J P French λssociates

In the Matter of

ETS TOEIC Fraud

Report on Testing of Samples Undertaken by ETS

Dated 5th February 2015

Author: Dr Philip Harrison

Prepared on the Instructions of

Bindmans
236 Gray's Inn Road
London
WC1X 8HB

by

J P French Associates
Forensic Speech & Acoustics Laboratory
86 The Mount
York
YO24 1AR

Tel: 01904 634821 Fax: 01904 634626
E-mail: enquiries@jpfrench.com

J P French Limited trading as J P French Associates
Registered in England and Wales, Registration No. 5095541
Registered Office: 86 The Mount, York, YO24 1AR

I confirm that I have made clear which facts and matters referred to in this report are within my own knowledge and which are not. Those that are within my own knowledge I confirm to be true. The opinions I have expressed represent my true and complete professional opinions on the matters to which they refer.

Dated 5th February 2015

Signed ....

1 PERSONNEL

I am a forensic consultant specialising in the analysis of speech, audio and recordings. A summary curriculum vitae is appended to this report (Appendix A).

2 MATERIAL RECEIVED

I was provided with the following items by Salima Budhani of Bindmans:

Table 1 Material received from Salima Budhani of Bindmands

Item	Date received	Delivery method	Item/ Format	Content
1	06/01/2015	Email	Statement	Witness statement of Peter Millington dated 23/06/14 in the matter of Zaheer Hussain Mohammed v Secretary of State for the Home Department
2	06/01/2015	Email	Statement	Witness statement of Rebecca Collings dated 23/06/14 in the matter of Zaheer Hussain Mohammed v Secretary of State for the Home Department
3	09/01/2015	Post	CD-R	Video of BBC Panorama program 'Immigration Undercover: The Student Visa Scandal'

3 INSTRUCTIONS

3.1 Instructions

J P French Associates received instructions from Ms Budhani in a letter dated 6th January 2015. These are summarised below.

1. Describe the analysis process that would be undertaken by my laboratory if instructed to analyse a potentially fraudulent TOEIC test, including equipment and software used and the time required to undertake the examinations.
2. Consider the approach taken by ETS in general but in particular:
    a. Comment on the validity and reliability of automatic software tools used in identifying recordings of the same speaker;
    b. In relation to the human verification process, comment on whether this is a reliable process;
    c. In relation to the human verification process, comment on what length of sample would be necessary and consider potential issues regarding accent and dialect;
    d. Consider the time taken to conduct the examinations;
    e. Comment on the likely number of false positives produced by the process.

3.2 Standard expert declaration of impartiality

I understand that my duty is to the court rather than to those instructing me. I have complied with this duty and will continue to do so. I am aware of the requirements of Part 35 of the Civil Procedure Rules, Practice Direction 35 and the Protocol for Instruction of Experts to give Evidence in Civil Claims.

4 SPEAKER COMPARISON METHOD EMPLOYED BY J P FRENCH ASSOCIATES

The most common type of examinations undertaken by J P French Associates is forensic speaker comparison analysis, usually for criminal proceedings. The outcome of the analysis is an opinion concerning the potential identity or not of an unknown speaker in a recording. A normal case involves comparing the known voice and speech patterns of a suspect, usually from a police interview recording, with the voice of an offender in a criminal recording; for example in a hoax 999 call, a fraudulent telephone call to a bank or a covert recording relating to the supply of drugs.

If we were instructed to examine materials from an individual potentially fraudulent TOEIC tests in which it was suspected that a proxy sitter had been used, we would follow the same method we currently use for *forensic* speaker comparisons. The method is known as the auditory-acoustic phonetic method and is described in the sections below. (More comprehensive descriptions of the method are provided in French and Stevens 2013 and Jessen 2008.) The general method is also followed by all other forensic speech analysts working within the UK, as well as a significant proportion in Western Europe, including government and police laboratories in countries such as Germany and the Netherlands. In a recent survey of practicing forensic speech scientists it was found to be the most commonly used method internationally (Gold and French 2011).

The method is componential in that it involves analysing different aspects of the voice and speech patterns found in a recording. The profiles of the features that are found are then compared across the recordings. The analysis process usually takes between 10 and 15 hours for a comparison of two samples.

4.1 Question Addressed

The availability of samples and information concerning the potential fraud would govern the questions that could be addressed. These might be:

1. Compare recordings of answers from two or more tests from purportedly different candidates to assess whether they have been taken by the same proxy sitter;
2. Compare recordings of answers from one test with a known sample of speech from the named candidate to assess if they are the same speaker;
3. Compare recordings of answers from a test (or tests) with a known sample of speech from the suspected proxy sitter to assess if they are the same speaker.

Each question would provide slightly different information but would ultimately assist in addressing the issue of whether fraud had been committed.

4.2 Preparatory Work

Before the analysis commences the samples need to be prepared. This involves ensuring that the recordings are in a suitable format for replay and analysis within a computer. This may involve re-recording (digitising) material from analogue cassette or converting audio data files from a proprietary or compressed format to a standard format. This is done using a standard PC with a high quality audio interface and high quality replay equipment. Following this, the samples are manually edited so that they only contain speech from the voice of interest in the recording. For interview recordings, this involves removing speech from the interviewing officers and legal representatives. This is again carried out using a standard PC and audio editing software, specifically *Sony Sound Forge*.

I have not had access to any audio files resulting from TOEIC tests, but it is likely that they would require some editing in order to remove any long pauses or silences between the speech of the test taker and to remove any potentially intrusive background sounds and speech. The editing is undertaken so that the analyst can focus on the speech of the voice being analysed and not be distracted by other voices or sounds, and so that analysis software is only producing measurements based on the voice of the speaker being examined.

4.3 Analysis

The analysis of the recordings involves two types of examination which are carried out in parallel: auditory-phonetic and acoustic-phonetic. Auditory-phonetic analysis involves listening analytically to individual voice and speech parameters and assigning them to categories laid out in frameworks established by the speech sciences. The second, acoustic-phonetic analysis, uses computer software to produce visual representations of the speech signal and make various measurements; both automatically and manually. Each type of analysis is described below.

4.4 Auditory-Phonetic Analysis

The components of the auditory-phonetic analysis are described below. They are all undertaken whilst listening to the material via high quality headphones in sound editing or sound analysis software (in the case of our laboratory, *Sony Sound Forge* and/or *Praat*).

4.4.1 Segmental analysis

Segmental analysis involves determining how individual vowel and consonant sounds (or speech 'segments') are pronounced. For example, whether or not someone produces the 'h' sound at the start of words such as 'hello' or 'help', whether they replace 't' sounds with a glottal stop as in words such as 'better', whether a long or short vowel is used in words such as 'bath' and 'path'. The way that the sounds are produced is recorded using a specialised system of notation developed by the International Phonetic Association for capturing the fine-grained nuances of speech.

A profile of the speaker's pronunciation patterns is compiled for each sample. This assists in determining an individual's regional accent, as well as assessing any first language influences if they are not a native speaker of English. Most importantly, the profile also allows potentially distinctive realisations to be highlighted which may not be standard for that speaker's variety of English; for example, lisped pronunciations of 's' or 'w'-like realisations of 'r'.

