experiment I is an analysis of which recordings our initial forensic-voice-comparison system "thinks" are similar to the mock offender recordings. We test a database which is diverse in terms of the speakers' first languages, language spoken, and transmission channel. Although we have argued that such categories are irrelevant for database selection, we would expect them to be correlated with similarity, and whether the highest ranked recordings have the same first language, language, spoken, and channel as the offender recording can serve as a diagnostic of similarity. Ultimately we are interested in accounting for more subtle aspects of similarity, but those would be inherently difficult to quantify, and the use of this database provides a convenient analysis tool. Experiment II applies the database selection procedure to select background, development, and test databases, and compares a forensic-voice-comparison system's performance on the test database when the background and development databases are instead selected randomly from the initial database.

6. Experiment I - Analysis of the recordings selected by the automatic procedure

6.1. Methodology

6.1.1. Forensic-voice-comparison system

The forensic-voice-comparison system was a basic automatic system: 16 mel-frequency-cepstral-coefficient (MFCC) values were extracted every 10 ms over the entire speech-active portion of each recording using a 20 ms wide hamming window. Delta coefficient values were also calculated and included in the subsequent statistical modeling (Furui [39]). Feature warping was applied (Pelecanos & Sridharan [40]) to both coefficients and deltas. A GMM-UBM (Reynold, Quatieri, & Dunn [30]) was built using the initial background data to train the background model. The number of Gaussians in the mixture was 128.

6.1.2. Data

The initial background database, $D_{initial}$, used for training the initial UBM consisted of 400 recordings from the NIST 08 database (short2-short3, [41]). We only use this initial background database in preliminary tests of how we might refine the selection of more relevant background, development, and test databases ($D_{background}$, $D_{development}$, D_{test}).

For testing purposes we used an extremely diverse database, $D_{diverse}$ (NIST 06 1conv4w, [42]), including speakers with different first languages speaking different languages on different recording channels (all speakers we selected were male) ($D_{diverse} \cap D_{initial} = 0$). The largest groups of speaker by first-language (L1) were US-English speakers and speakers of Standard Chinese (Mandarin), see Table 1. The L1-US-English speakers were only recorded speaking English but the L1-Standard-Chinese speakers produced recordings in Chinese and in English. Other L1s in the database were (alphabetically) Australian English, Bengali, Cantonese, Farsi, Gbe, Hindi, Japanese, Korean, Russian, Spanish, Tagalog, Urdu, Vietnamese, and Wu. Most of these speakers only produced recordings in English, but some Bengali, Cantonese, Farsi, Hindi, and Russian speakers also produced recordings in their L1. After voice activity detection, recordings ranged in length from 15 to 120 seconds with a mean of 85 seconds. This diverse database is a source from which we will select recordings for inclusion in the relevant background, development, and test databases ($\{D_{background}, D_{development}, D_{test}\} \in D_{initial}$; $D_{background} \cap D_{development} = 0$; $D_{background} \cap D_{test} = 0$; $D_{development} \cap D_{test} = 0$).

For a number of speakers (25 L1-US-English speakers and 46 speakers of other L1s), there were multiple recording sessions on different transmission channels: mobile phone and landline (for our experiments we have conflated landline cordless and landline corded).

Table 1: First language and language spoken in the test database. Number of recordings (number of speakers).

First Language	Language Spoken		
	English	Standard Chinese	Other
US English	171 (46)	–	–
Standard Chinese	184 (87)	183 (83)	–
Other	182 (74)	–	65 (35)

6.1.3. Procedures

A model was built for each recording in $D_{diverse}$ adapted from the initial UBM via the maximum a posteriori probability (MAP) procedure (Reynold, Quatieri, & Dunn [30]).

A series of mock-offender recordings were selected from $D_{diverse}$. These recordings fulfilled the requirements that they were a produced in English by L1-US-English speakers who produced at least one mobile-telephone recording and at least one landline-telephone recording in English. The number of speakers fulfilling these criteria was 25 (for speakers of other L1s the proportion fulfilling the mobile-plus-landline criterion in either their L1 or in English was smaller and for simplicity we do not explore these here). One mobile and all the landline recordings were selected from each of the mock-offender speakers.

