Proc. *Int'l Conf. on Acoustics Speech & Signal Processing* (ICASSP) pp. 5900-5903 Prague 2011.

FORENSIC VOICE COMPARISON WITH SECULAR SHIBBOLETHS - A HYBRID FUSED GMM-MULTIVARIATE LIKELIHOOD RATIO-BASED APPROACH USING ALVEOLO-PALATAL FRICATIVE CEPSTRAL SPECTRA

Phil Rose

School of Language Studies, Australian National University

ABSTRACT

The suitability of voiceless fricative spectra for forensic voice comparison is explored within a Likelihood Ratio-based framework. Non-contemporaneous landline telephone recordings of 99 male Japanese speakers are compared using only tokens of their voiceless alveolo-palatal fricative [ɕ]. A subset of mean-cepstrally-subtracted LPC CCs from the fricative spectrum from dc to 5 kHz is used. GMM/UBM and multivariate likelihood ratios are extracted for the 99 target and 4851 non-target trials, and fused with logistic regression. An EER of 7.4% and log-LR cost of 0.26 is demonstrated. It is concluded that the [ɕ] spectrum does have some individualising potential.

Index Terms— Forensic Voice Comparison, GMM/UBM, Multivariate Likelihood Ratio, coronal fricative spectra, cepstrum.

1. INTRODUCTION

In the last decade, forensic speaker recognition has experienced a remarkable paradigm shift [9]. Built around the likelihood ratio (LR) of Bayes' Theorem, approaches have been developed with both automatic speaker recognition and traditional phonetic methods which now make it possible to claim [5] emulation, at least in methodology, of what is now considered the gold-standard of forensic identification-of-the-source sciences, namely DNA profiling [2]. In Spain and Australia, LR-based forensic voice comparison evidence has also now been received in court.

As currently practiced, traditional and automatic LR-based forensic voice comparison (FVC) differ in several ways: in the use of generative *vs* discriminative LRs; the amount of processable information and the time required to complete an analysis; sensitivity to channel effects; and the potential strength of evidence achievable. The main difference, however, is in the information used. Traditional approaches are essentially local in both time and frequency domain, with traditional phonetic features more closely associated with linguistic units, like formant centre frequencies extracted from linguistically comparable items, e.g. the same word or phoneme. In traditional FVC the expert first scours the suspect and offender samples for comparable linguistic units, like vowel phonemes. The traditional acoustic properties of these units, e.g. formant centre-frequencies, are then quantified and LRs estimated. LRs for the different linguistic units are then combined to obtain an overall LR, e.g. [15].

The most commonly used traditional LR-based FVC feature is vowel formant centre-frequencies, in real case-work usually parametrised as point estimates of so-called formant target values. Their individualising potential has been fairly extensively tested e.g. [8]. Language contains many speech sounds other than vowels, however, and their forensic potential remains largely unexplored. In this paper I investigate the forensic potential of a segment from another major class of speech sounds – voiceless coronal fricatives (*coronal* means made by the tip and/or blade of the tongue). According to the Old Testament, these sounds were involved in possibly the earliest speaker recognition experiment, with dire consequences for many [10], but my choice here is motivated by a real-world case involving an AU$150 million telephone fraud which went to trial in 2008 [11]. The incriminating recording contained little offender speech, but it was highly comparable with the suspect samples in the phrase *not too bad*, and several tokens of the word *yes*, and I was able to derive a multivariate LR from this material based on vocalic F-pattern and intonational F0.

One conspicuous feature which I was unable to exploit, other than with a crude categorical LR, was the obvious similarity between offender and suspect /s/ spectra, both of which had a lower than normal cut-off frequency. The lack of a proper methodology for deriving LRs from fricatives, which are common speech sounds, suggested that this was an important FVC research question. This paper thus asks: what strength of FVC evidence can be expected from the spectral features of voiceless fricatives like [s] or [ʃ]? On the one hand, one expects coronal fricative spectra to have a certain amount of individualising potential by virtue of an advantageous between- to within-variance ratio arising from between-speaker differences in the dimensions of the rigid structures involved in their articulation – dentition, alveolar ridge, hard palate – provided that speakers have not lost teeth or had new dentures in the interim. On the other hand, these sounds' most well-defined spectral prominences are located in frequency regions likely to be compromised by telephone channel bandpassing.