4.4.2 Voice quality

Voice quality is the overall tone or timbre of someone's voice. The analysis undertaken during a comparison involves uses a modified version of the Vocal Profile Analysis Scheme developed at the University of Edinburgh, which can be broken down into three main strands:

1. phonation types - features originating from the larynx e.g., breathiness, creakiness, harsh- or rough-ness;
2. overall muscular tension - i.e., whether someone has very lax or very tight control of their larynx and articulators;
3. features relating to the vocal tract - i.e., features originating above the larynx, such as tongue body/blade/tip position, jaw position, degree of nasal resonance and larynx height.

This allows the voice to be assessed under 12 main categories with 38 different sub-categories.

4.4.3 Pitch, intonation, rhythm and tempo

Several other aspects of a person's speech which occur at a more general level are also considered. These include speaking pitch, i.e. how low or high pitched their voice is in general and also their intonation patterns, i.e. how their pitch changes within utterances - the melody of speech. The speed or tempo of speech is noted and sometimes measured along with rhythmic features. A speaker's level of fluency will also be noted, with particular focus on distinctive disfluency features if present.

4.4.4 Lexical/grammatical choices and conversation management

If elements of the spoken content are considered to be potentially distinguishing, then patterns of language and grammar use may also be analysed. This might also involve their use of hesitation markers, e.g. 'ums' and 'ers', as well as specific phrases, interruption strategies, telephone call opening or closing patterns.

4.5 Acoustic-Phonetic Analysis

Acoustic-phonetic analysis entails the use of specialised speech analysis software to produce visual representations of the speech signal and allows the analyst to make objective measurements. The different types of analysis are discussed below. The software used at J P French Associates is called *Praat* and it is commonly used within the forensic field. It is also widely used by phoneticians and speech scientist more generally.

4.5.1 Spectrographic Analysis

One of the main tools within speech analysis software produces what is known as a spectrogram. This is a graphical representation of the different frequencies of sound energy over time. Different types of speech sounds show different configurations and patterns of energy. Such displays are often viewed during the segmental auditory analysis as they can provide corroboration of what is heard.

Measurements in both time and frequency can be taken from spectrograms. This could include measuring the duration of specific speech sounds or the frequency at which certain features occur, for example the concentration of energy is an 's' sound.

4.5.2 Formant Analysis

Formants are resonances or concentrations of energy at different frequencies that occur when vowel sounds are produced. Different vowel sounds have different characteristic frequency patterns but individuals also display variation of their patterns of usage because formant frequencies result from the configuration of the articulators (the tongue, the lips etc.) but also the shape and size of a person's vocal tract. The frequencies that formants occur at are measured in a manual or semi-automatic way for multiple instances of the same vowel sound occurring in different words. These values are logged by the computer program. They can then be compared across the samples and patterns of overlap or deviation can be seen.

4.5.3 Fundamental Frequency Analysis

Speech analysis software automatically measures the fundamental frequency or pitch of the voice across the entire speech sample and produces various statistical measures, such as the mean, standard deviation and the distribution of the measurements. This provides an objective measure of the speech that can be easily compared across the samples and against population data.

4.5.4 Articulation Rate

Articulation rate is an objective measure of the speed that a person speaks. To calculate this, the number of syllables spoken in a short utterance is counted and the duration of the utterance is measured. This is done across several utterances and an average syllables per second rate is calculated. The articulation rate can be compared across samples and also against population data.

4.6 Conclusion

In reaching the final conclusion, the findings of the analyses are considered in two ways. Firstly, an assessment is made of the similarity of the voices across the samples. Consideration must be given to any factors which may influence the speech within them, such as emotional state, intoxication, technical characteristics of the recordings. Secondly, the distinctiveness of the voice in the criminal sample is considered in order to assess the number of potential speakers who may also exhibit the voice and speech patterns found. These two factors are weighed and the final conclusion is given as a statement of support for either the view that the speakers are the same or that they are different. For example:

On the basis of these assessments, my opinion is that the evidence provides strong support for the view that the questioned speaker is Mr Smith.

Forensic speech analysts generally do not provide categorical statements of identity or non-identity such as 'Mr Smith made the call'. This is because there are no speech features either individually or in combination that can uniquely identify individuals. This is a fundamental limitation of the science and the method. Also, forensic analysts do not produce statements of likelihood along the lines of 'it is highly likely that Mr Smith made the call'. Statements such as this stray into the province of the triers of fact and fall foul of the prosecutor's fallacy (French and Harrison 2007[1]).

4.7 Educational & Training Requirements for Analysts

There are no specific regulations or statutory requirements governing the qualifications and training for forensic speech analysts. However, within the field it is generally accepted that at least a master's level university education in phonetics/linguistics is required, if not a doctorate. Such qualifications should involve substantial components of phonetics, socio-phonetics and speech acoustics.

Specific forensic training should ideally take place over a period of at least one to two years and cover a large number of different forensic cases. This would involve shadowing the work of an established expert as well as conducting work independently that is checked and subject to detailed scrutiny by a mentor.

[1] The framework for expressing conclusions presented in French and Harrison (2007) has since been superseded by the support statement framework discussed above.

4.8 Combined Use with Automatic Systems

At present, we do not routinely use an automatic speaker recognition (or 'ASR') system in casework. This is because it has not yet been possible to satisfactorily test system performance in a sufficient range of realistic case conditions due to the lack of representative speech databases. The variation in the types of criminal recordings encountered in the UK is much greater than in many other countries in which ASRs are more frequently used. In these countries recordings of intercepted telephone calls are admissible as evidence and are frequently encountered. However, we have used an ASR system (*Batvox*) in a recent Criminal Court of Appeal case in which its results provided further evidence of incorrect identifications and supported the findings of earlier auditory-acoustic phonetic analyses. In this case we were able to test the performance of the automatic system using samples equivalent to the disputed ones in which the defendants were agreed to be present.

In cases where such software is used, it is generally considered as an additional tool rather than being used as the sole test. The results from the system are considered in conjunction with the results of the other auditory and acoustic phonetic tests. This approach is advocated by Becker *et al* (2012):

> 'Using automatic forensic voice comparison systems without any further investigation of the recording material results in a considerable proportion of errors. This proportion can be reduced if forensic phonetic experts are involved to judge the material as well as [the ASR].' (page 5)

Indeed it seems that ETS have, to some extent, followed this approach in adding a human verification element to their process (albeit not one involving forensic phonetic experts).

4.9 Other Factors

4.9.1 <u>Appeal Court Rulings</u>

Forensic speaker comparison analysis undertaken for the criminal courts in the UK is subject to two specific Appeal Court Ruling Rulings. The first of these is R -v- Anthony O'Doherty ([2002] NICA 20). This relates specifically to the jurisdiction of Northern Ireland and states:

> “in the present state of scientific knowledge no prosecution should be brought in Northern Ireland in which one of the planks is voice identification given by an expert which is solely confined to auditory analysis. There should also be expert evidence of acoustic analysis … which includes formant analysis.”

This ruling essentially makes acoustic analysis, and specifically formant analysis, a compulsory aspect of any voice comparison exercise undertaken for the courts in Northern Ireland. Whilst it is not binding on the courts in England, Wales and Scotland, analysts in these jurisdictions do carry out acoustic analysis, including formant analysis.