For each of the mock offenders, all of the recordings (mobile and landline) from that speaker were removed from $D_{diverse}$ and each mock offender recording, $R_{offender}$, was then compared with all the remaining recordings in $D_{diverse}$ and a score for each of those comparisons calculated. There were two methods of comparison: (1) $R_{offender}$ was used to train a speaker model, and each of the remaining recordings in $D_{diverse}$ was used as probe data, see Equation 1. (2) Each of the remaining recordings in $D_{diverse}$ was used to train a speaker model, and $R_{offender}$ was used as probe data, see Equation 2. In both cases a score was calculated as the mean of the logarithm of the ratio of the probability-density-function values of the speaker model versus the UBM for each feature vector in the probe set, see Equations 1 and 2.

$$S_{diverse,m} = \frac{1}{N_m}\sum_{n=1}^{N_m}\log\left(\frac{p\left(X_{diverse,m,n}\,|\,M_{offender}\right)}{p\left(X_{diverse,m,n}\,|\,M_{initial}\right)}\right) \quad (1)$$

$$S_{diverse,m} = \frac{1}{N}\sum_{n=1}^{N}\log\left(\frac{p\left(X_{offender,n}\,|\,M_{diverse,m}\right)}{p\left(X_{offender,n}\,|\,M_{initial}\right)}\right) \quad (2)$$

Where $S_{diverse,m}$ is the score for recording m in $D_{diverse}$, $X_{diverse,m,n}$ is the feature vector (MFCC coefficients plus deltas) for frame n in recording m in $D_{diverse}$ which is N_m frames long, $X_{offendern}$ is the feature vector (MFCC coefficients plus deltas) for frame n in $R_{offender}$ which is N frames long, $M_{offender}$ is the model trained on $R_{offender}$, $M_{diverse,m}$ is the model trained on recording m in $D_{diverse}$, and $M_{initial}$ is the model trained on $D_{initial}$. $p(X|M)$ is the probability-density-function value of model M evaluated at values X.

The tests were repeated, once with feature warping and once without. The pattern in the results was stronger when $R_{offender}$ was used as probe data (Equation 2) rather than to build a speaker model (Equation 1), and also when feature warping was used. For simplicity, only these results are reported below.

6.2. Results and Discussion

The results were rather complicated, given the possible interactions between L1, language spoken, and channel. We

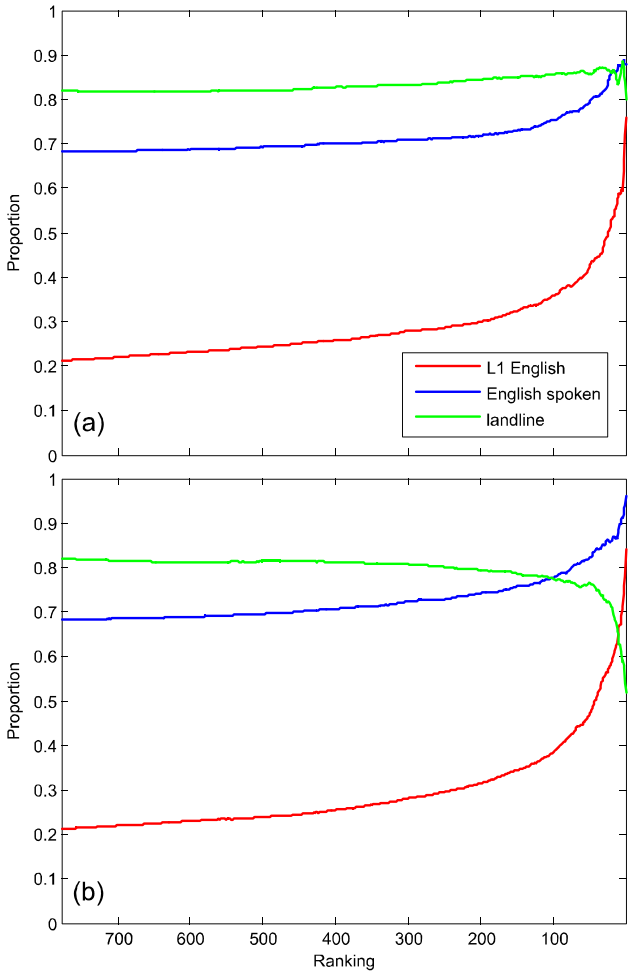

Figure 1: Main effects: Proportion of top ranked scores which are from L1-US-English speakers (red), where the language spoken is English (blue), and where the channel is a landline (green), when the mock-offender recording is of an L1-US-English speaker speaking English and is used as probe data, and is recorded on a landline (a) or a mobile telephone (b).

