Forensic Science International 208 (2011) 59–65

Contents lists available at ScienceDirect

Forensic Science International

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

An empirical estimate of the precision of likelihood ratios from a forensic-voice-comparison system

Geoffrey Stewart Morrison [a,b,*], Cuiling Zhang [c,a], Philip Rose [a]

[a] School of Language Studies, Australian National University, Canberra, ACT 0200, Australia
[b] Forensic Voice Comparison Laboratory, School of Electrical Engineering & Telecommunications, University of New South Wales, Sydney, NSW 2052, Australia
[c] Department of Forensic Science & Technology, China Criminal Police University, Tawan Street 83, Huanggu District, Shenyang, Liaoning 110854, China

ARTICLE INFO

Article history:
Received 7 January 2010
Received in revised form 27 October 2010
Accepted 2 November 2010
Available online 4 December 2010

Keywords:
Forensic voice comparison
Validity
Reliability
Accuracy
Precision
Credible interval

ABSTRACT

An acoustic–phonetic forensic-voice-comparison system was constructed using the time-averaged formant values of tokens of 61 male Chinese speakers' /i/, /e/, and /a/ monophthongs as input. Likelihood ratios were calculated using a multivariate kernel density formula. A separate set of likelihood ratios was calculated for each vowel phoneme, and these were then fused and calibrated using linear logistic regression. The system was tested via cross-validation. The validity and reliability of the results were assessed using the log-likelihood-ratio-cost function (C_{llr}, a measure of accuracy) and an empirical estimate of the credible interval for the likelihood ratios from different-speaker comparisons (a measure of precision). The credible interval was calculated on the basis of two independent pairs of samples for each different-speaker comparison pair.

1. Introduction

In a typical forensic-voice-comparison scenario, a forensic scientist is provided with two speech recordings, one a recording of the voice of a known speaker and the other a recording of the voice of a speaker whose identity is in question. Within the likelihood-ratio framework for the evaluation of forensic evidence, the task of the forensic scientist is to calculate a likelihood ratio (*LR*) which is an evaluation of the probability of observing the acoustic (or other relevant) differences between the voice samples under the hypothesis that the questioned-voice sample was produced by the same speaker as the known-voice sample versus under the hypothesis that the questioned-voice sample was produced by a different speaker, Eq. (1).

$$LR = \frac{p(\text{observed acoustic difference} \mid \text{same speaker})}{p(\text{observed acoustic difference} \mid \text{different speakers})} \quad (1)$$

The same-speaker hypothesis usually relates to the prosecution's proposition that the accused is the offender, and the different-speaker hypothesis usually relates to the defence's proposition that the offender is someone other than the accused. A likelihood ratio greater than one lends support to the same-speaker hypothesis, e.g., if the trier of fact is provided with a likelihood ratio of 100 based on the voice evidence, then, whatever their belief before the evidence was presented, they should now be one hundred times more likely than before to believe that the two voice samples were produced by the same speaker. A likelihood ratio less than one lends support to the different-speaker hypothesis, e.g., if the trier of fact is provided with a likelihood ratio of 1/100 based on the voice evidence, then, whatever their belief before the evidence was presented, they should now be one hundred times more likely than before to believe that the two voice samples were produced by different speakers. The likelihood ratio framework is recommended by numerous forensic statisticians and forensic scientists, both for forensic comparison in general [1–7] and for forensic voice comparison in particular [8–15].

Presently, the issue of validity and reliability is of great concern in forensic science [16–19]. In statistics and general scientific literature "reliability" is synonymous with "precision" and "validity" is synonymous with "accuracy"; however, in judicial and forensic-science literature "reliability" has often been discussed without explicit definition, or has been defined in terms of a measure of validity: classification-error rates, i.e., the proportion of same-origin comparisons in a test set which a forensic-comparison system classifies as different-origin comparisons (misses), and the proportion of different-origin comparisons in the test set which the system classifies as same-origin comparisons (false alarms). Such classification-error rates are

* Corresponding author at: Forensic Voice Comparison Laboratory, School of Electrical Engineering & Telecommunications, University of New South Wales, Sydney, NSW 2052, Australia. Tel.: +61 2 9385 6544.
E-mail address: geoff-morrison@forensic-voice-comparison.net (G.S. Morrison).

0379-0738/$ – see front matter © 2010 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.forsciint.2010.11.001

based on thresholded posterior probabilities, and are inconsistent with the likelihood-ratio framework [12]. Even if applied to likelihood-ratios rather than posterior odds, a loss function which assigns a penalty of 1 to a result on the "wrong" side of the $LR = 1$ boundary and 0 to a result on the "right" side of the boundary, would be incompatible with the likelihood-ratio framework. A likelihood ratio which supports a contrary-to-fact hypothesis by a large amount is worse than a likelihood ratio which supports a contrary-to-fact hypothesis by a small amount. Likewise a likelihood ratio which supports the consistent-with-fact hypothesis by a larger amount is better than one which supports it by a small amount. One could debate the exact shape of an appropriate loss function, but it is clear that it should be continuous and more heavily penalise likelihood ratios which are worse in the sense that they provide less support for the consistent-with-fact hypothesis or more support for the contrary-to-fact hypothesis.

In automatic speaker recognition and in forensic voice comparison an increasingly popular measure of *accuracy*[1] for a system outputting likelihood ratios is the *log-likelihood-ratio cost* (C_{llr}) [10,12,20–26]. This measure has the desired properties of being based on likelihood ratios, being continuous, and more heavily penalising worse results.