A second ruling R -v- Flynn & Another ([2008] EWCA Crim 970) also concerns voice comparison evidence given by experts as well as evidence given by lay listeners, in this case police officers. The relevant paragraphs from the postscript of the ruling are:

> 63. The increasing use sought to be made of lay listener evidence from police officers must, in our opinion, be treated with great caution and great care. In our view where the prosecution seek to rely on such evidence it is desirable that an expert should be instructed to give an independent opinion on the validity of such evidence. In addition, as outlined above, great care should be taken by police officers to record the procedures taken by them which form the basis for their evidence. Whether the evidence is sufficiently probative to be admitted will depend very much on the facts of each case.
>
> 64. It goes without saying that in all cases in which the prosecution rely on voice recognition evidence, whether lay listener, or expert, or both, the judge must give a very careful direction to the jury warning it of the danger of mistakes in such cases.

This ruling highlights the court’s view on the potential issues associated with speaker comparison evidence given by lay listeners.

5 PERFORMANCE OF AUTOMATIC SPEAKER COMPARISON SYSTEMS

5.1 Principles of Automatic Systems

Automatic speaker recognition systems work on the principle that individual voices may be distinguished from one another by virtue of the different anatomical dimensions and proportions of different speakers' vocal tracts, and that individuals also exhibit different speaking behaviours. These factors give rise to acoustic differences; namely, differences in the structure of the frequency characterises of speech found across individuals. To capture these differences, automatic systems take the recorded voices of individuals, perform complex mathematical operations on them, and reduce them to statistical representations or models.

In conducting an analysis, a system compares the models generated from two different recordings or samples and produces a score which is a measure of the similarity/difference between the two[2]. In order to determine whether the score is indicative of the two samples having come from the same or different speakers, one of the models is also compared with a set of statistical models from a reference population of other speakers held within the system. The characteristics of the reference population, including gender, language and recording conditions, should ideally be the same as those of one of the speaker models in the initial comparison. This is a generic description of automatic systems and the steps followed by specific systems may be different. A comprehensive overview of automatic speaker recognition is provided by Kinnunen and Li (2010).

The statistical models generated by automatic systems are sometimes referred to as 'voice-prints', as in paragraph 27 of Mr Millington's report. This terminology is problematic and potentially misleading. Speaker models are not the equivalent of finger-prints. Fingerprints are a stable physiological attribute of an individual. Speech is the product of a dynamic bio-mechanical process, which is affected by many factors leading to variation in the speech of individuals. In view of this, speaker models aim to capture a general representation of an individual's speech and they are not considered as uniquely identifying.

[2] This is the current state of the art approach employed by systems that use a speaker models known as 'i-vectors'. In previous generations of ASR technology the speech features extracted from one sample were compared with a speaker model obtained from a second sample.

5.2 *Determining the Performance of Automatic Systems*

One of the benefits of automatic speaker comparison systems is that it is relatively easy to assess their performance, i.e. how good they are at correctly determining if two samples of speech are from the same person. This is because the comparison process is automatic and the comparisons are done very quickly. To assess a system's performance, a large database of recordings is required in which the true identity of the speakers is known. Two different types of comparison are carried out by the system. One set of tests involves pairs of samples where the same speaker is in each sample, known as same speaker tests, and a second set is where different speakers are in each pair of samples, known as different speaker tests.

In its simplest form the output of an automatic system is a simple binary 'yes' or 'no' answer to the question of whether the two speech samples being compared are of the same speaker.

For the same speaker tests, if the system gives a 'yes' result, then it is correct and this is known as a 'true positive' or 'true acceptance'. If the system gives a 'no' result, then it has made an error known as a 'false negative' or 'false rejection'.

In the case of the different speaker tests, if the answer given is 'no', then this is the correct result referred to as a 'true negative' or 'true rejection'. If the result is a 'yes' then the system has made an error known as a 'false positive' or 'false acceptance'.

The different types of correct results and errors can also be expressed in table form as shown below.

Table 2 Correct responses and errors results obtained when testing the performance of an automatic speaker comparison system

System Result	**Same Speaker Pair**	**Different Speaker Pair**
Yes	True Positive – Correct Result	False Positive – Error
No	False Negative – Error	True Negative – Correct Result

If a large number of same speaker and different speaker pairs are tested by the system then it is possible to calculate errors rates for the system which reflect its performance. For example, if 100 same speaker pairs are tested and the correct result of 'yes' is obtained for 95 of the 100 pairs, then it has a 'true positive' rate of 95%. For 5 of the pairs the incorrect result of 'no' was given meaning that the 'false negative' error rate was 5%.

The results produced by automatic systems are numeric scores which reflect the degree of similarity between two samples – larger numbers reflect greater similarity and smaller numbers reflect a greater dissimilarity between samples. For a yes/no decision to be made a score threshold must be established so that all scores above the threshold are considered as a 'yes' and all those below it are a 'no'. This is discussed in paragraph 32 of Mr Millington's witness statement. Changing the threshold alters the errors rates of the system since results from some pairs will change from a 'yes' to a 'no' or vice-versa. There is a trade-off in the error rates. As the threshold increases the false negative error rate increases whilst the false positive error rate decreases. Conversely, if the threshold is decreased then the false negative error rate decreases and the false positive error rate increases. Therefore the choice of threshold is crucial in determining the errors rates and performance of the system. The changes in error rate resulting from different thresholds can be represented as shown in the plot below.

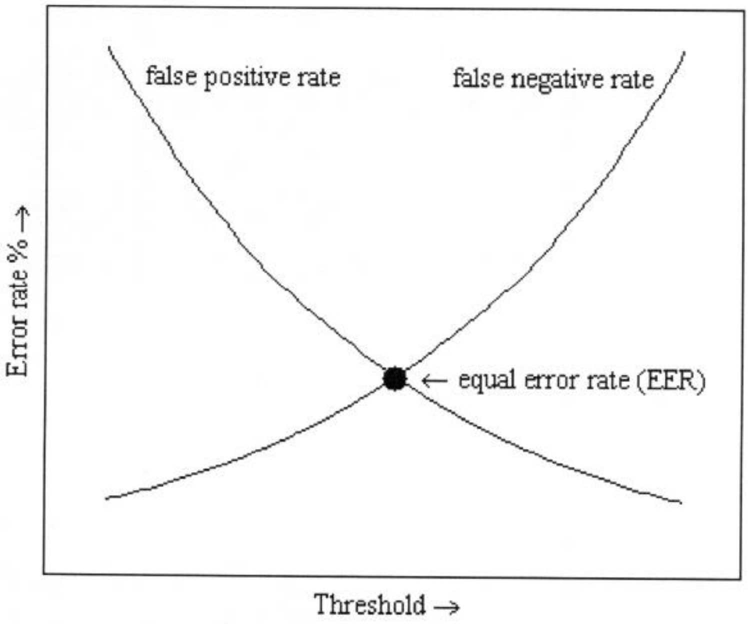

Figure 5.1 Plot showing how false positive rate and false negative rate changes for different thresholds.