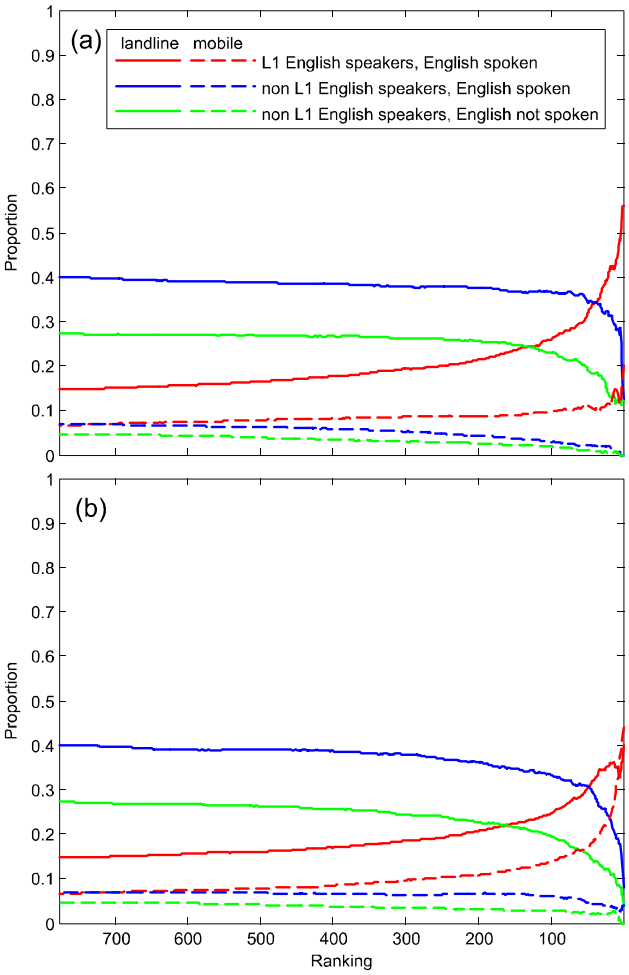

Figure 2: Interactions: Proportion of top ranked scores which are from L1-US-English speakers speaking English on a landline (solid red), L1-US-English speakers speaking English on a mobile telephone (dashed red), non-L1-US-English speakers speaking English on a landline (solid blue), non-L1-US-English speakers speaking English on a mobile telephone (dashed blue), non-L1-US-English speakers speaking a language other than US English on a landline (solid green), non-L1-US-English speakers speaking a language other than US English on a mobile telephone (dashed green), when the mock-offender recording is of an L1-US-English speaker speaking English and is recorded on a landline (a) or a mobile telephone (b).

proceed by describing only the subset of the main effects and the interactions of interest. Results were averaged across the 25 L1-US-English mock-offender speakers.

Figures 1 and 2 display results from when the R_{offender} were landline recordings, panel (a), and when they were mobile-telephone recordings, panel (b). For R_{offender} the scores from the comparison of this recording with each of the remaining speakers in D_{diverse} were ranked. In each panel in the figures, the x axis indicates the x top-ranked scores. The left edge of the x axis includes the scores from all remaining recordings in D_{diverse}, the point marked 700 includes the 700-top ranked scores, etc., and the right edge of the x axis includes only the single top-ranked score. The y axis represents the proportion of recordings from the remaining recordings in D_{diverse} with the indicated category (L1-US-English speakers, speakers speaking on a landline, etc.) included in the x top-ranked scores. The proportions at each rank are averaged across the results from the 25 mock offender speakers.

Figure 1 shows the main effects, i.e., looking at L1, language spoken, and channel without considering possible interactions between these categories.

– **L1** (red lines): The proportion of L1-US-English speaker recordings in the entire database is 0.21, but this rises rapidly towards the top of the ranking. 0.36–0.39 of the top 100 ranked recordings are of L1-US-English speakers.

– **Language spoken** (blue lines): The proportion of recordings which are in English also increases with the rank of the score: The proportion of recordings which are in English in the entire database is 0.68 and rises to 0.75–0.77 of the top 100 ranked recordings.
– **Channel** (green lines): The proportion of recordings on a landline recording in the entire database was 0.82. This rose to 0.86 of the top 100 ranked recordings when the mock offender was talking on a landline, and dropped to 0.77 of the top 100 ranked recordings when the mock offender was talking on a mobile telephone.

There were strong effects for L1 and language spoken, especially strong for L1, and weaker effects for channel. The relatively weak results for channel, especially when the mock offender was speaking on a landline, were due to an interaction between channel and language as will be illustrated below.

Figure 2 shows the three-way interactions between L1, language spoken, and channel.

– **L1-US-English speakers** (red lines): Irrespective of channel, the proportion of recordings of L1-US-English speakers speaking English increases as the rank of the recordings increases. The increase is greater when the channel is the same as that of the mock offender recording, solid red line in panel (a) and dashed red line in panel (b).
– **non-L1-US-English speakers** (blue and green lines): Irrespective of channel and irrespective of language spoken, the proportion of recordings of non-L1-US-English speakers decreases as the rank increases.

A relatively large effect for channel condition is apparent when the channel by L1 interaction is taken into account, particularly when the speakers are L1-US-English speakers.

