STRENGTH OF FORENSIC SPEAKER IDENTIFICATION EVIDENCE: MULTISPEAKER FORMANT AND CEPSTRUM-BASED SEGMENTAL DISCRIMINATION WITH A BAYESIAN LIKELIHOOD RATIO AS THRESHOLD

Phil Rose[1] Takashi Osanai[2] and Yuko Kinoshita[1,3]

[1]Phonetics Laboratory, Linguistics (Arts), Australian National University.
[2]Forensic Science Laboratory, Miyagi Prefectural Police H.Q., Japan.
[3]Division of Communication and Education, University of Canberra.

ABSTRACT: A forensic-phonetic speaker identification experiment is described which tests to what extent same-speaker pairs from a 60 speaker Japanese database can be discriminated from different-speaker pairs using the forensically appropriate Bayesian likelihood ratio (LR) as discriminant function. The strength of forensic-phonetic speaker identification evidence is quantified from the results. Non-contemporaneous telephone recordings are used, with comparison based on mean values from three segments only: a nasal, a voiceless fricative, and a vowel. It is shown that discrimination using the LR-based distance is better than with a conventional distance, and that the cepstrum outperforms the formants. A LR for the test of 50 is obtained for formant-based discrimination, compared to ca. 900 for the cepstrum, and the tests are thus shown to be capable of yielding a probative strength of support for the prosecution hypothesis that is conventionally quantified as 'moderate' for formants but 'moderately strong' for the cepstrum.

INTRODUCTION

Undoubtedly the most typical scenario in Forensic Speaker Identification (FSI) involves the comparison of one or more samples of an unknown voice with one or more samples of a known voice. Often the unknown voice is that of the individual alleged to have committed an offence, and the known voice belongs to the suspect. Obviously, both prosecution and defence are then concerned with being able to say on the basis of the evidence - the similarities and differences between the suspect and offender speech samples - whether the two samples have come from the same person, and thus be able either to identify the suspect as the offender, or eliminate them from suspicion.

However, stating the probability of the hypothesis given the speech evidence (p(H|E)) in this way is quite definitely *not* the aim of forensic speaker identification, and is both legally and logically incorrect (Aitken 1995: 4; Robertson and Vignaux 1995: Ch2; Rose 2002: Ch4). Rather, it is accepted that the aim of the forensic identification expert must be to determine the probability of the *evidence* under competing prosecution and defence hypotheses (H_p, H_d), and present the court with their ratio. This ratio - $p(E|H_p)/p(E|H_d)$ - is called the Likelihood Ratio (LR), is part of Bayes' theorem, and quantifies how much more likely one is to get the evidence assuming that the prosecution hypothesis is true than assuming that the defence hypothesis is true. Values of the LR greater than unity give support to the prosecution hypothesis that the forensic speech samples came from the same speaker; values less than unity indicate support for the defence hypothesis; the more the LR deviates, either way, from unity, the stronger the support. Unity, or values very close to unity, indicate that one is just as likely to get the evidence whether the speech samples come from the same speaker or from different speakers, and that it therefore has little or no probative value.

Since values of the LR greater than one indicate same-subject data, and values less than one indicate different-subject data, the LR is a potential discriminant function, and its ability to discriminate same-subject from different-subject data has to date been successfully tested on three different types of forensically common evidence: DNA (Evett et al. 1993: 503), glass fragments (Brown 1996), and speech. Of recently reported forensic speaker identification experiments using Bayesian methods on forensically realistic speech material, Meuwly and Drygajlo (2001: 150) for example, with Swiss-French speakers, obtained values of about 86% for same-speaker pairs with LRs greater than one, and 86% for different-speaker pairs with LRs less than one. Gonzalez-Rodriguez et al. (2001) for Spanish, and Nakasone and Beck (2001) with American English also show similar results. These experiments have all used a conventional automatic approach, with cepstral parameters, after channel normalisation, on long term speech data undifferentiated with respect to segments. In spite of its discriminant power, such automatic forensic speaker identification is still not a realistic option for real case-work, however, and is only used for investigative purposes (Nakasone and Beck 2001). Of importance and relevance to this paper is the LR-based experiment reported in Kinoshita (2001) with traditional parameters (i.e. formant centre-frequencies) on linguistically defined speech segments. Kinoshita (2001) showed with 11 male Japanese speakers how even with only six acoustic parameters from the formants of selected vowel and fricative segments, 90% of same-speaker pairs were resolved with LRs greater than one, and 97% of different-speaker pairs had LRs smaller.