Another way in which the performance of a system is often expressed is known as the 'equal error rate' (EER). On the plot above there is a threshold value at which the false positive and false negative error rate curves cross over and both have the same value. This is the equal error rate since both errors are equal. Whilst this is a commonly used measure of system performance since it is a single number, the actual errors rates of a system will be different if the threshold used in the operation is not the one that gives the equal error rate.

The descriptions above show that systems can make two types of errors, a false positive or a false negative. For a quoted error rate to be meaningful, the type of error that it refers to must be stated. There are other graphical methods and numeric measures that can be used to show the performance of a system. However, the ones discussed above are the most straight-forward and demonstrate the concept of system performance for the present purpose.

5.3 *Factors Affecting Performance*

There are many factors that affect the performance of automatic systems. Those that are most relevant to the current case are discussed below.

5.3.1 Duration of samples

The duration of a sample reflects how much speech material is contained within it. The longer a speech sample, the more speech is available to the automatic system to generate a speaker model. In general terms, the longer the samples that are compared, the better the performance of the system. Table 3 below shows the minimum and recommended sample durations for a commercially available automatic speaker comparison system, Batvox 4, which is widely used by forensic analysts. The durations given refer to the net amount of speech that the system extracts from a sample. Systems will pre-process each sample to automatically remove non-speech sections such as silences or pauses. Separate sets of values are given for both the known reference recording and the unknown criminal sample. These values have been determined by the manufacturer and are based on their own testing of the system. However, these values do not imply that a specific level of performance can be achieved since the other factors discussed below also influence performance.

Table 3 Absolute minimum and recommended net speech length for speech samples used in Batvox 4

	Net speech length	
	Known Sample	**Test Sample**
Absolute minimum	30 seconds	7 seconds
Recommended	60 seconds or greater	15 seconds or greater

5.3.2 Quality of samples

The quality of samples also has a marked influence on system performance. Quality can refer to many different aspects of a recording. One of the key aspects is the amount of noise in a recording. Noisier recordings will lead to a reduced performance relative to quieter ones. Different types of noise will also affect the performance to differing degrees. For example music has been shown to decrease performance more than road noise within a car (Künzel and Alexander 2014). Representing noise in a meaningful quantitative way is somewhat problematic since the different kinds of noise have varying time and frequency structures, which influence performance in different ways that cannot be simply expressed as a single figure. However, one commonly used measure is known as signal to noise ratio (SNR), which reflects how loud the signal of interest is, in this case the speech, relative to the noise. The minimum and recommended SNR values for both known and criminal recordings to be used in Batvox 4 are given in Table 4 below.

Table 4 Absolute minimum and recommended SNR for speech samples used in Batvox 4

	Signal to noise ratio (SNR)
Absolute minimum	10 dB
Recommended	15 dB or greater

Quality can also relate to the recording format used. If a lossy compression format such as MP3 is used to make a recording, this can also influence the performance. Recordings of a limited bandwidth, such as those made over the telephone also tend to reduce the performance of ASR systems.

5.3.3 Reference population mismatch

A reference population of speakers is used in automatic systems to assess whether the degree of similarity between two samples is indicative of them having come from the same speaker or not. For the reference population to do this job correctly it must match various characteristics of the materials being testing, including recording quality and language. If there is a mismatch with these characteristics then the system is likely to make more errors.

5.4 Assessing the Performance and Suitability of Automatic Systems

Since there are no automatic systems that perform perfectly, there are no systems that are completely reliable. In other words, all systems will make errors. To assess the degree of reliability of a system its performance must be measured by conducting large numbers of comparisons of pairs of same speaker and different speaker recordings where the ground truth answer is known. The issue of whether a system is reliable enough depends on the context in which it is being operated and how the results from the system are used in reaching decisions. The potential consequences of false positive and false negative errors are different in telephone banking authentication, intelligence gathering exercises and in criminal court proceedings. It is not possible, therefore, to provide a single threshold figure for either the quality or the duration of a sample at which the results of an automatic system change from being unreliable to reliable.

A very important consideration is that quoted performance and error rates only relate to the specific recordings involved in the testing and their characteristics such as duration and quality. If a different set of recordings are used, even with the same characteristics, the performance will be slightly different. Therefore quoted error rates can only be used to infer the performance of the system and are not absolute concrete, unchangeable values.

5.5 Typical performance of ASR systems

The discussion above in Section 5.3 introduces three factors that can influence the performance of ASRs: duration, quality and reference population mismatch. Other important factors are the underlying method employed by the system and mismatch in recording conditions between the samples being compared, such as a microphone interview recording with a telephone recording. Given the number of influencing factors and the interaction between them, a wide range of performances are reported in the research literature. In very general terms, the *best* performance for state of the art systems, expressed as equal error rate, is typically between 1% and 3% (Hautamaki *et al* 2010). However, commonly reported equal error rates can range from between 2% to 10% (Kinnunen and Li 2010 and Gonzalez-Rodriquez 2014). In non-ideal conditions, equal error rates can be as high as 20 or 30% (Hautamaki *et al* 2010).

6 RELIABILITY OF PROCESS USED BY ETS

In principle, given the number of TOEIC tests undertaken in the UK and the amount of material to be analysed, the general approach adopted by ETS is reasonable. To subject thousands of samples to the type of detailed analysis described in Section 4 would be entirely impractical.

However, the level of detail provided in the witness statements of Mr Peter Millington and Ms Rebecca Collings is not sufficient for me to be able to properly scrutinise ETS's approach. The circumstances of the present case are different from that usually encountered in that witness statements describing an analysis process are normally authored by the person conducting the analysis, not a third party. Such statements tend to contain a greater level of detail. Also, the recordings that have been examined are normally made available for analysis by experts representing the other party involved in the matter.

6.1 Specific Issues with Automatic Analysis

6.1.1 Lack of Information Concerning Initial Testing

Paragraphs 28 to 34 of Mr Millington's statement[3] describe a "proof of concept" exercise undertaken by ETS to assess whether automatic speaker comparison technology would be suitable for their purposes in detecting fraud. The tests involved comparing a large number of same speaker pairs (285) and different speaker pairs (over 70,000). There is certain information not presented about this testing which makes it difficult to assess. This includes:

[3] All further references to paragraph numbers concern the report of Mr Millington.

1. Other than the fact that the testing involved 'representative data' no information is provided about the quality or duration of the samples, or the type of spoken material, i.e. spontaneous or read.
2. Paragraph 28 lists three key questions that ETS felt should be addressed before the technology should be used. They relate to the sensitivity of the results of the system to i) first languages spoken, ii) duration and quality of samples and iii) suitable thresholds. There is no further mention of whether these issues were addressed sufficiently by the testing other than with the very general statement that the trial was successful. No mention is made of any threshold or limits that were chosen for example on sample duration or quality.
3. An error rate of less than 2% is quoted in paragraph 31. However, the type of error rate i.e. false positive, false negative or EER, is not given.
4. Paragraphs 32 to 33 discuss the trade-off between false positive and false negative errors and selecting an appropriate threshold. However, the resulting error rates from the selected threshold are not provided.