The results therefore indicate that, using the automatic procedure, L1 is the most important factor in selecting recordings similar to the offender recording. Speaking English per se was not a particularly important factor, even when the non-L1-US-English speakers were speaking English the proportion of their recordings in the highest ranked recordings went down. We assume that the non-L1-US-English speakers spoke with accents which differed more from the mock offenders' accents than did those of other L1-US-English speakers, and that this is therefore an accent effect. The effect for channel was less than the language-plus-accent effect.

There have been claims (see Jessen [43] p. 700) that automatic systems using non-language non-accent specific background databases can be used for forensic voice comparison, at the extreme even when the language spoken in the suspect and offender recordings is unknown. These results should serve as warning that language and accent are important and hence one must collect data from speakers of the relevant language and accent. We would recommend that end users of commercially-produced forensic-voice-comparison software or of opinions proffered by forensic-voice-comparison experts be highly skeptical of any claims made as to language independence and demand demonstration of the degree of validity and reliability of proffered systems system under conditions reflecting those of the case at trial.

7. Experiment II - The effect of the automatic database-selection procedure on the performance of a forensic-voice-comparison system

We simulate a situation in which the offender recording is a mobile telephone recording and the suspect recording is a landline recording. In forensic voice comparison telephone intercept recordings are common and it is usually known whether the intercepted telephone is a landline or a mobile telephone.

7.1. Methodology

The same data and forensic-voice-comparison system used in Experiment I were also used in Experiment II.

7.1.1. Procedure

For the first mock offender a mobile telephone recording, R_{offender}, was selected. The remaining recordings from this speaker were removed from the diverse database, D_{diverse}, and the procedure for ranking the similarity of recordings described above was then applied to the remaining *landline recordings* in D_{diverse}. These recordings were than ranked according to their scores. Recordings from the 45 speakers with the highest ranked recordings were then divided among the background, development, and test databases ($D_{\text{background}}$, $D_{\text{development}}$, D_{test}) as follows:

– The speaker who produced the top-ranked recording was identified and all the *landline recordings* produced by this speaker were included in $D_{\text{background}}$ and removed from the ranked set of recordings.
– In the remaining ranked recordings, the speaker who produced the top-ranked recording was identified and all the *landline and mobile recordings* produced by this speaker were included in $D_{\text{development}}$ and removed from the ranked set of recordings.
– In the remaining ranked recordings, the speaker who produced the top-ranked recording was identified and all the landline recordings produced by this speaker were included in D_{test} and removed from the ranked set of recordings.

These three steps were repeated until each of the three databases, $D_{\text{background}}$, $D_{\text{development}}$, D_{test}, contained recordings from 15 speakers. All the *landline recordings* produced by the mock offender were also included in D_{test}, hence D_{test} included same-speaker as well as different-speaker comparison pairs. 15 speakers per database is small, but there were only 46 L1-US-

English speakers in D_{diverse} including the mock suspect. This database size allowed for the possibility that the database-selection procedure would pick only L1-US-English speakers.

$D_{\text{background}}$ and $D_{\text{development}}$ were used to calculate scores comparing each mobile recording in $D_{\text{development}}$ with each landline recording in $D_{\text{development}}$. This included some same-speaker comparisons and a larger number of different-speaker comparisons. In order to ensure a reasonable number of mobile recordings in $D_{\text{development}}$, if the number of speakers in $D_{\text{development}}$ who had both landline and mobile recordings was less than 8, a speaker in $D_{\text{background}}$ or D_{test} who had both types of recordings was randomly selected and all that speaker's recordings moved to $D_{\text{development}}$, and a speaker in $D_{\text{development}}$ who only had landline recordings was randomly selected and all that speaker's recordings moved to replace those removed from $D_{\text{background}}$ or D_{test}. The procedure was repeated until there were 8 speakers in $D_{\text{development}}$ who had both landline and mobile recordings.

The scores from the development set were then used to calculate weights for logistic-regression calibration (Brümmer & du Preez [11]; van Leeuwen & Brümmer, 2007). Calculations were performed using Brümmer [44] and Morrison [45], and using the pooled procedure (Morrison, Thiruvaran, & Epps [46]).

$D_{\text{background}}$ was used to train a background model ($M_{\text{background}}$) which was used to calculate scores comparing R_{offender} (a mobile-telephone recording) with every landline recording in D_{test} (mock suspect recordings), see Equation 3.

$$S_{suspect,m} = \frac{1}{N}\sum_{n=1}^{N}\log\left(\frac{p\left(X_{offender,n}\,|\,M_{suspect,m}\right)}{p\left(X_{offender,n}\,|\,M_{background}\right)}\right) \quad (3)$$

Where $S_{suspect,m}$ is the score for the comparison between the mock offender and suspect recording m from D_{test}, and $M_{suspect,m}$ is the model trained on suspect recording m from D_{test}.

These scores were then converted into likelihood ratios using the logistic-regression weights calculated using $D_{\text{development}}$ as described above.