It is unlikely that this essentially traditional FVC question can be satisfactorily answered without recourse to automatic signal processing methods, and a second aim of my paper is to explore how traditional and automatic methods can be profitably combined. In order to do this, some common features and back-end processing are recruited from automatic forensic speaker recognition. This paper will assess the FVC potential of just one type of voiceless coronal fricative: the alveolo-palatal (also called palatalized post-alveolar) [ɕ] in Japanese.

2. VOICELESS ALVEOLO-PALATAL FRICATIVE

The voiceless alveolo-palatal fricative [ɕ] is auditorily similar to the English so-called palato-alveolar sibilant fricative [ʃ] in *sheet*, but its constriction is further back, with a longer, narrower channel formed from a higher tongue body and blade [7]. The X-rays of [ɕ] in figure 1 show this fairly long palatal channel formed between

the antero-dorsum of the tongue and the middle of the hard palate. This results in a short front cavity and a much longer back cavity. Similar articulations can be seen in the nice sagital MRIs in [17]. Information on the spectral characteristics of [ɕ] can be found in the acoustic theory of speech production modeling in [16], and the estimation, in [17], of transfer functions from MRI-derived area functions of two Polish speakers' /ɕ/, compared with mean spectra of their actual productions. The important spectral characteristic of [ɕ] is a compact prominence located between ca. 2.5 and 4.0 kHz. This prominence is contributed by the λ/2 resonance of the palatal channel, tuned-up somewhat by its finite constriction and posterior flaring, and the λ/4 resonance from the front cavity, tuned up by its tapered shape and the non-finite impedance of the palatal channel. Below this main prominence the supralaryngeal articulation predicts two further poles. One is the λ/2 resonance associated with the back cavity (if of a sufficiently high amplitude to be spectrographically visible, this should be continuous with vocalic F2). Below that will be a Helmholz resonance which will be continuous with vocalic F1, but below the dynamic range in a normal spectrogram. A 17.5 cms. long vocal tract, with a constricted, ca. 4.5 cms. palatal channel and a 3 cms. front cavity will thus have poles at about 1.8 kHz (rear cavity), 2.9 kHz (front cavity) and 3.8 kHz (palatal channel). Zeros are also expected from the supralaryngeal configuration, mostly from the location of a source at the teeth. In addition to the supralaryngeal articulation, contributions are also expected as a consequence of the laryngeal gesture associated with voicelessness, namely abducted vocal cords. An open glottis means acoustic coupling with the trachea, and one would therefore also expect low frequency spectral peaks (for a male, the lowest just below ca. 1 kHz), and zeros, from the sub-glottal system.

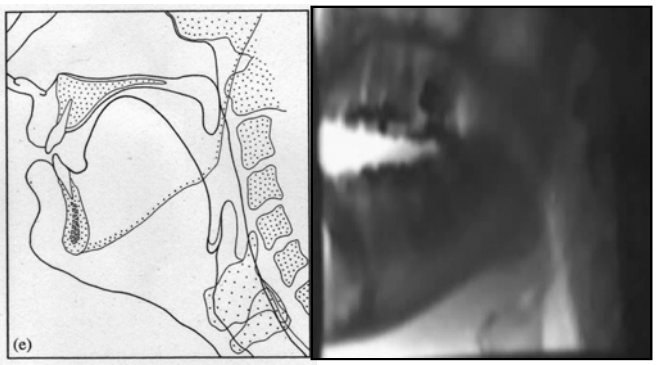

Figure 1. Left: mid-sagital section of tongue position in medial phase of alveolo-palatal fricative [ɕ] estimated from X-ray (reproduced from [6]). Right: X-ray movie of [aˈɕa] produced by author at the former *Adelaide Childrens' Hospital*. Click on figure to view movie.