6.1.2 Comparability of Initial Testing with Testing for Current Purposes

No reference is made to how comparable the samples used for the proof of concept testing are to the samples used for the TOEIC testing. Differences in technical attributes of the samples such as duration and quality may have resulted in differing performance of the software. ETS acknowledged that the variety of first languages spoken by their clients (paragraph 28 i)) may have influenced the performance of the system but there is no mention as to whether this was found to be the case and whether the first languages encountered in the proof of concept testing was the same as those in the TOEIC testing.

There is no explicit statement that the configuration of the automatic system used in the proof of concept testing was the same as that used for the TOEIC testing. It is possible that the system received updates from the manufacturer that may have resulted in changes to its performance. Also, there may have been user-configurable parameters that were different which could affect performance, such as the selection of reference populations.

6.1.3 Description of Analysis Method

In general, only limited information is provided about how the automatic system was used. No detail is given about the system itself since 'the procurement of the technology remains subject to a confidentiality agreement' (paragraph 29). It is highly unusual for fundamental information such as this not to be provided in a forensic report. If the manufacturer and model of the system were given, it would potentially allow the underlying analysis approach of the tool to be known and further information concerning its general performance could be obtained by consulting other studies.

Paragraphs 36 to 38 provide some detail of the analysis process followed. It states that six audio files were selected for comparison from each test, based on the largest/longest audio files, those providing the clearest responses or those involving the reading of a set text. No mention is made of specific durations of the files selected and whether the audio from the six files was combined to form a longer recording. Since the term 'or' is used within the list it is possible that short, poor quality samples of reading may have been selected.

No mention is made as to whether this selection process was done manually or automatically and whether any editing of the samples was done prior to the analysis to remove non-speech events such as coughs, or loud instances of background speech.

There is no indication whether the consistency of the speaker across the six files was assessed by the automatic system. If it is assumed that the same test taker answered all parts of a test then this testing across the six samples would have provided an indication as to how reliable the system was at identifying the specific voices in individual tests.

As mentioned above, it is not apparent whether each of the six audio files from each test was considered separately or combined into a single file. It is therefore not clear whether the flagging of a test as a 'match' was based on the result from a single audio file or not. If the results were based on the single files then the overall performance of the system might be expected to be less than if the six files had been combined. This is because the individual files would be shorter and there would be the potential for individual files to be noisy or have significant amounts of background speech.

No explicit mention is made about the scoring method used by the system and whether a reference population was employed in the calculation of the final comparison score. If a reference population was used, it is not stated how well this was matched to the test materials and whether it was varied for individual tests.

6.2 General Issues with Human Verification

6.2.1 Factors Affecting Performance

Like automatic systems, humans also make errors when attempting to identify speakers. There are many factors that can affect the performance of humans when conducting these tasks. There is a significant body of academic research literature which investigates these factors and their influence on performance, for example Bull and Clifford (1984), Foulkes and Barron (2000), Kerstholt *et al* (2004), Shirt (1984) and Yarmey (2004). The studies that have been conducted address a wide variety of factors and employ different experimental designs.

In summary, some of the factors that have been shown to affect performance include:

1. Duration of speech material – shorter samples result in worse performance;
2. Quality of speech material – worse quality results in worse performance;
3. Individual ability of the listener – amongst lay-listeners some people are better than others and phonetically trained people perform better than lay people;
4. Familiarity with the voice of the speaker – greater familiarity can lead to better performance but not always;
5. Distinctiveness of the voice – more distinctive voices are normally easier to identify and result in better performance.

6.2.2 Typical Performance of Human Verification

Many of the studies concerning the ability of humans to identify voices are focused on the performance of lay-listeners and are designed to provide insight into lay-witness identification i.e. the situation where a lay-witness claims to have recognised the voice of a criminal. The test conditions in these studies are not representative of the task undertaken by the ETS analysts so they are only able to provide insight into factors that can affect performance rather than provide comparable performance results.

However, there is one particular body of research which aligns more closely with the human verification task undertaken by ETS. This is the Human Assisted Speaker Recognition (HASR) task which was part of the American National Institute of Standards and Technology (NIST) Speaker Recognition Evaluation (SRE) project undertaken in 2010. Since 1996 NIST has conducted a series of evaluations of automatic recognition systems in which they provided companies and academic institutions around the world with large quantities of speech material to allow them to test the performance of their automatic systems and compare it with others. In 2010 NIST incorporated the HASR task to test the performance of humans.

For the HASR task two sets of recordings were made available to the human analysts. The first, HASR1, consisted of 15 pairs of speakers, 6 were same speaker pairs and the other 9 were different speaker pairs that were judged to be difficult to distinguish. The second set HASR2 had 150 test pairs, 51 same speaker pairs and 99 different speaker pairs, and was somewhat less difficult than HASR1. The pairs of recordings involved one telephone and one microphone recording and both were approximately 3 minutes or greater in length. The human listeners could be a single person or a team of listeners, and they were free to use just human expertise or combine it with an automatic system.

Averaged over the 8 human systems that took part in the HASR2 test, the false positive rate was 42%. The best false positive error rate for any system was 3% (system 20), whilst the worst was 75% (system 11). Unfortunately the details of each human system are not made available, so it is not possible to know the specific method that was used by any of the systems. Further details, including the complete set of results, can be found in Greenberg *et al* (2010)[4] and Schwartz *et al* (2011).

[4] The complete set of results is provided in the presentation slides rather than the proceedings article. The slides are available to download from: http://www.nist.gov/itl/iad/mig/upload/hasr_od10_webpage.pdf

6.2.3 Defining Thresholds

Similarly to automatic systems, it is not possible to define a set threshold for duration or quality of material at which a method changes from being unreliable to being reliable. Similarly it is not possible to prescribe a specific amount of training that someone must undergo or a set amount of knowledge that they must have before being able to reliably conduct a comparison. However, research has shown that lay-listeners are more prone to errors than trained and experienced phoneticians (Schiller and Köster 1998).

6.3 Specific Issues with Human Verification Method Employed by ETS

6.3.1 Explicit Acknowledgement of Human Errors

The reason for using human analysts as a second stage of analysis was to "avoid the occurrence of false positives" (paragraph 40) in the results from the automatic comparison process. There is further acknowledgement that "ETS accepted that voice biometric technology is currently imperfect" (paragraph 32). However, there is no explicit acknowledgement that the human verification process will almost certainly have resulted in false positive errors.

ETS have clearly taken steps to attempt to reduce the likelihood of the human analysts making false positive errors. These include the use of two analysts working independently, the training of new analysts and the requirement for analysts to reject any matches where they had any "doubt about the validity of a match".

6.3.2 No Testing of Analysts Performance

There is no mention that human analysts were subject to any specific testing of their performance. In relation to the training of new analysts, their work was "peer reviewed by an experience analyst until a level of confidence was reached that they were capable of carrying out the work on their own". However, there is no indication of what this level was in objective terms (see further discussion in Section 7.3.3 below).

Performance testing would have been possible as the OTI had a database of TOEFL recordings that they had used when assessing the suitability of the automatic system. Whilst it would be very time consuming to run the same number of tests as that undertaken by the automatic system a smaller number of samples would still provide an indication of performance. Without knowing the performance of the human analysts it is not possible to make an objective assessment of the overall reliability of the process used to ETS to identify potentially fraudulent tests.