In contrast to the work on the development of a measure of accuracy appropriate for forensic likelihood ratios, there has been little work on the development of a measure of the *precision* of forensic likelihood ratios (although the need for such a measure has been recognised [12,27]). The present paper introduces the use of an empirical estimate of a *credible interval*[2] as a measure of the precision of a forensic likelihood ratio. The aim is that when a forensic scientist presents a likelihood-ratio in court, they would be able to make a statement of the following sort:

> "Based on my evaluation of the evidence, I have calculated that one would be X times more likely to obtain the acoustic properties of the voice samples if the questioned-voice sample had been produced by the accused than if it had been produced by someone other than the accused. Based on my calculations, I am 95% certain that it is at least X_{lower} times more likely and not more than X_{upper} times more likely".

Another aim is to be able to compare the precision of different forensic-comparison systems. In terms of forensic voice comparison, the system includes the choice of which parts of the acoustic signal to measure and which acoustic properties to measure, the procedures for the extraction of numeric representations of the acoustic information, the composition of the background sample, and the statistical procedures used for the calculation of the likelihood ratios on the basis of the numeric values. Although analytic solutions for calculating the precision of sub-components of the system may be possible, an analytic solution for the entire system may be intractable. In addition, different systems can differ in any or all sub-components, and an analytic solution developed for one system may not be applicable to another system, making it hard to compare systems. An empirical solution which treats the system as a black box and its output as point values is therefore preferred. The input to the black-box system is a pair of voice samples, and in a test situation we know the origin of each of those samples, and the output of the system is a likelihood ratio. We can obtain several samples from each speaker, and can therefore construct several independent (i.e., non-overlapping) pairs of samples for each pair of speakers compared. By entering multiple independent sample pairs from the same pair of speakers, one can estimate the within-group sample distribution of the likelihood-ratio output for this pair of speakers. Although samples from different pairs of speakers will lead to different average likelihood-ratio outputs, we can subtract the within-group mean from the individual outputs and pool data across different groups to estimate the pooled within-group distribution of the likelihood-ratio output. Given an estimate of this distribution, we can then calculate an estimate of the credible interval for the likelihood-ratio output of the system. This can be used as a summary measure for comparison with other systems or for presentation in court (if we assume uniform priors for the credible interval calculations the sample distribution becomes the posterior distribution). Because of practical limitations on the cost of data collection and analysis, in the present study we will only be using two independent sample pairs per pair of speakers, and only calculate the credible interval for different-speaker pairs (we would need to collect and analyse four samples from a speaker to generate two independent sample pairs for a same-speaker comparison).

There have been two basic approaches to forensic voice comparison based on acoustic measurement: *automatic* and *acoustic–phonetic* ([12] briefly summarises these two approaches, for longer descriptions see [11,13,14,20,29,30]). A typical automatic approach is based on features such as cepstral-coefficient values which can easily be extracted from voice recordings—measurements are taken every few milliseconds over the whole of the recorded speech. A traditional automatic system treats linguistic information as noise (unwanted variability), but deals with this signal-to-noise problem by automatically analysing massive amounts of data. In contrast, an acoustic–phonetic approach explicitly exploits linguistic information, e.g., tokens of a given phoneme in one voice sample will be compared with tokens of the same phoneme in the other voice sample. In an acoustic–phonetic approach the features extracted from the voice recordings are typically measurements such as vowel formant frequencies which are traditionally used by acoustic phoneticians in relation to theories of speech production and perception. Unlike automatic procedures, acoustic–phonetic procedures are intensive in terms of human labour since the identification of the segments to compare is typically performed manually and reliable measurement of features such as formants often requires substantial human supervision. The study reported in the present paper is part of a larger project aimed at improving the validity and reliability of forensic voice comparison by combining automatic and acoustic–phonetic approaches within the likelihood-ratio framework. The present study represents a relatively modest combination of automatic and acoustic–phonetic approaches similar to that reported in Ref. [21]. The methodology is essentially acoustic–phonetic but some analysis procedures developed for use with automatic approaches are applied.

Forensic-voice-comparison analyses were conducted on the monophthongs /i/, /e/, and /a/[3] in a database of recordings of

[1] Note that we are using the term "accuracy" here with respect to the extent of support of a likelihood ratio for the true hypothesis (same speaker or different speaker), which is known in the case of a test set. This "accuracy" could potentially be improved by measuring different acoustic properties, by improving the measurement procedures, or by improving the statistical modelling procedures. The term "accuracy" could also be applied to whether an estimated likelihood ratio is close to the true value given the true distribution of the variables (which could only be known if synthetic data were used). Two speakers could have very similar values for the acoustic properties measured and the true value of the likelihood ratio could then support the same-speaker hypothesis, or two samples from the same speaker could be very different and the true value of the likelihood ratio could then support the different-speaker hypothesis. The latter sense of "accuracy" is therefore quite different from the former.

[2] A "credible interval" is the Bayesian analogue of a frequentist "confidence interval", but is interpreted as a range of values which is believed to have a specified posterior probability, e.g., 95%, of containing the true value [28].

[3] The phonetic realisations of Standard Chinese /i/ and /a/ are [i] and [a], but /e/ has a number of discrete allophones conditioned by surrounding tautosyllabic segments [31, p. 144]. The present study made use of only the basic/unconditioned allophone [ə] occurring in words such as sè "colour" [sə\̀].