FORENSIC VOICE COMPARISON WITH JAPANESE VOWEL ACOUSTICS – A LIKELIHOOD RATIO-BASED APPROACH USING SEGMENTAL CEPSTRA

Phil Rose

School of Language Studies, Australian National University, Australia
philip.rose@anu.edu.au

ABSTRACT

The suitability of vowel cepstral spectra for forensic voice comparison is explored within a likelihood ratio-based framework. Non-contemporaneous landline telephone recordings of 297 male Japanese speakers are compared using only two replicates each of their five vowels. 14 cepstrally-mean-subtracted LPC CCs from dc to 5 kHz are used as features. Multivariate likelihood ratios estimated for the 297 target- and 43956 non-target trials give good results: an equal error rate of 0.28% and log likelihood ratio cost of 0.013. It is concluded that the approach has some merit.

Keywords: forensic voice comparison, likelihood ratio, vowel spectra, cepstrum

1. INTRODUCTION

In forensic voice comparison (FVC) the expert typically compares suspect and offender speech samples to help the trier of fact decide whether the suspect said the incriminating speech. The evidence in FVC is the ensemble of observed differences between the suspect and offender speech samples, and the crucial concept is its strength, quantified as a likelihood ratio (LR) [7]. This is the ratio of the probabilities of observing the evidence under the competing hypotheses – usually the prosecution hypothesis that the suspect said both speech samples, and the defence hypothesis that they were said by different speakers. The LR is crucial not only because the usefulness of an expert is considerably restricted if they are unable to say how likely the evidence is under both prosecution and defence hypotheses. The LR has also been shown to be a powerful function in the essential testing, as is done in this paper, of the discriminability of various forensic media, e.g DNA [3] and speech [5, 8].

The strength of the evidence/magnitude of the LR in a particular case depends on many factors, but *ceteris paribus* on the features used to compare the samples. Cepstral coefficients (CCs) have long been the feature of choice in automatic speaker and speech recognition, and automatic forensic speaker recognition [5], where they are applied globally. The general aim of this paper is to advance current investigations as to how automatic speaker recognition methods can be used to enhance traditional FVC; specifically it is to explore the potential of CCs when used locally, to characterise segmental phones such as vowels.

Cepstral coefficients represent a way of parametrising a spectrum with a greater degree of smoothing than with linear prediction: a degree of smoothing which generally exhibits '… strong immunity to non-information variabilities in the speech spectrum' [9], and thus turns out to be optimum for speech and speaker recognition. The main forensic advantage of the cepstrum is its power. For example, it has been shown capable, albeit with a small sample of 60 speakers, of delivering much stronger evidence, i.e. greater magnitude LRs, for the same data than formant centre-frequencies alone [12]. This is probably simply because, in capturing the whole of the spectral envelope, more potentially speaker-specific information is able to be exploited. For example, the cepstral-spectral envelope of a vowel can be expected to reflect not only the dimensions of the tract that produced it (in its F-pattern centre-frequencies - to the extent they are resolved), but also aspects of the phonatory activity of the source (in its spectral slope). A second advantage over formants is that the cepstrum is much easier to extract. Formants, including forensically important sub-glottal and nasal resonances, can be notoriously difficult to identify and extract in the often degraded speech encountered in real case-work. The major drawback of the cepstrum is of course its sensitivity to transmission effects (compared, say, with vocalic F2); another is its general lack of interpretability in terms of speech production.