6.3.3 Relationship Between Confidence & Accuracy

As mentioned above, one of the methods used to reduce the likelihood of false positive errors was for analysts to reject any matches where they had any "doubt about the validity of a match". This implies that they only verified matches where they had a very high degree of confidence in the accuracy of their opinion. However, many studies have found that there is no correlation - or at best a weak correlation - between level of confidence and accuracy of recognition. Studies by Yarmey *et al* (2001), Yarmey (2004), Broeders and Rietveld (1995), Hollien *et al.* (1983), and Sørensen (2012) found very limited, if any, correlations between correct judgements and confidence. Although Rose and Duncan (1995) showed a significant positive correlation between confidence and accuracy, this was only for familiar speakers and in individual cases listeners were incorrect despite being 'very certain'.

In court cases involving ear-witness identifications, Bull and Clifford (1999) recommend that witness confidence ratings should not be used by courts as an indication of correctness as there are only limited and variable correlations between confidence and accuracy.

The implication of these research findings for ETS's testing is that although the analysts only verified matches where they had no doubt about their validity – i.e. where they were certain about their judgements – this should not be taken as a reliable indicator of the accuracy of those judgements. This approach does not remove the risk of false positive results.

6.3.4 Details of Analysis Method

Paragraph 24 makes references to a procedure that OTI analysts would use to compare "two voice samples to establish whether the tests were taken by the same person". It states that these were developed 7 years ago with input from a former FBI officer. It must be assumed that these are the procedures that were followed in the current matter. However, no information is given about the actual process. Also, no specific details are given about the background, skills and knowledge of the former FBI officer. Without this information it is very difficult to attempt to make an assessment of the method.

In relation to the specific task undertaken, there is no indication of how much time was spent by each analyst on the samples from each test. There are no indications of how many times each sample was listened to. There is no information about the level of detail that was included in any analysis notes that might have been made. Again, this information would provide insight into the analysis process followed and the level of detail in the analysis.

6.3.5 Degree of Experience & Knowledge of "Experienced Analysts"

A distinction is made between experienced analysts within OTI and other members of ETS's staff who were brought in to assist. Little specific information is given in relation to the experience, abilities and knowledge that the experienced analysts have. It is not clear what training they have received, how often they perform this kind of analysis and their familiarity with different varieties of foreign accented English that they might encounter. Whilst they may have greater experience and skills than a lay-listener, it would appear that they cannot be considered the equivalent of phonetically trained forensic experts.

6.3.6 Training of New Analysts

Paragraph 40 states that the additional ETS staff "received mandatory training in voice recognition analysis, and were initially mentored by experienced OTI analysts". However, no information is given about the content of this mandatory training, how long it lasted, what the mentoring involved or how long it lasted. Given the timeframe involved, the training can only have been very limited relative to that undertaken by forensic phoneticians.

In order to determine which staff to redeploy or to allow others to decide not to continue with the checking of samples, information on their performance must have been available. Again, no information is provided about the threshold above which staff were considered suitable for the task.

6.3.7 Familiarity with language varieties

In paragraph 42 Mr Millington states that as part of the demonstration of samples he received he "was able to compare tone, accent, the distinctive and indistinctive expressions used to fill hesitation in speech". It is not clear whether these are specific features that the OTI analysts also used as part of their comparison process. In order to be able to assess the usefulness and distinctiveness of features one must know what the norms are for a particular variety of a language. For example, a non-native speaker of English might use a particular pronunciation of a vowel sound that, when compared with standard English, appears to be very distinctive. However, the use of that non-standard pronunciation might be very common in other speakers with the same first language because their English pronunciations are influenced by their first language. Without a detailed knowledge of these pronunciation patterns, speech features might be classed as distinctive and form a significant part of an identification decision when in fact they are features shared by a large number of speakers. These patterns of influence and the specific features affected will vary across different first languages.

A small number of studies have addressed whether the speaker recognition ability of lay-listeners is influenced by foreign accented speech. The general finding is that a foreign accent makes the identification task harder. The difference in performance of listeners between non-foreign accented speech and accented speech varies across the studies and is also influenced by the duration of the samples (Doty 1998 and Goldstein *et al* 1981).

6.3.8 Independence of Results from Two Analysts

The decision to use two analysts working independently is a sensible approach as it prevents some types of bias which can be introduced when one analysts checks the results produced by another. However, the occurrence of errors produced by one analyst cannot be assumed to be independent of the second analyst. This is because their ability to accurately detect the same or different voices is in part governed by the degree of similarity or dis-similarity and the distinctiveness of the voices examined. If two samples from two different people sound very dis-similar then there is a high likelihood that both analysts will mark them as being two different people, i.e. their false identification rate will be low. If there are samples from two different people who sound similar and non-distinctive then the likelihood of a false identification error being committed is greater. This is demonstrated in various studies which have found that identification rates vary across speakers (for example Foulkes and Barron 2000, Mullennix *et al* 2011, Orchard and Yarmey 1995, Papcun *et al* 1989 and Sørensen 2012).

6.4 Further Issues

6.4.1 Examples Played to Mr Millington & Colleagues

In paragraphs 42 and 43 of Mr Millington's statement he explains that during his visit to ETS he was played samples which "we concluded were the same person speaking" as well as "non-verified" matches (i.e. false positives that the automatic system showed to be matches but which the human analysts rejected as a match). No information is given on how many samples they listened to or how representative they were of the material listened to by the analysts in terms of degree of similarity between speakers, duration or recording quality. Whilst this was no doubt a useful exercise to demonstrate the task undertaken by ETS, the exposure to a limited number of samples of unknown representativeness may have misrepresented the difficulty of the comparison task.

6.4.2 Basis for Mr Millington's Opinion

In Mr Millington's statement he sets out the information that he has been provided with in order to reach his view that where "ETS have identified positive voice matches among two candidates" "this is clear evidence that both candidates have fraudulently obtained their TOEIC certificate" (paragraph 49). However, it is not apparent whether Mr Millington has sufficient expertise, scientific knowledge or experience to properly assess and interpret the information that he has been provided with in order to reach his opinion. No information is given about his qualifications or any expertise that he may possess.

6.4.3 Interpretation of ETS's Results

In paragraph 48 the process undertaken by ETS has given results that are summarised as showing "that where matches have been identified the individuals taking those tests are highly likely to be the same person" [my emphasis]. This is not a categorical statement and as such it implies that there is a small probability that errors, i.e. false identifications, have occurred. Since a large number of tests have been analysed it is therefore not unreasonable to suggest that within the set of verified matches provided by ETS that there will be some false identifications, albeit an unknown number of them. This issue is addressed further in the following section.

6.5 Number of False Positive Results

As discussed at various points in this report there is a limited amount of information and detail provided about the process used by ETS. This means it is not possible on the basis of the information provided to accurately assess its reliability and estimate how many false positive results may exist within the 25,000 verified matches. In order to do this it would be necessary to know the performance of the human analysts as well as that of the automatic system.

However, it is possible to demonstrate in a hypothetical manner how the number of false positive verified matches is influenced by the performance of both the automatic system and the human analysts. For the sake of this exercise it is assumed that the automatic system produced 33,000 same speaker matches. Table 5 shows the number of false positive test results from the total 33,000 same speaker matches for different false positive error rates.