Some of these features can be seen in the conventional wide-band spectrogram in figure 2, of the Japanese word kaisha [kaiɕa] *company*, said over the phone. The band-limiting of the telephone channel is clear, especially in the high frequency region, where no signal energy is evident above 4 kHz. The effect in the low frequencies is clear from the up-shifted F1 centre-frequency in the part corresponding to [i] – it is about 450 Hz, whereas one would expect a value closer to 250 or 300. The portion corresponding to the [ɕ] is clear, lasting about 10 csec., from ca. csec. 32 to csec. 44. It is dominated by narrowband energy constituted from two poles, from front cavity and palatal channel, at ca. 3.0 and 3.5 kHz (due to apparent variable formant-cavity association [17] one cannot say which pole reflects which cavity). There is also weaker energy centered at about 700 Hz from a sub-glottal resonance, and even weaker energy at 1.8 kHz from the rear cavity. Of note is the clear downwards perturbation at the end of the 3.0 kHz resonance as the tongue moves from [ɕ] to [a], and the drop in noise amplitude immediately preceding [a] periodicity which reflects removal of constriction with cords still abducted.

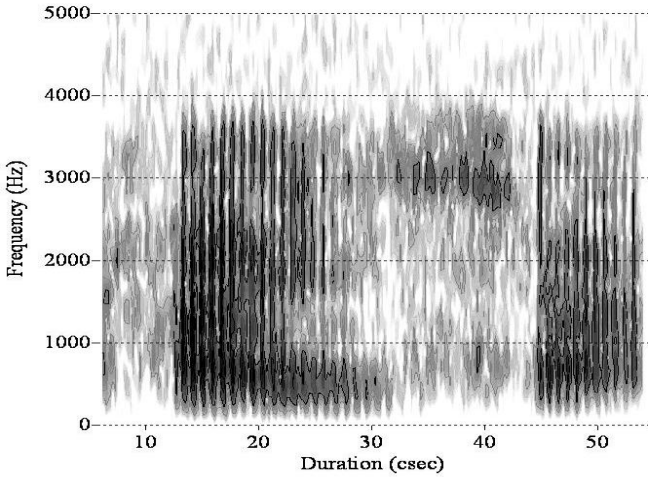

Figure 2. Spectrogram of Japanese word 会社 /kaiɕa/ *company*, showing acoustics of alveolo-palatal fricative. Listen here.

3. PROCEDURE

3.1. Recording, speakers, corpus.

Tokens of [ɕ] were taken from the first 99 speakers of a Japanese *National Research Institute of Police Science* database. This database contains recordings, digitized at 10 kHz with 12 bit quantization, of 297 adult male Japanese from 11 different prefectures around Japan. 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 equipment, of landline telephone calls. Most importantly, 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 two repeats of the set of five Japanese vowels, and several single-word and many-word utterances like kodomo wa daijōbu da *the kid's safe*, or bakudan o shikaketa *I've planted a bomb*.

The words and expressions from which the [ɕ] tokens were extracted are given in table 1, together with typical phonetic realisations (*typical*, because speakers do not invariably conform to prescribed Standard Japanese phonological behavior, and in real speech there is actually considerable between- and within-speaker variation). As can be seen, there were nine tokens of [ɕ] per repeat, seven occurring as the realisation of the palatal fricative phoneme /ɕ/ (romanised as 'sh'), and two as part of the realisation of the palatal affricate phoneme /tɕ/ (romanised as 'ch'). As there were two repeats per recording session, there were 18 tokens of [ɕ] in each recording. The long [ɕː] tokens occur as the result of so-called vowel devoicing, whereby high vowels in unaccented syllables become voiceless between two voiceless segments (the end of a word is also considered voiceless). When this occurs, the constriction for a preceding [ɕ] is simply continued for the

duration of the devoiced vowel. Thus in ashita /aɕitaˀ/ *tomorrow*, the high vowel /i/ occurs in an unaccented syllable between two voiceless consonants /ɕ/ and /t/, and so devoices. The constriction of the /ɕ/ is then prolonged, resulting in a long fricative segment [ɕː] occurring before a [t]. As a result of devoicing, the [ɕ] actually occurs in a variety of contexts, before [i a t k], and after [a i o t]. This contributes some within-speaker variation, as the [ɕ] assimilates a little to its surroundings.