2. PROCEDURE

2.1.1. Database, speakers, corpus

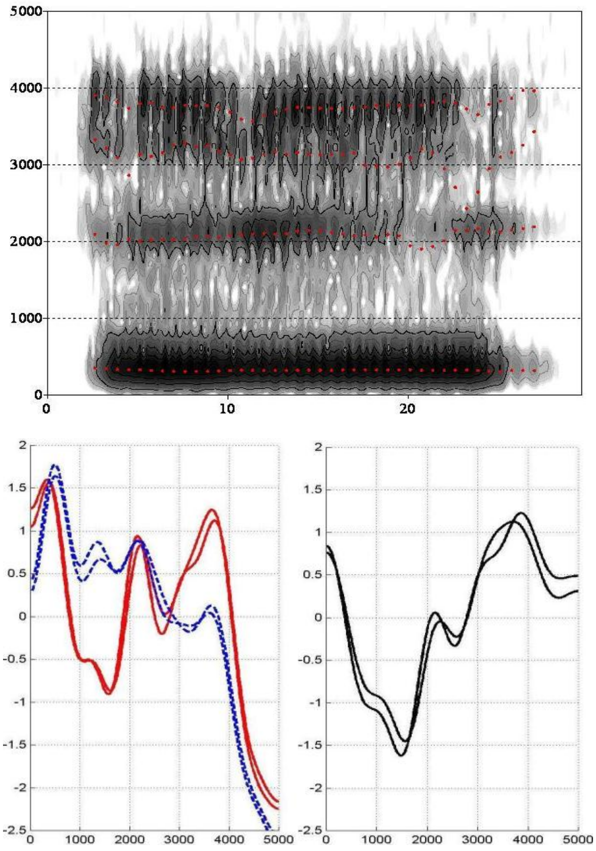
The database used in this experiment was collected some time ago by the Japanese *National Research Institute of Police Science* (NRIPS) for forensic speaker recognition tests, and allows testing with a reasonably large number of speakers. It comprises recordings, digitized at 10 kHz with 12 bit quantization, of 297 adult male Japanese from 11 different prefectures around Japan. This means nearly 300 same-speaker comparisons and 44,000 different-speaker comparisons can be made. All speakers were members of the Japanese police force and were uncontrolled for age, which ranged from ca. 20 to 50 years. The recordings were made centrally, on the same NRIPS equipment, of incoming landline telephone calls. Most importantly for realistic forensic testing, two non-contemporaneous recordings were made for each speaker, separated by three to four months. Each recording for each speaker contains about 70-80 seconds net speech comprising single- and many-word utterances, and a set of five Japanese vowels read out from hiraganaい え あ お う representing the five Standard Japanese vowel phonemes /i e a o ɯ/. It is these vowel readings that were used in the experiment. In each recording all the data was repeated, giving just two replicates of each vowel per non-contemporaneous recording session. Click here for examples from the first part of two speakers' recordings (vowels, numbers).

The CCs used in this experiment had already been extracted from the speakers' vowels in Khodai-joopari's forensically motivated Ph. D thesis [6]. In order to extract representative CCs for the vowels' spectra, Khodai-joopari first identified a given vowel token's *best continuous interval* (BCI). This contained all the speech wave-form within 10% of the maximum amplitude of the vowel, and, by visual examination, no extraneous material (e.g. non-speech transients). From the 14th order autocorrelation LP of the vowel's BCI, four consecutive single 25.6 ms. frames were then identified such that a single frame could be chosen from them which simultaneously (1) best represented the whole of the BCI, and (2) minimized the variance across a speaker's four vowel tokens. The CCs from this frame were then used to characterise the vowel token in question. In this way the within-speaker spectral variability was claimed to be minimised. This final part of the procedure was of course forensically unrealistic, because it rendered the two non-contemporaneous recordings for each speaker as similar as possible. In reality, one does not of course know whether suspect and offender speech samples are from the same speaker or not, and one can imagine what defence counsel would have to say if it transpired that the forensic voice comparison expert had tried to make them as similar as possible! Nevertheless, this minimization of same-speaker variance means that we can expect the discrimination of same-speaker speech samples to probably be optimal, with consequent degradation in different-speaker comparisons.

2.1.2. Further processing

For this experiment, the CCs extracted by Khodai-joopari were further processed in the following way. In order to at least partially compensate for the inevitable spectral distortion caused by all aspects of the 'phone transmission, a set of cepstrally-mean subtracted CCs (cms-CCs) was obtained by subtracting each speaker's CCs for a given vowel token from the mean cepstral vector obtained from their whole repeat. This process is illustrated in figure 1, with data from speaker 52's [i] vowels. A *Praat* wideband spectrogram of his first [i] token, with superimposed formant centre-frequencies (*Burg,* 4 formants below 5 kHz), shows the typical diffuse-acute F-pattern for [i]. F1 is at ca. 300 Hz; F2 at just over 2 kHz; a weak F3 appears to meander around 3.3 kHz, and F4 is at about 3.8 kHz. Khodai-joopari's cepstral spectra for speaker 52's non-contemporaneous [i] means (shown with solid red lines in the bottom left panel) correspond fairly well with the spectrographic F-pattern. Some of these spectral characteristics will be due to the transmission effect, and these will be contained in the speaker's mean cepstral spectra over their whole recordings shown in the dotted blue lines in the bottom left panel [4]. The most obvious feature is of course the upper passband cutting-in at just below 4 KHz. In subtracting the mean cepstrum from the [i] CCs one hopes at least partially to deconvolve the original signal from the channel [4]: the result is shown in the cms-CC spectra for the two non-contemporaneous means in the bottom right panel. Some changes in the vowels' overall spectral shape can be seen, notably in the increased spectral slope, which now corresponds better to the the auditorily slightly tense phonation type of the tokens. The [i] spectra difference shown in this comparison was ca. 390 times more likely with same rather than different speaker provenance ($\log_{10}$LR = 2.59).