Table 5 Number of false positive results produced by the automatic system for different hypothetical false positive error rates from a total of 33,000 same speaker matches.

	Automatic System False Positive Error Rate				
	1%	**2%**	**5%**	**10%**	**20%**
Number of false positive results	330	660	1650	3300	6600

If it is assumed that the false positive errors made by the pairs of human analysts are independent of the false positive errors made by the automatic system, then the same exercise can be performed to determine the total number of false positive errors made by the combined automatic and human verification process. The results from Table 5 are reproduced at the top of Table 6. The bottom section of Table 6 shows the number of false positive results after the human verification at different false positive error rates.

Table 6 Number of false positive results produced by the combined automatic and human verification process for different hypothetical false positive error rates.

		Automatic System False Positive Error Rate					
		1%	**2%**	**5%**	**10%**	**20%**	
	Number of false positive results after automatic analysis	330	660	1650	3300	6600	
Human Verifiers' False Positive Error Rate	**1%**	3	7	17	33	66	Number of false positive results after human analysis
	2%	7	13	33	66	132	
	5%	17	33	83	165	330	
	10%	33	66	165	330	660	
	20%	66	132	330	660	1320	
	30%	99	198	495	990	1980	

Table 6 shows that if both the automatic system and the human verifiers have a false positive error rate of 1 percent then only 3 out of the 33,000 tests would be false positives. If they both had an error rate of 5 percent then that number would increase to 83. If the error rate of the automatic system was 10 percent and the human verification process was 20 percent then there would be 660 false positive results. If the automatic system false positive error rate is 20 percent and that of the human analysts is 30 percent then the number of false positive results would be 1,980.

7 SUMMARY & CONCLUSIONS

Section 4 of this report has set out how individual speaker comparison tests are carried out by J P French Associates, and forensic phoneticians more widely, using the auditory-acoustic phonetic approach. The method involves a componential analysis of a range of speech features including individual speech segments, voice quality, pitch, intonation, rhythm, tempo and spectral aspects of the speech signal. The features are assessed and compared using a range of both auditory-phonetic and acoustic-phonetic analysis techniques. The time taken to analyse and compare a reference sample with one unknown sample is normally between 10 and 15 hours. Skills in using these analysis techniques and knowledge of the patterns of occurrence and variation of speech features are obtained via a postgraduate-level university education in linguistics and phonetics and experience with forensic case materials.

Due to the time required to undertake auditory-acoustic phonetic comparisons, it would not have been practical for ETS to have adopted the method to examine the very large number of TOEIC tests undertaken in the UK in order to determine which ones may have been the subject of fraud. They used a method in which an automatic speaker comparison system compared many pairs of tests and identified those in which the same speaker occurred in more than one test. As described in Section 5 of this report, automatic systems do not perform perfectly and false positive errors can occur i.e. recordings from two tests can be identified by the system as being from the same speaker when in fact they are different speakers. In order to reduce the likely number of false positive results, ETS required each of the 33,000 pairs of tests that had been identified as a match to be verified by two human analysts.

At the level of principle, the overall method adopted by ETS is a reasonable approach, given the magnitude of the task. However, there is a lack of technical information and detail within the statements of Mr Millington and Ms Collings concerning the specific implementation. This means that it is not possible to adequately scrutinise the method or assess its overall reliability. Also, there is no explicit acknowledgment that the human verification method – and therefore the overall exercise - is almost certain to have resulted in false positive results.

Relevant factors which would influence the reliability of the method and for which there is insufficient information provided includes:

1. Technical characteristics of the TOEIC test recordings including duration and background noise levels
2. The performance of the automatic system when operated with recordings representative of the TOEIC samples
3. The comparability of the TOEIC tests materials with the TOEFL materials used in the trial of the automatic system
4. The specific operation and configuration of the automatic system including whether samples from individual tests were combined, the use of appropriate reference populations and chosen thresholds
5. The performance of the human analysts
6. The experience, training and knowledge of the 'experienced' OTI analysts
7. The training given to the members of ETS staff drafted in to assist
8. Details of the analysis method used by the human analysts

9. The time spent by human analysts on comparisons
10. The familiarity of the human analysts with the range of foreign accented English and specific knowledge of common features within those varieties.

Without this information it not possible to provide a detailed objective assessment of the overall reliability of ETS's method and the likely number of false positive results, i.e. how many tests considered as verified matches were not the result of a fraud.

Table 6 illustrates the influence of the performance of the automatic system and human analysts on the number of false positive results. In the example, the number of false positives range from 3 to 1,980 when the automatic system false positive error varies from 1 percent to 20 percent and from 1 percent to 30 percent for the human analysts.

Since there are an unknown number of false positive results there are also an unknown number of test takers who have been incorrectly identified as having fraudulently taken the TOEIC test. At present, for any specific case, there is no independent way to assess whether the individual in question committed fraud or whether their result is a false positive. The only information available is the verified match result from ETS. Since the performance of ETS's process is unknown it is not possible to assess the degree of confidence that can be placed in the results provided by ETS. If the audio material from individual tests were made available then it would be possible to use the auditory-acoustic phonetic approach to independently examine and compare the audio material. It would also be possible to use auditory-acoustic phonetic testing to assist in assessing the overall reliability of ETS's method by conducting spot checks on a selection of tests. However, such checking would only be valid if the material examined was sufficiently representative of that which was tested.

In conclusion:

1. Given the large number of tests examined and the limitations of both automatic and human speaker comparison methods it is almost certain that the set of verified match results from ETS will contain false positive errors;
2. Insufficient information has been provided to allow an assessment of the likely reliability of the method employed by ETS and the potential number of false positive results;

3. Making recordings available from tests in which concern has been raised about the accuracy of the result would allow them to be subject to independent scrutiny via auditory-acoustic phonetic speaker comparison analysis.

8 QUALITY CONTROL/CHECKING

As part of the quality control procedure at J P French Associates, the content of this report was discussed with and agreed by my colleague Dr Richard Rhodes[5].

Dr Philip Harrison BEng MA PhD MIOA 5th February 2015

[5] Forensic consultant with experience of over two hundred and fifty cases, BA (First Class Hons) in Applied English Language Studies, MSc in Forensic Speech Science and PhD in Linguistics - Forensic Speech Science.

Reference List

Becker, T., Solewicz, Y., Jardine, G., and Gfrörer, S. (2012). Comparing Automatic Forensic Voice Comparison Systems under Forensic Conditions. *In Audio Engineering Society Conference: 46th International Conference: Audio Forensics*. Audio Engineering Society.

Broeders, A. P. A. and Rietveld, A. C. M. (1995). Speaker identification by earwitnesses. In A. Braun and J. P. Köster (eds.), *Studies in Forensic Phonetics.* Wissenschaftlicher Verlag: Trier, 24-40.

Bull, R., and Clifford, B. R. (1984). Earwitness voice recognition accuracy. In G. L. Wells and Loftus, E. F. (eds.), *Eyewitness Testimony: Psychological perspectives.* Cambridge University Press: New York, 92-123.