Table 1. Corpus. Column 3: Standard Japanese romanisation sh = /ɕ/, ch = /tɕ/. Column 4: typical phonetic realisation.

もしもし	*hello*	moshimoshi	moɕimoɕi
私	*me*	watashi	wataɕi
会社	*company*	kaisha	kaiɕa
爆弾を仕掛けた	*I've planted a bomb*	bakudan o shikaketa	o ɕːkaketa
明日の朝	*tomorrow morning*	ashita no asa	aɕːta
金を用意しろ	*get the money ready*	kane o yōishiro	joːiɕiro
一	*one*	ichi	itɕː
八	*eight*	hachi	hatɕː

3.2. Front-end processing

Utterances containing [ɕ] were located and saved as separate audio files. Conventional wideband spectrograms of the utterances with superimposed intensity contours and formants (*Burg*, six formants below 5 kHz) were generated with *Praat*, and a spectrally maximally homogeneous portion of [ɕ] selected by eye. This portion was then saved in a separate .wav audio file and 12th order LPC CCs extracted. Cepstral subtraction was then performed using the mean cepstral vector obtained from the speaker's whole repeat. This process is illustrated in figure 3, which shows the cepstral spectrum for the [ɕ] token in figure 2 both before and after mean cepstral subtraction. The band-limiting of the telephone is clear in both the mean cepstral spectrum of the whole repeat and the spectrum of the segment in the left panel. Of interest is the cepstrally-mean-subtracted (cms) spectrum in the right panel, which appears to represent a partially restored typical spectrum for a fricative of this kind. Four peaks are obvious: the two highest amplitude peaks at ca. 2.9 and 3.6 kHz presumably reflect front cavity and palatal channel resonances; the two lowest peaks may reflect sub-glottal and back cavity resonance.

3.3. Back-end processing

An optimum set of cms CCs was empirically selected using a version of the multivariate LR formula developed at Edinburgh University's *Joseph Bell Centre for Forensic Statistics and Legal Reasoning* [1], followed by calibration [12] using Brümmer's *Focal* toolkit [3]. The CCs were optimized for the calibrated log likelihood ratio cost Cllr and equal error rate. Cllr is currently the evaluation metric of choice for the performance of LR-based detection systems. It is a simple scalar mean of two hypothesis-dependent logarithmic functions, one for all target LRs and one for all non-target LRs, which severely penalizes highly misleading LRs [4]. It was found that a improvement in evaluation of target LR performance was obtained by discarding four cms CCs (4 6 8 9), and so the remaining six cms CCs were used.

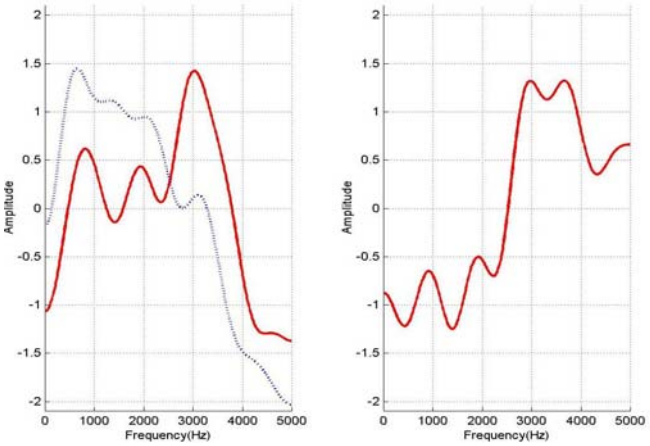

Figure 3. Illustration of cepstral mean subtraction in [ɕ]. Left panel: red line = 12th order LPC cepstral spectrum of phone-recorded [ɕ] in [kaiɕa] from figure 2, dotted line: speaker's mean cepstral spectrum. Right panel: cepstrally-mean-subtracted cepstral spectrum for [ɕ].