Figure 1: Illustration of cepstral mean subtraction in [i]. Top: wideband spectrogram of [i] (axes = csec., Hz). Bottom left panel: red line = 12th order LPC cepstral spectrum of phone-recorded [i]; dotted blue line = speaker's mean cepstral spectrum. Right panel: cepstrally mean subtracted spectrum for [i]. Spectral axes = Hz, arbitrary amplitude units.



LRs for separate vowels were then estimated from the cms-CCs with the generative multivariate kernel density LR formula developed at Edinburgh University's *Joseph Bell Centre for Forensic Statistics and Legal Reasoning* [1]. This formula estimates multivariate LRs (MVLRs) taking into account any correlation between variables within a segment. (Although the variables in this case are CCs, which are orthogonal by definition, there is always the chance that correlation will arise by virtue of the spectral shape of the actual sound being modeled, as was shown for [ɕ] in [12]). With two non-contemporaneous recordings, two independent non-target trials are possible. Only one was tested, thus giving in all 43,956 non-target trials. LRs were also estimated with a Gaussian mixture model approach [8], but this, surprisingly, gave markedly poorer performance and is not discussed further.

Although the MVLR takes into account any correlation between features within a segment, any between-segment correlation must be also be handled [11]. This was done by logistic-regressive fusion [10], another common automatic FVR procedure, which combines the LRs from the different vowels according to the correlation between the vowels' LRs. The output is a set of calibrated LRs for all five vowels combined. The performance of the system was then quantified with its equal error rate (EER), and calibrated log likelihood-ratio cost Cllr [2]. Currently the evaluation metric of choice for the performance of LR-based detection systems, Cllr is a simple scalar which severely penalizes highly counterfactual LRs. Various subsets of the 14 cms-CCs were tested and it was found that optimum Cllr and EER were obtained with all 14.

3. RESULTS

Figure 2: Tippett plot for multivariate LRs derived from comparisons using 14 cms LPC CCs from all five vowels. X axis = log10LR greater than …; y axis = cumulative proportion of non-target trials ~ 1-cum.prop. target-trials. Solid black/red lines = cms-CCs; dotted green/blue lines = raw CCs. Inset = EER detail.

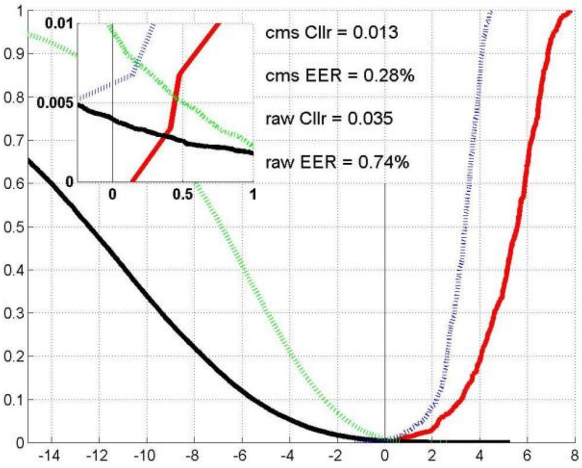


Figure 2 shows, with a conventional Tippett plot, the results for the MVLR-based discrimination using all five vowels (to show the beneficial effect of cepstral mean subtraction, results using the 14 raw CCs have also been included and plotted with dotted lines). Rising towards the right are cumulative Log_{10}LR curves for comparisons between samples from the same speaker; curves for different-speaker comparisons rise towards the left. The Cllr, at 0.013, is encouragingly low (values below unity indicate that the system is delivering some information; values below 0.1 are to be hoped for). The results show that vocalic segmental cms-CCs can be used within a likelihood-ratio based approach to discriminate rather well between same-speaker and different-speaker speech samples: the EER of ca. 0.28% for the cms-CCs reflects the fact that all of the 297 same-speaker comparisons were correctly evaluated with a LR that would be more likely had they come from same speakers; and out of the 43956 different-speaker comparisons, just 173 had counterfactual LRs. Of these, however, 14 had Log_{10}LRs bigger than 1000, and one comparison was over 100000 times more likely given same-speaker provenance (this is the price you pay for higher-order CCs: the magnitude of the worst different-speaker LRs drops considerably with lower-order analyses). It can be seen that the raw CCs also perform well, but with EER and Cllr not quite as good as the cms-CCs. Clearly, the MVLR likes this kind of data.