7.2. Results and Discussion

The database-selection procedure described above was repeated for all 25 R_{offender} and the log-likelihood-ratio cost, C_{llr} (Brümmer & du Preez [11]; van Leeuwen & Brümmer [12]), and the 95% credible interval, 95% CI (using the parametric procedure, Morrison [16]; Morrison, Thiruvaran, & Epps [47]), were calculated using the likelihood-ratio results pooled over all 25 R_{offender} sets. C_{llr} and the 95% CI provide metrics of the degree of validity and reliability (accuracy and precision) of the performance of a forensic-voice-comparison system (Morrison [16]). As a measure of validity, we calculated C_{llr} using the means of the likelihood ratios in each of a number of groups. In each group each likelihood ratio was the result of comparison of a pair of recordings, in each pair the mock offender recording was always the same and the mock suspect recording was different, but all the mock suspect recordings within a group were produced by the same speaker (see Morrison [16]). This metric we designate C_{llr}*-mean*. We also calculated C_{llr}*-pooled*, calculated using the individual likelihood ratios from all comparison pairs without reference to group membership. The latter is a single-value system-performance metric which

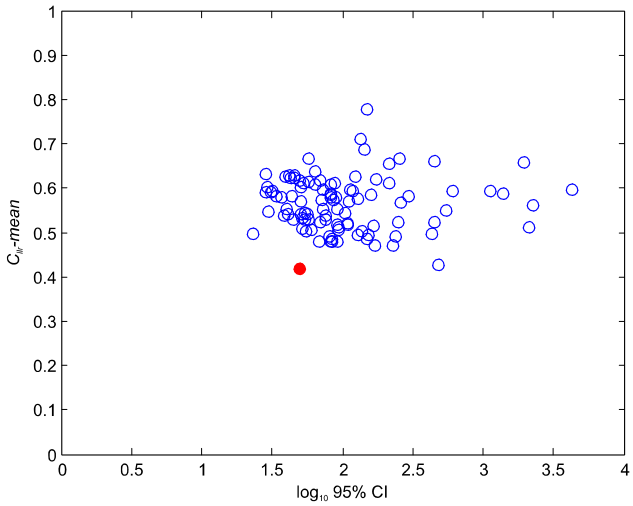

Figure 3: Comparison of the validity and reliability of the forensic-voice-comparison system when the database-selection procedure is used for the background and development databases (red filled circle) and when the speakers in the background and development databases are chosen at random (blue unfilled circles). Validity measured by C_{llr}*-mean* and reliability measured by the 95% CI in log$_{10}$ units (orders of magnitude). The *x* axis is truncated and excludes a randomization-test result with a log$_{10}$ 95% CI of 10.

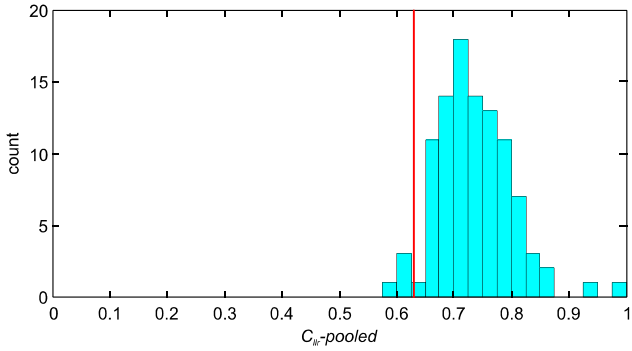

Figure 4: Comparison of system performance, in terms of C_{llr}*-pooled*, when the database-selection procedure is used for the background and development databases (vertical red line) and when the speakers in the background and development databases are chosen at random (histogram of results from 100 randomization tests).

conflates both validity and reliability.

To determine whether the $D_{background}$ and $D_{development}$ selection procedure improved the performance of the forensic-voice-comparison system a set of randomization experiments were conducted. The D_{test} for each of the 25 $R_{offender}$ were kept as they were, but $D_{background}$ and $D_{development}$ were replaced with recordings of randomly selected speakers from $D_{diverse}$ (excluding recordings of any speakers already in D_{test}). $D_{development}$ was constrained to have the same number of speakers with mobile and landline recordings and the same number of speakers with landline only recordings as had been the case when the speakers in $D_{development}$ were selected on the basis of similarity with $R_{offender}$ (for all $R_{offender}$ this was 8 speakers with both types of recordings and 7 with only landline). Likelihood-ratios were pooled over all 25 sets of results and C_{llr} and the 95% CI calculated. The randomization test was repeated 100 times.

