Defining the *relevant population* according to class and age when computing numerical likelihood ratios for forensic voice comparison

The likelihood ratio (LR) is increasingly accepted as the appropriate framework for the evaluation of forensic voice comparison (FVC) evidence. The LR involves an assessment of the similarity between suspect and offender samples, and their typicality in the *relevant population* (Aitken and Taroni, 2004). However, previous attempts to define the *relevant population* for FVC have consistently overlooked the complex sources of structured between-speaker variation in natural speech, which are known to affect the frequency of given phonetic variants in the population at large. This paper considers the effects on LR output of different definitions of the *relevant population,* with regard to socio-economic class and age.

Dynamic, time-normalised measurements of the F1, F2 and F3 of FACE /eɪ/ were extracted for 101 male speakers in the Canterbury Corpus of the Origins of New Zealand English (ONZE) database (Gordon et al., 2007). The raw Hz values were fitted with cubic polynomial curves and the coefficients used as input for LR comparisons. Three sets of homogeneous test data were created, defined by (1) class (working class), (2) age (younger) and (3) both class and age (young working class). In each of the three experiments, typicality was assessed using three sets of reference data: (a) 'match' (WC and/ or younger speakers), (b) 'mismatch' (MC and/ or older speakers), (c) 'mixed' (no class and/ or age controls). The distributions of calibrated $\log_{10}$ LRs and error rates (C_{llr}) were compared against the baseline results from the match condition.

The strength of mismatch LRs was generally overestimated compared with mixed or match LRs, although underestimation was found. Mixed LRs typically displayed wider interquartile and overall ranges, although their distributions were closer to the baseline than the mismatch results. Greater differences were found in experiment (2) than in experiment (1), due to the larger differences in FACE variants correlating with age versus class in NZE. Differences in error rates were, however, systematic across experiments (Fig. 1), with C_{llr} overestimated in the mismatch conditions. There was marginal underestimation of the C_{llr} in the mixed conditions but generally the value was much closer to that of the baseline.

These results have important implications for the application of the LR to FVC. A narrow but incorrect definition of the *relevant population* (mismatch) is, for this data set at least, more problematic than no controls (mixed). The magnitude of the effects depends on the source of between-speaker variation, the linguistic-phonetic variable of interest, the community from which the speaker(s) come and the individual pair of comparison samples. Further, there is an interaction between class and age, such that the magnitude of the difference between the match and mixed C_{llr} in experiment (3) is intermediate to that in experiments (1) and (2) (Fig. 1). As such, it is essential that experts are aware of the complex social stratification of phonetic variables, in order to understand the potential effects of different definitions of the *relevant population* on strength of evidence.

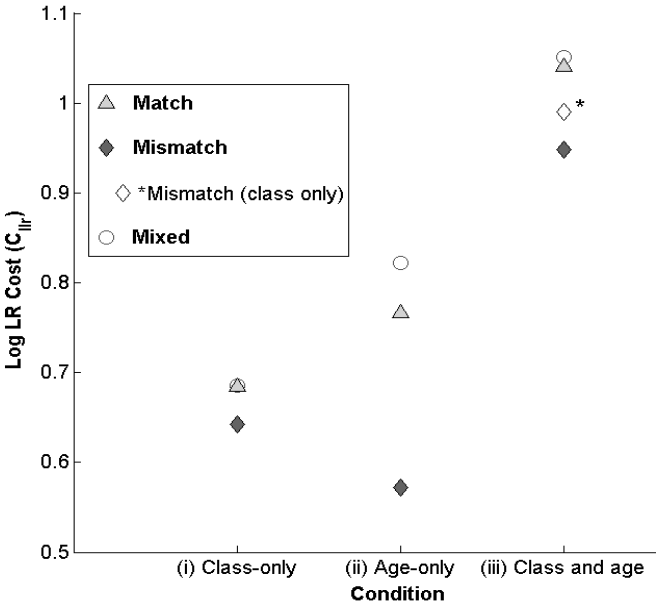

Figure 1: C_{llr} based on each set of reference data in each of the three experiments (where optimum performance = 0)