4. SUMMARY & DISCUSSION

This paper has used a large database of speakers to show how same-speaker vowel samples can be well discriminated under forensically realistic conditions of non-contemporaneity and telephone recording from different speakers' vowels using likelihood ratios derived from their segmental cepstrum. (Lest it be thought that isolated vowels are totally unrealistic, the author has more than once had to deal with forensic speech samples containing them when speakers spell out names). The good results undoubtedly reflect the favourable conditions of the comparison: considerable care was taken to ensure *a priori* minimal within-speaker variation; and isolated vowels, with their spectra unperturbed by consonantal effects, are also highly comparable. (There is minimal contribution to the good results from the Japan-wide sampling: vowels from different dialect areas do not differ, other than in the expected F2 difference for /high back vowels/ correlating with the well-known East-West difference in rounding [13].) The experiment has thus shown that the segmental cepstrum as a forensic tool deserves some further study – perhaps with non-vocalic sonorants; definitely with mobile phones. And once again the power of the LR is clear as forensic discriminant function.

5. ACKNOWLEDGEMENTS

This paper was written as part of Australian Research Council Discovery Grant DP0774115. Many thanks to Drs. Osanai, Khoodai-Joopari, Clermont: they did all the really hard work to collect the data and extract CCs. Thanks also to my two reviewers for their careful reading and advice. Usual disclaimers apply.

6. REFERENCES

[1] Aitken, C.G.G., Lucy, D. 2004. Evaluation of trace evidence in the form of multivariate data. *Applied Statistics* 53(4), 109-122.

[2] Brümmer, N., du Preez, J. 2006. Application independent evaluation of speaker detection. *Computer Speech and Language* 20(2-3), 230-275.

[3] Evett, I.W., Scrange, J., Pinchin, R. 1993. An illustration of the advantages of efficient statistical methods for RFLP analysis in forensic science. *Am. J. Human Genetics* 52, 498-505.

[4] Garcia, A.A., Mamone, R.J. 1999. Channel-robust speaker identification using modified-mean cepstral mean normalisation with frequency warping. *Proc. ICASSP* 1, 325-328.

[5] Gonzalez-Rodriguez, J., Rose, P., Ramos, D. Torre, D., Ortega-García, J. 2007. Emulating DNA: Rigorous quantification of evidential weight in transparent and testable forensic speaker recognition. *IEEE Trans. Audio Speech & Language Processing* 15(7), 2104-2115.

[6] Khodai-joopari, M. 2006. *Forensic Speaker Analysis and Identification by Computer. A Bayesian Approach Anchored in the Cepstral Domain*. Unpublished Ph.D. thesis, University of New South Wales.

[7] Morrison, G.S. 2010. Forensic voice comparison. In Freckelton, I., Selby, H. (eds.), *Expert Evidence*. Sydney: Thomson Reuters, Ch. 99.

[8] Morrison, G.S. 2011. A comparison of procedures for the calculation of forensic likelihood rations from acoustic-phonetic data: Multivariate kernel density (MVKD) versus Gaussian mixture model-universal background model (GMM-UBM). *Speech Comm*. 53, 24-256.

[9] Rabiner, L., Juang, B.-H.J. 1993. *Fundamentals of Speech Recognition*. Englewood Cliffs: Prentice-Hall.

[10] Ramos-Castro, D. Gonzalez-Rodriguez, J., Ortega-Garcia, J. 2006. Likelihood ratio calibration in a transparent and testable forensic speaker recognition framework, *Proc. IEEE Odyssey*.

[11] Rose, P. 2010. The effect of correlation on strength of evidence estimates in forensic voice comparison: Uni- and multivariate likelihood ratio-based discrimination with Australian English vowel acoustics. *Intl. J. Biometrics* 2(14), 316-329.

[12] Rose, P., Osanai, T., Kinoshita, Y. 2003. Strength of forensic speaker identification evidence - Multispeaker formant and cepstrum based segmental discrimination with a Bayesian likelihood ratio as threshold. *Intl. J. Speech Language and the Law* 10(2), 179-202.

[13] Shibatani, M. 1990. *The Languages of Japan*. Cambridge: CUP.