Effects of variability on numerical likelihood ratio calculations for forensic voice comparison

Vincent Hughes[1] and Paul Foulkes[1,2]
[1]*Department of Language and Linguistic Science, University of York*
[2]*J P French Associates*

Forensic voice comparison (FVC) accounts for the vast proportion of casework undertaken by forensic phoneticians (ca. 70%, French p.c.). Typically, the expert is provided with an incriminating sample of the offender and a known sample of the suspect, and asked to assess the possibility that the same individual is present in both. The likelihood ratio (LR) is generally accepted as the "logically and legally correct" (Rose & Morrison 2009: 143) framework for the assessment of such comparison evidence (Robertson & Vignaux 1995). The LR ensures that the expert provides a gradient estimation of the probability of the evidence under prosecution (same-speaker) and defence (different-speakers) hypotheses, thus preserving the *ultimate issue* (Lynch & McNally 2003:96) of guilt for the trier-of-fact.

In FVC these competing hypotheses are reduced to an assessment of the similarity and typicality of features across the two evidential samples. Typicality is dependent on patterns within the *relevant population*, and is quantified relative to a sub-section of that population. However the definition and delimitation of the relevant population in FVC has largely been overlooked as an issue which may compromise the reliability of LR output. Previous studies display a lack of consensus over the size of the reference sample and at best only an implicit awareness of the range of social and stylistic variability in speech production.

This study therefore assesses the effects on LRs when key dimensions are varied: (i) the number of speakers, (ii) the number of tokens per speaker, and (iii) the dialect mixture of the reference sample. Using polynomial estimations of F1 and F2 trajectories from spontaneous GOOSE vowels, LR comparisons were performed against a reference set of up to 120 speakers and 13 tokens per speaker. Results suggest that same-speaker LRs are robust until sample size decreases below 20 speakers and fewer than three tokens. However, variance and error are continually reduced with the inclusion of more data.

The dialect composition of the reference distribution was varied to test Rose's default assumption that, in the absence of a specific alternative proposition, the relevant population should be "same-sex speaker(s) of the language" (2004: 4). LRs based on GOOSE are presented for four sets of mock suspect and offender data where only one set matches the reference distribution for dialect. Although the magnitude of same-speaker LRs and severity of error are considerably higher for the 'mismatch' test sets, results indicate that LRs are more stable when dialect defining acoustic information, in this case F1, is removed. The positive practical implication of this is that reliable LR estimations may be possible using a broader definition of the relevant population for features which are expected to carry less region-specific information, such as short vowels and hesitation markers.