The results of the database-selection procedure were then compared with the results of the randomization tests, see Figures 3 and 4. In the two-dimensional validity by reliability space shown in Figure 3 there was complete separation between the database-selection result and the random-selection results. For the database-selection procedure C_{llr}*-mean* was 0.421, smaller than any of the C_{llr}*-mean* values from the randomization tests, and the 95% CI was ±1.69 orders of magnitude, smaller than 81% of the 95% CI values from the randomization tests. For the database-selection procedure C_{llr}*-pooled* was 0.630, smaller than 96% of the C_{llr}*-pooled* values from the randomization tests.

A Tippett plot of the results from the database-selection procedure is provided in Figure 5. Performance was poor, but compared to random database selection, the results reported above lend very strong support to the hypothesis that the automatic database-selection procedure improves system validity and reliability.

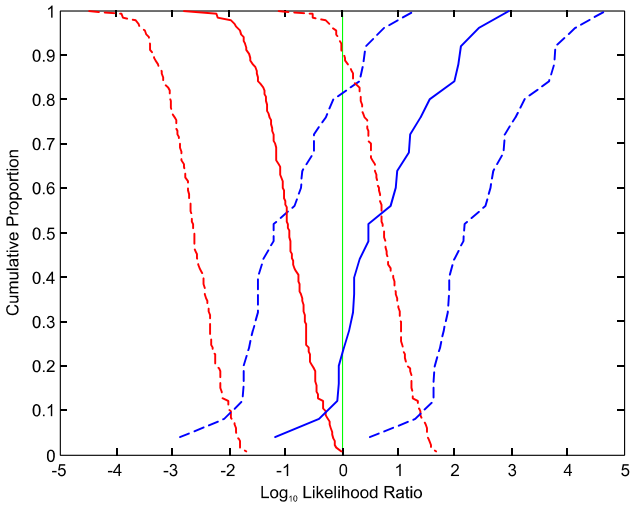

Figure 5: Tippett plot of system performance when the database-selection procedure was applied. The blue lines rising to the right represents the cumulative distribution of likelihood ratios from same-speaker comparisons, and the red lines rising to the left represents the cumulative distribution of likelihood ratios from different-speaker comparisons. Solid lines indicate group-mean values and the dashed lines to the left an right of the solid lines indicate the 95% CI.

8. Conclusion

We have presented a logical argument with respect to selecting the appropriate population to sample in order to construct background, development, and test databases for forensic voice comparison. We have argued that because suspect and offender recordings are usually only submitted for forensic comparison if they sound sufficiently similar to a layperson such as a police officer, the appropriate defense hypothesis for the forensic scientist to adopt is that the suspect is not the same speaker as on the offender recording, but is a member of a population of speakers who sound sufficiently similar to the voice on the offender recording that a layperson such as a police officer would submit them for forensic comparison. We proposed that appropriate databases could be selected by panels of listeners, and presented three casework examples to illustrate how this might work in practice. The panel of listeners should be as similar as possible in linguistic exposure to the person who originally generated the same-speaker hypothesis and should listen under conditions as similar as possible to the conditions under which the same-speaker hypothesis was originally generated. An automatic database-selection procedure was also proposed, which, although not fully consistent with the logical argument and human-listener system, could immediately be used to provide a proof-of-concept empirical test of the importance of database selection. Compared to random database selection, the automatic database-selection procedure resulted in better validity and reliability for an MFCC GMM-UBM forensic-voice-comparison system.

9. References

[1] French, J.P., Harrison, P., "Position statement concerning use of impressionistic likelihood terms in forensic speaker comparison cases", *Int. J. Speech Lang. and Law*, 14:137–144, 2007. doi:10.1558/ijsll.v14i1.137

[2] French, J.P., Nolan, F., Foulkes, P., Harrison, P. McDougall, K., "The UK position statement on forensic speaker comparison: a rejoinder to Rose and Morrison", *Int. J. Speech Lang. and Law*, 17:143–152, 2010. doi:10.1558/ijsll.v17i1.143

[3] Rose, P., and Morrison, G. S., "A response to the UK position statement on forensic speaker comparison", *Int. J. Speech Lang. and Law*, 16:139–163, 2009. doi:10.1558/ijsll.v16i1.139

[4] Morrison, G. S., "Forensic voice comparison and the paradigm shift", *Sci. and Just*., 49:298–308, 2009. doi: 10.1016/j.scijus.2009.09.002

[5] Morrison, G. S., "Forensic voice comparison", Freckelton, I. and Selby, H. (Eds.), *Expert Evidence*, Thomson Reuters, Sydney, Australia, 2010, ch 99.

[6] Champod, C. and Meuwly, D., "The inference of identity in forensic speaker recognition", *Speech Comm*. 31:193–203, 2000. doi:10.1016/S0167-6393(99)00078-3

[7] Rose, P., *Forensic Speaker Identification*, Taylor and Francis London, UK, 2002.

