Comparative performance of real cepstral coefficients vs mel frequency cepstral coefficients for forensic voice comparison

Annie I-AN Lu[1], Yu Liu[1], Bernard J. Guillemin[1]*

[1]Department of Electrical & Computer Engineering, The University of Auckland, New Zealand

*bj.guillemin@auckland.ac.nz

The Mel frequency cepstral coefficients (MFCCs) have long been used for both speech and speaker recognition [1]. Some researchers have now started using cepstral coefficients for forensic voice comparison (FVC) purposes [2]. The aim of this preliminary study was to investigate whether MFCCs give improved performance over Real Cepstral Coefficients (RCCs) for FVC. The MFCCs and RCCs differ in respect to how information is extracted from the frequency spectrum. With MFCCs the frequency bands are equally spaced on the mel scale, which aims to approximate the human auditory system's response. Specifically it incorporates a frequency spacing which is approximately linear below 1000Hz and approximately logarithmic above it. With RCCs the frequency bands are linearly-spaced throughout.

The database used in this study was provided by Dr. Phil Rose and consists of forty-three male Australian speakers conversing with each other about an unclear fax message that was received. The speech samples were recorded on three different occasions over a period of one to two months. The speech was therefore non-contemporaneous, spontaneous, and coming from a background population of similar sounding people, thus meeting a number of the requirements for being forensically realistic. The database was divided into three sub groups: 12 speakers for Development, 12 speakers for Testing, and 19 speakers for the Background population. Comparisons were based upon sixteen MFCCs and sixteen RCCs computed from multiple tokens (at least two per recording session) of vowel segments extracted for common words/phrases spoken by all speakers, namely the diphthong /ai/ from *size* and the monophthongs /ɒ/ from *model*, and /i/ from *PTZ*. The Development group was used to both calibrate results for individual vowels and fuse results for multiple vowels using logistic-regression fusion. The Bayesian likelihood-ratio framework was used for comparison and the accuracy of a FVC experiment was determined using log-likelihood ratio cost (C_{llr}). Likelihood ratios (LRs) were computed using principal component analysis kernel likelihood ratio (PCAKLR), a routine recently developed at The University of Auckland for computing LRs [3]. Results were also analysed using Tippett plots.

Results in respect to C_{llr} were consistent for all three vowels and showed that MFCCs gave better FVC performance than RCCs. In respect to fused results for all three vowels, for RCCs $C_{llr} = 0.76$, and for MFCCs $C_{llr} = 0.26$. Fig. 1 shows the Tippett plot of these fused results. Though for same-speaker comparisons RCCs outperformed MFCCs, the reverse was true for different-speaker comparisons, particularly for log-likelihood ratios close to zero.

These results are very preliminary and much more investigation is required, but they do suggest that MFCCs might be a better choice than RCCs for FVC.

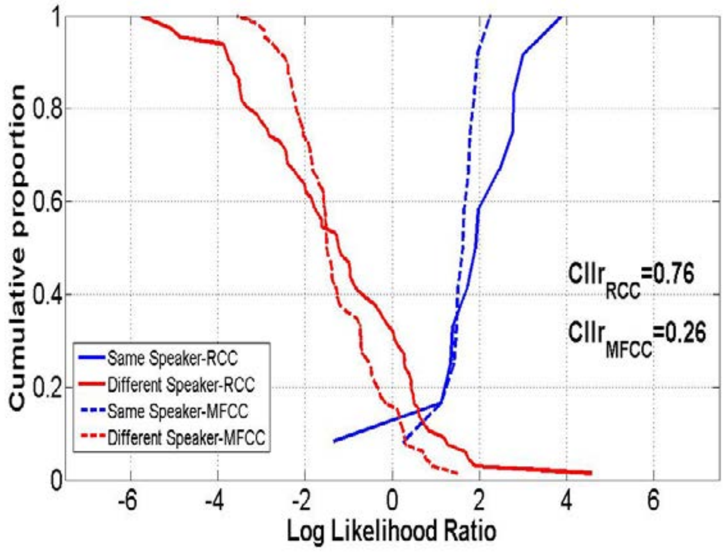

Figure 1: Tippett plot of fused results for all three vowels

References

[1] Muda, L., Begam, M., Elamvazuthi, I., “Voice recognition algorithms using Mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques”, Journal of Computing, Vol. 2, issue 3, 2010.

[2] Rose, P., Forensic Voice Comparison With Japanese Vowel Acoustics - A Likelihood Ratio-Based Approach Using Segmental Cepstra. The 17th International Congress of Phonetic Sciences(ICPHS XVII), 2011.

[3] Nair, B., Alzqoul, E.A., Guillemin, B.J., Detertemination of likelihood ratios for forensic voice comparison using principal component analysis, Forensic Science International (currently under review).