LR-based forensic voice comparison was then performed on the 99 speakers' non-contemporaneous [ɕ] cms CCs using both multivariate (MV) and GMM/UBM likelihood ratios [13]. With two non-contemporaneous recordings, two independent non-target trials are possible. Only one was used - between speakers' first recording sessions - thus giving 99 target and 4851 non-target trials.

4. RESULTS

Figure 4 shows, by means of the now conventional Tippett, or reliability plot, the results of the FVC using both MV and GMM/UBM LRs. Although both show an EER of ca. 10% and also have effectively the same calibrated Cllr values (given in the figure), their overall performances differ considerably, involving a quasi 20º rotation around threshold. (This brings into question the adequacy of Cllr as a metric, of course.) The multivariate analysis is clearly superior to the GMM in non-target trial resolution: half the non-target MV LRs are smaller than ca. log10 -2, compared to log10 -1 for the GMM LRs. More importantly, the maximum counterfactual non-target MV LR is ca. log10LR = 2, compared to just under log10LR of 5 for GMM. From a legal perspective (avoid convicting the innocent rather than exonerating the guilty) the higher GMM values are clearly undesirable. The superiority is reversed for target trials, at least as far as the higher LR values are concerned. The performance of MV and GMM systems is almost identical for the lowest 50% of the cases, with the worst counterfactual values remaining nicely below log10LR = ca. -1.5. The GMM/UBM is clearly capable of very much higher strength of evidence for the upper 50% of target trials.

FVC is ultimately about estimating the strength of voice evidence in a specific case. Despite the encouraging similarity in the EER and Cllr of the MV and GMM/UBM results, the individual LR estimates of the two approaches differ in many respects, and, since Cllr indicates no obvious preference, it is best to fuse rather than chose. As shown in figure 5, the result is an

overall improvement: Cllr reduces to 0.26, EER to ca. 7.4%. The magnitude of the worst counterfactual non-target LRs remains a worry, however.

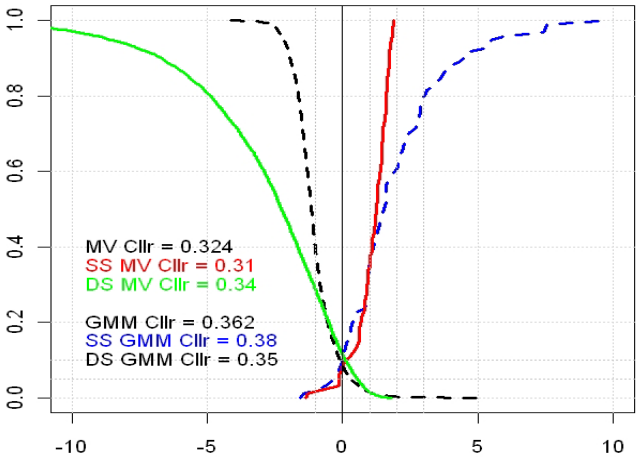

Figure 4. Tippett plot for multivariate (solid) and GMM/UBM (dashed) LRs derived from comparisons using cms LPC CCs from [ɕ]. X axis = log10LR greater than …; Y axis = proportion of non-target trials ~ 1/proportion of target trials. SS = target, DS = non-target trials.