[8] Alexander, A. and Drygajlo, A., "Scoring and direct methods for the interpretation of evidence in forensic speaker recognition", *Proceedings of Interspeech 2004 - 8th International Conference on Spoken Language Processing (ICSLP)*, 2004, p 2397–2400.

[9] González-Rodríguez, J., Drygajlo, A., Ramos-Castro, D., García-Gomar, M., and Ortega-García, J., "Robust estimation, interpretation and assessment of likelihood ratios in forensic speaker recognition", *Comput. Speech and Lang*., 20:331–355, 2006. doi:10.1016/j.csl.2005.08.005

[10] Neumann, C., Evett, I. W., and Skerrett, J., "Quantifying the weight of evidence from a forensic fingerprint comparison: A new paradigm", *J. Royal Statist. Soc. A* 175(2): 371–415, 2012. doi:10.1111/j.1467-985X. 2011.01027.x

[11] Brümmer, N., and du Preez, J., "Application independent evaluation of speaker detection," *Comput. Speech Lang*. 20:230–275, 2006. doi:10.1016/j.csl.2005.08.001

[12] van Leeuwen, D. A., and Brümmer, N., "An introduction to application-independent evaluation of speaker recognition systems", Müller C. (Ed.), *Speaker Classification I: Selected Projects*, Springer, Heidelberg, Germany, 2007, p 330–353. doi:10.1007/ 978-3-540-74200-5_19

[13] Pigeon, S., Druyts, P., and Verlinde, P., "Applying Logistic Regression to the Fusion of the NIST'99 1-Speaker Submissions", *Digit. Sig. Process*., 10:237–248, 2000.

[14] Brümmer, N., Burget, L., Černocký, J., Glembek, O. Grézl, F. Karafiát, M. van Leeuwen, D. A.. Matějka, P. Schwarz, P., and Strasheim, A., "Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST speaker recognition evaluation 2006," *IEEE Trans. Audio, Speech Lang. Process*., 15:2072–2084, 2007. doi:10.1109/TASL.2007.902870

[15] González-Rodríguez, J., Rose, P., Ramos, D., Toledano, D. T., and Ortega-García, J., "Emulating DNA: Rigorous quantification of evidential weight in transparent and testable forensic speaker recognition", IEEE *Trans. Audio, Speech and Lang. Proc*., 15:2104–2115, 2007. doi:10.1109/TASL.2007.902747

[16] Morrison, G. S., "Measuring the validity and reliability of forensic likelihood-ratio systems",. *Sci. and Just*., 51:91–98, 2011. doi:10.1016/j.scijus.2011.03.002

[17] Aitken, C. G. G., and Taroni, F., *Statistics and the evaluation of forensic evidence for forensic scientist*, 2nd ed, Wiley, Chichester, UK, 2004.

[18] Champod, C., Evett, I. W., and Jackson, G., "Establishing the most appropriate databases for addressing source level propositions", *Sci. and Just*., 44:153–164, 2004. doi:10.1016/S1355-0306(04)71708-6

[19] Lucy, D., *Introduction to statistics for forensic scientists*, Wiley, Chichester, UK, 2005.

[20] Robertson, B. and Vignaux, G. A., *Interpreting Evidence*, Wiley, Chichester, UK, 1995.

[21] State of Western Australia v Cameron James Mansell, WA Dist Ct, No 665 of 2008.

[22] Labov, W. and Harris, W. A., "Addressing social issues through linguistic evidence", Gibbons J. (Ed.), *Language and the Law*, Longman, Harlow, UK, 1994, p 265–305.

[23] Gruber, J.S., and Poza, F., "Voicegram identification evidence", vol. 54, *American Jurisprudence Trials*, Westlaw, 1995.

[24] Meuwly, D., *Reconnaissance de locuteurs en sciences forensiques: L'apport d'une approche automatique*. PhD diss., Université de Lausanne, 2001.

[25] Biedermann, A., Gittelson, S., Taroni, F., "Recent misconceptions about the 'database search problem': A probabilistic analysis using Bayesian networks", *Forensic Sci. Int*., 212:51–60, 2011. doi:10.1016/ j.forsciint.2011.05.013

[26] Becker, T., Jessen, M., Alsbach, S., Broß, F., and Meier, T., "Automatic forensic voice comparison using recording adapted background models", *Proceedings of the 39th International Audio Engineering Society (AES) Conference – Audio Forensics: Practices and Challenges, Hillerød, Denmark*, 2010, p 162–166.

[27] Reynolds, D., "Comparison of background normalization methods for text-independent speaker verification", *Proceedings of Eurospeech 1997, Rhodes*, 1997, p 963-966.