Bull, R. and Clifford, B. (1999). *New Law Journal,* Expert Witness Supplement (Feb.), 216-20.

Doty, N. D. (1998). The influence of nationality on the accuracy of face and voice recognition. *American Journal of Psychology*, 111(2), 191-214.

French, J. P. and Harrison, P. (2007). Position Statement concerning use of impressionistic likelihood terms in forensic speaker comparison cases. *International Journal of Speech, Language and the Law*, 14(1), 137-144.

French, J. P. and Stevens, L. (2013) Forensic speech science. Chapter Twelve of M. Jones & R. Knight (eds.) *Bloomsbury Companion to Phonetics*. London: Continuum.

Foulkes, P. and Barron, A. (2000). Telephone speaker recognition amongst members of a close social network, *Forensic Linguistics*, 7, 181-198.

Gold, E. & French, J. P. (2011). International practices in forensic speaker comparison. *International Journal of Speech, Language and the Law*, 18 (2), 293-307.

Goldstein, A. G., Knight, P., Bailis, K. & Conover, J. (1981). Recognition memory for accented and unaccented voices. *Bulletin of the Psychonomic Society*, 17(5), 217-220.

Gonzalez-Rodriguez, J. (2014). Evaluating Automatic Speaker Recognition systems: An overview of the NIST Speaker Recognition Evaluations (1996-2014). *Loquens*, 1(1), e007.

Greenberg, C., Martin, A., Brandschain, L., Campbell, J., Cieri, C., Doddington, G. and Godfrey, J. (2010). Human Assisted Speaker Recognition (HASR) in NIST SRE10. *Proceedings of Odyssey 2010*, 180-185.

Hautamaki, V., Kinnunen, T., Nosratighods, M., Lee, K., Ma, B. and Li, H. (2010). Approaching human listener accuracy with modern speaker verification. *Proceedings of Interspeech 2010*, 1473-1476.

Hollien, H., Bennett, G., and Gelfer, M. P. (1983). Criminal identification comparison: Aural versus visual identifications resulting from a simulated crime, *Journal of Forensic Sciences*, 28, 208-221.

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

Kerstholt, J. H., Jansen, N. J. M., Van Amelsvoort, A. G. and Broeders, A. P. A. (2004). Earwitnesses: effects of speech duration, retention interval and acoustic environment. *Applied Cognitive Psychology*, 18(3), 327-336.

Kinnunen, T. and Li, H. (2010). An overview of text-independent speaker recognition: From features to supervectors. *Speech Communication*, 52, 12-40.

Künzel, H. & Alexander, P. (2014). Forensic Automatic Speaker Recognition with Degraded and Enhanced Speech. *Journal of the Audio Engineering Society*, 62(4), 244-253.

Mullennix, J. W., Ross, A., Smith, C., Kuykendall, K., Conard, J. and Barb, S. (2011). Typicality effects on memory for voice: implications for earwitness testimony. *Applied Cognitive Psychology*, 25(1), 29-34.

Orchard, T. L. and Yarmey, A. D. (1995). The effects of whispers, voice-sample duration, and voice distinctiveness on criminal speaker identification. *Applied Cognitive Psychology*, 9(3), 249-260.

Papcun, G., Kreiman, J, and Davis, A. (1989). Long-term memory for unfamiliar voices, *Journal of the Acoustical Society of America*, 85, 913-925.

Rose, P. and Duncan, S. (1995). Naive auditory identification and discrimination of similar voices by familiar listeners, *Forensic Linguistics: The International Journal of Speech, Language and the Law,* 2, 1-17.

Schiller, N. O. and Köster, O. (1998). The ability of expert witnesses to identify voices: a comparison between trained and untrained listeners. *International Journal of Speech, Language and the Law,* 5(1), 1-9.

Schwartz, R., Campbell, J. P., Shen, W., Sturim, D. E., Campbell, W. M., Richardson, F. S., Dunn, R. B. and Granville, R. (2011). USSS-MITLL 2010 Human Assisted Speaker Recognition. *Proceedings of ICASSP 2011*, 5904-5907.

Shirt, M. (1984). An auditory speaker recognition experiment. *Proceedings of the Institute of Acoustics Conference*, 6, 101–4.

Sørensen, M. H. (2012). Voice line-ups: speakers' F0 values influence the reliability of voice recognitions. *The International Journal of Speech, Language and the Law*, 19(2), 145-158.

Yarmey, A. D., Yarmey, A. L., Yarmey, M. J., and Parliament, L. (2001). Commonsense beliefs and the identification of familiar voices, *Applied Cognitive Psychology*, 15, 283-299.

Yarmey, A. D. (2004). Common-sense beliefts, recognition and the identification of familiar and unfamiliar speakers from verbal and non-linguistic vocalizations. *The International Journal of Speech, Language and the Law,* 11(2), 267-277.

Court of Appeal Cases

The Queen v Anthony O'Doherty [2002] NICA 20, 19/4/02 ref: NICB3173 Court of Criminal Appeal Northern Ireland.

Regina v Ronald Flynn and Joe Philip St John [2008] EWCA Crim 970 2nd May 2008.

Appendix A – Dr Philip Harrison: Summary Curriculum Vitae

Philip Harrison is a forensic consultant specialising in the analysis of speech, audio and recordings. He has seventeen years' experience of forensic work at J P French Associates, having worked on over 1000 cases in the areas of authentication, enhancement, transcription and speaker comparison, as well as many miscellaneous cases.

He holds a first class honours degree (BEng) in Acoustical Engineering from the Institute of Sound and Vibration Research, University of Southampton, an MA with Distinction in Phonetics and Phonology and a PhD in Linguistics from the Department of Language and Linguistic Science at the University of York.

He is an elected Member of the Institute of Acoustics (MIOA), of the International Association for Forensic Phonetics and Acoustics, and of the British Association of Academic Phoneticians. He is a committee member of the Speech and Hearing Group within the Institute of Acoustics.

He was appointed a specialty assessor in the area of Audio Analysis for the Council for the Registration of Forensic Practitioners (CRFP) before its closure and was responsible for producing the assessor documentation for the Speech and Audio Analysis specialty. He is currently part of a group writing the appendix for Speech and Audio analysis for the UK Forensic Regulator's Codes of Practice and Conduct.

He holds the position of Teaching Fellow in the Department of Language and Linguistic Science at the University of York and is involved in lecturing on a post-graduate course on forensic speech science. He has been employed as a tutor on forensic speech analysis at the International Summer School in Forensic Linguistic Analysis.

He is actively involved in research in the areas of forensic speech and audio analysis, and regularly delivers lectures and presentations to academic conferences and universities in the UK and abroad. He was also centrally involved in the formulation of a new framework for expressing conclusions in forensic speaker comparison cases which has now been universally adopted by experts within the United Kingdom.

He has appeared as an expert witness in the Court of Appeal (Criminal Division), Crown Courts in England, including the Central Criminal Court (the Old Bailey), the Sindh High Court, Karachi (court removed to England to hear expert evidence) and the High Court in Accra, Ghana. In criminal cases he undertakes work for both prosecution and defence.

Relevant publications, high profile cases and further information can be found in a more detailed CV at: http://www.jpfrench.com/staff/philip-harrison/