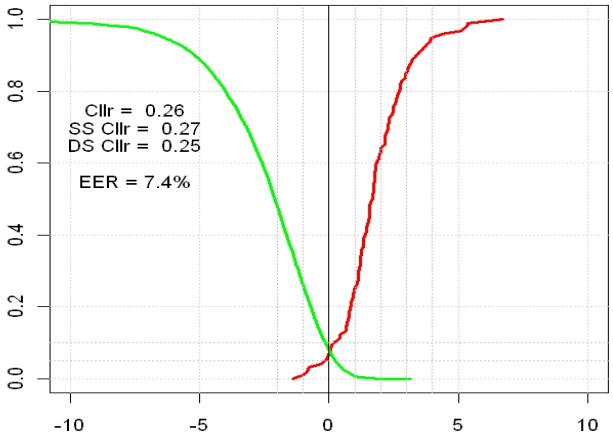

Figure 5. Tippett plot for logistic-regressively fused MV & GMM/UBM LRs derived from comparisons using mean-cepstrally-subtracted LPC CCs from [ɕ]. X & Y axes as for fig. 4.

9. SUMMARY & CONCLUSION

A part traditional, part automatic, FVC procedure has been described using spectra from voiceless alveolo-palatal fricatives. By fusing already reasonable LR estimates from both GMM/UBM and MV approaches, it has shown that [ɕ] is likely to be of use in FVC when its LR is combined with LRs from other segments. Other common coronals, e.g. [s] and [ʃ], should now be tested, and refinements to the method sought, in particular focusing on improved channel compensation, band-limited cepstra, and alternative features (e.g. MFCCs, spectral peak frequencies).

10. ACKNOWLEDGEMENTS

This paper was written as part of *Australian Research Council Discovery Grant* No. DP0774115. Many thanks, but no criticism, are due to Drs. Osanai, Clermont, Kinoshita for advice and help in processing. Thanks also to my four reviewers, one of whose advice I was able to take to increase my subjects (from an original 75).

11. REFERENCES

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

[2] D.J. Balding, *Weight of Evidence for Forensic DNA Profiles*, Wiley, Chichester, 2005.

[3] N. Brümmer, "Focal Toolkit" http://www.dsp.sun.ac.za/nbrummer/focal

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

[5] J. Gonzalez-Rodriguez, P. Rose, D. Ramos, D. Torre and J. Ortega-García, "Emulating DNA: Rigorous Quantification of Evidential Weight in Transparent and Testable Forensic Speaker Recognition", *IEEE Transactions on Audio Speech and Language Processing* 15/7, pp. 2104 – 2115, 2007.

[6] J.M.D. Laver, *Principles of Phonetics*, University Press, Cambridge UK, 1994.

[7] P. Ladefoged and I. Maddieson, *Sounds of the World's Languages*, Blackwell, Oxford, 1996.

[8] G.S. Morrison, "Likelihood Ratio forensic voice comparison using parametric representation of the formant trajectories of diphthongs", *JASA* 125, pp. 2387-2397, 2009.

[9] G.S. Morrison, "Forensic voice comparison and the paradigm shift", *Science & Justice* 49, pp. 298–308, 2009.

[10] Old Testament, Judges 12: 5,6.

[11] R v Hufnagl, Ernst [2008] NSWDC. See also http://www.abc.net.au/pm/content/2008/s2451596.htm

[12] D. Ramos-Castro, J. Gonzalez-Rodriguez and J. Ortega-Garcia, "Likelihood Ratio Calibration in a transparent and Testable Forensic Speaker Recognition Framework", *Proc. IEEE Odyssey*, 2006.

[13] D.A. Reynolds, T.F. Quaterieri and R.B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models", *Digital Signal Processing* 10, pp. 19-41, 2000.

[14] P. Rose, "Technical Forensic Speaker Recognition: Evaluation, Types and Testing of Evidence", *Computer Speech and Language* 20/2-3, pp. 159-191, 2006.

[15] P. Rose, "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", *International Journal of Biometrics* 2/14, pp. 316-329, 2010.

[16] K.N. Stevens, *Acoustic Phonetics*, MIT Press, Cambridge Mass., 1998.

[17] M. Toda, S. Maeda and K. Honda, "Formant-cavity affiliation in sibilant fricatives". In S. Fuchs, M. Toda, M. Żygis (eds.) *Turbulent Sounds*, Mouton De Gruyter, New York, pp. 343-374, 2010.