[28] Auckenthaler C., Carey, M., and Lloyd-Thomas, H., "Score normalization for text-independent speaker verification Systems", *Digit. Sig. Process*., 10:42–54, 2000. doi:10.1006/dspr.1999.0360

[29] Ramos-Castro D., Fierrez-Aguilar J., González-Rodríguez J., and Ortega-García J., "Speaker verification using speaker- and test-dependent fast score normalization", Pattern Recog. Letters, 28:90–98, 2006. doi:10.1016/j.patrec.2006.06.008

[30] Reynolds, D. A., Quatieri, T. F., Dunn, R. B., "Speaker verification using adapted Gaussian mixture models", *Digit. Sig. Process.*, 10;19–41, 2000. doi:10.1006/dspr.1999.0361

[31] Kinnunen, T. and Li, H., An overview of text-independent speaker recognition: From features to supervectors", *Speech Comm*., 52:12–40, 2010. doi:10.1016/j.specom.2009.08.009

[32] Rosenberg A., Delong J., Lee C., Juang B., and Soong F., "The use of cohort normalized scores for speaker recognition", *Proceedings of the 2nd International Coference on Spoken Language Processing (ICSLP), Banff*, International Speech Communication Association, 1992, p 599–602.

[33] Ramos Castro D., Forensic evaluation of the evidence using automatic speaker recognition systems, PhD diss., Universidad Autónoma de Madrid, 2007.

[34] Campbell, J. P. Jr., "Speaker recognition: A tutorial", *Proc. of the IEEE*, 85:1437–1462, 1997. doi:10.1109/5.628714

[35] Hasan T. and Hansen J. L. H., "A study on universal background model training in speaker verification",. IEEE *Trans. Audio, Speech and Lang. Process*., 19:1890–1899, 2007. doi:10.1109/TASL.2010.2102753

[36] McLaren, M., Baker, B., Vogt, R., and Sridharan, S., "Improved SVM speaker verification through data-driven background dataset selection", *Proceedings of the IEEE Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2009, p 4041–4044. doi:10.1109/ICASSP.2009.4960515

[37] Suh, J.-W., Lei, Y., Kim, W., and Hansen, J. H. L., "Effective background data selection in SVM speaker recognition for unseen test environment: More is not always better", *Proceedings of the IEEE Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, p 5304–5307. doi:10.1109/ICASSP.2011.5947555

[38] Zhang, W.-Q., Shan, Y., and Liu, J., "Multiple background models for speaker verification," Cernocký, H. and Burget, L. (Eds.), *Proceedings of Odyssey 2010: The Language and Speaker Recognition Workshop, Brno*, International Speech Communication Association, 2010, pp. 47–51.

[39] Furui, S., "Speaker-independent isolated word recognition using dynamic features of speech spectrum", *IEEE Trans. Acoust., Speech and Sig. Process*. 34:52–59, 1986. doi:10.1109/TASSP.1986.1164788

[40] Pelecanos, J., and Sridharan, S., "Feature warping for robust speaker verification", *Proceedings of 2001: A Speaker Odyssey – The Speaker Recognition Workshop, Crete*, International Speech Communication Association, 2001.

[41] National Institute of Standards and Technology, *The NIST year 2008 speaker recognition evaluation plan*, 2008. http://www.itl.nist.gov/iad/mig/tests/sre/2008/sre08_evalplan_release4.pdf

[42] National Institute of Standards and Technology, *The NIST year 2006 speaker recognition evaluation plan*, 2006. http://www.itl.nist.gov/iad /mig/tests/spk/2006/sre-06_evalplan-v9.pdf

[43] Jessen M., "Forensic phonetics", Lang. and Ling. Compass, 2:671–711, 2008.. doi:10.1111/j.1749-818x.2008.00066.x

[44] Brümmer, N., *Tools for fusion and calibration of automatic speaker detection systems*, 2005. http://niko.brummer.googlepages.com/focal/

[45] Morrison, G. S., *multivar_kernel_LR: Matlab implementation of Aitken & Lucy's (2004) forensic likelihood-ratio software using multivariate-kernel-density estimation*, 2007. http://geoff-morrison.net/

[46] Morrison, G. S., Thiruvaran, T., and Epps, J., "An issue in the calculation of logistic-regression calibration and fusion weights for forensic voice comparison", Tabain, M., Fletcher, J., Grayden, D., Hajek, J., and Butcher, A. (Eds.), *Proceedings of the 13th Australasian International Conference on Speech Science and Technology, Melbourne*, Australasian Speech Science and Technology Association, 2010, p 74–77.

[47] Morrison, G. S., Thiruvaran, T., and Epps, J., "Estimating the precision of the likelihood-ratio output of a forensic-voice-comparison system", Cernocký, H. and Burget, L. (Eds.), *Proceedings of Odyssey 2010: The Language and Speaker Recognition Workshop, Brno*, International Speech Communication Association, 2010, p 63–70.