An Issue in the Calculation of Logistic-Regression Calibration and Fusion Weights for Forensic Voice Comparison

Geoffrey Stewart Morrison[1,2], *Tharmarajah Thiruvaran*[1], *Julien Epps*[1,3]

[1]Forensic Voice Comparison Laboratory, School of Electrical Engineering and Telecommunications, University of New South Wales, Sydney, Australia
[2]School of Language Studies, Australian National University, Canberra, Australia
[3]National ICT Australia, Sydney, Australia
geoff-morrison@forensic-voice-comparison.net, thiru@ee.unsw.edu.au, j.epps@unsw.edu.au

Abstract

Logistic regression is a popular procedure for calibration and fusion of likelihood ratios in forensic voice comparison and automatic speaker recognition. The availability of multiple recordings of each speaker in the database used for calculation of calibration/fusion weights allows for different procedures for calculating those weights. Two procedures are compared, one using pooled data and the other using mean values from each speaker-comparison pair. The procedures are tested using an acoustic-phonetic and an automatic forensic-voice-comparison system. The mean procedure has a tendency to result in better accuracy, but the pooled procedure always results in better precision of the likelihood-ratio output.
Index Terms: logistic regression, calibration, fusion, weights, forensic voice comparison, likelihood ratio

1. Introduction

1.1. The issue

Logistic regression has become a popular technique for calibration and fusion of scores in automatic speaker recognition and in forensic voice comparison [1–6].

It is a requirement of the procedure that the data used to calculate the weights for calibration/fusion include at least two recordings of each speaker (hereafter *A* and *B*) such that same-speaker (target) scores can be calculated. It is essential in forensic voice comparison that these be non-contemporaneous recordings so as to achieve a reasonable estimate of within-speaker variability – in casework the suspect and offender recordings are non-contemporaneous. The pairs of recordings should also ideally be matched to the speaking styles and channels of the suspect- and offender-recording pair.

Having two recordings per speaker affords the opportunity for multiple different-speaker (non-target) comparisons, as shown in Table 1. If we assume that there are differences in speaking style and/or channel between recording *A* and recording *B*, and we want to maintain the distinction for consistency with the suspect and offender recordings, then we will limit our use of different-speaker pairs to those for which the suspect model is based on a recording *A* and the offender data comes from a recording *B* (those marked with an asterisk in Table 1). Only making use of these two pairs also has the additional advantage that there is no overlap in membership between the pairs and they can therefore be treated as statistically independent, see [7].

Having two non-overlapping pairs of different-speaker comparisons presents a choice as to which scores to use for the calculation of weights for logistic regression calibration/fusion; those from the first pair, those from the second pair, those from the first and second pair pooled together, or the means of the scores from the two pairs. It may be that common practice is to pool the scores without giving the issue any thought; however, there are theoretical and potentially empirical factors to consider. The key point of this paper is that one should consider the issue and make an informed conscious decision.

Table 1: Possible different-speaker comparison pairs

	Suspect model	Recording	Offender data	Recording
	Spk01	*A*	Spk02	*A*
*	Spk01	*A*	Spk02	*B*
	Spk01	*B*	Spk02	*A*
	Spk01	*B*	Spk02	*B*
	Spk02	*A*	Spk01	*A*
*	Spk02	*A*	Spk01	*B*
	Spk02	*B*	Spk01	*A*
	Spk02	*B*	Spk01	*B*

2. Theoretical arguments

2.1. Argument favoring the mean procedure

In many automatic-speaker-recognition applications the ultimate task may be to make a categorical decision, and post-decision the value of the score or likelihood ratio (*LR*) used to make that decision is no longer of interest. In forensic voice comparison, however, accurate and precise estimation of the *LR* is essential. It is an *LR* which the forensic scientist presents to the court as a strength-of-evidence statement which the trier of fact will then consider when making their determination on the ultimate issue of guilt or innocence. The *LR* expresses the relative likelihood of obtaining the acoustic differences between voices on the suspect and offender recordings under the hypothesis that they were both produced by the same speaker versus under the hypothesis that the voice on the offender recording was produced by someone other than the suspect.

If there are two *LR* values, each based on a non-overlapping pair of recordings, then each can be considered an independent estimate of the true *LR* for the comparison of the pair of speakers (or single speaker). According to the central-limit theorem, a more accurate estimate of the true *LR* can be obtained by taking the mean of the two individual estimates. There is therefore an argument to be made that the appropriate set of log-likelihood-ratio scores to use for calculating the weights for logistic-regression calibration/fusion should be the (more accurate) means of each pair.

ISBN 978-0-9581946-3-1
Accepted after peer review of full paper