Kinoshita (2001: 324) points out that her experiment was unrealistic in using a small number of speakers, and it is one obvious aim of the experiment reported here to see how well same-speaker pairs can be discriminated from different-speaker pairs with a LR-based approach with a considerably larger data base of 60 speakers. As an additional forensically realistic condition, we have retained the segment-based approach typical of proper forensic case-work and looked only at the discriminant power from a few (three) selected phonetic segments. However, we have tried to have the best of both traditional and automatic worlds by quantifying the segments with respect to both automatic (cepstrum) and traditional (formant centre-frequencies) parameters. This enables us to answer a further question concerning the relative discriminant power, and the associated strength of evidence, of these two approaches. Another realistic condition, of course, is the use of telephone recordings.

The most important question to be answered in the present experiment, however, is this: what is the average strength of forensic evidence to be expected from such an approach? That is, if, say, suspect and offender speech samples were to be compared using a test like this, and the value for the LR was determined to be bigger than unity, what is the strength of evidence to be communicated to the court?

$$\text{Strength of evidence} = \text{LR}_{test} = \frac{p(\,(\text{LR}>1) \mid \text{same-speaker pair})}{p(\,(\text{LR}>1) \mid \text{different-speaker pair})} \qquad (1)$$

The answer to this question is the likelihood ratio associated with the test (LR$_{test}$). This is the probability of the evidence assuming same-speaker data, relative to the probability of the evidence assuming different speaker-data, where the evidence is that the LR obtained from the comparison was greater than unity (1).The magnitude of LR$_{test}$ will clearly determine whether the strength of evidence is probative and such an approach is worth pursuing.

PROCEDURE

Part of the speaker database of the Japanese National Research Institute of Police Science (NRIPS) was used. This database contains recordings, digitised at 10 kHz with 12 bit quantisation, of 300 adult male speakers of Japanese from 11 different prefectures. All speakers are members of the Japanese police force aged between 20 to 50 years). Recordings were made centrally, on the same equipment, from landline telephone calls made from each prefecture. Two non-contemporaneous recordings were made for each speaker, separated by ca. three to four months. The recordings consisted of four types of material: the five Japanese short vowels /i/ /e/ /a/ /o/ /ɯ/; the numbers from /dzero/ *zero* to /jɯɯ/ *ten*; 26 individual forensically common words, like /giNkoo/ *bank*, /kane/ *money*; and 14 short polyword utterances like /asita no asa/ *tomorrow morning*. Each speaker repeated the material twice in both recordings.

The first 60 speakers of the database were chosen, and each repeat within each recording was treated separately. Two non-contemporaneous recordings, each containing two repeats, gave 240 same-speaker pairs, and 28,320 different-speaker pairs to be discriminated. The recordings of several speakers were listened to, and three segments selected for analysis on the basis of (1) high incidence in the data and (2) putative high individual-identifying potential. These segments were: (1) the so-called mora nasal /N/, (2) the voiceless alveopalatal fricative [ç], and (3) the long back mid rounded vowel [ɔː]. The mora nasal is a phoneme whose occurrence is restricted to the syllable coda. It counts as a separate timing unit and can carry a pitch accent. Its basic, prepausal, allophone is a nasal sonorant ranging between velar [ŋ] and uvular [N], but it also shows place assimilation with following segments, and is often realised without closure as a nasalised vowel (Vance 1987: 34-38). The mora nasal was chosen because nasal sonorants are assumed to be among the best segments for speaker identification / discrimination (Nolan 1983: 75-77), and the mora nasal in Japanese can be expected to have a fairly long, stable spectrum due to its phonological properties *qua* separate mora and pitch-accent carrier. There is also the intriguing observation (Vance 1987: 35) that the production of /N/ in some way directly reflects the speaker's articulatory setting. The [ɕ] occurs allophonically in Standard Japanese, at least under one phonemic analysis, as the realisation of the voiceless palatal fricative phoneme /š/, and the realisation of the fricated part of the voiceless palatal affricate /č/ (Vance 1987: 21-30). The long vowel [ɔː] is a sequence of two /o/ phonemes. The palatal fricative and [ɔː] were chosen because similar sounds had shown high F-ratios for one or more of their formants in the word moshimoshi *hello* (F3 in /o/; F2, F3 in [ɕ]) in previous forensic speaker identification experiments (Kinoshita 2001: 219). For convenience, these segments will be referred to below as "N", "sh" and "oo". Each sample consisted of seven tokens of N, 10 of sh, and 10 of oo, The words containing these tokens are listed in table 1.

Table 1. Corpus.

N (7)	sa**ɴ**, yo**ɴ**, de**ɴ**wa, bakɯda**ɴ**, bakɯda**ɴ** o šikaketa mooičido de**ɴ**wa sɯrɯ, gyakɯda**ɴ** sɯrɯna.
sh (10)	i[t**ɕ**ʰ]i, ha[t**ɕ**ʰ]i, mo[**ɕ**]imo[**ɕ**]i, wata[**ɕ**]i, kai[**ɕ**]ya, bakɯdaɴ o [**ɕ**]ikaketa, mooi[t**ɕ**ʰ]ido deɴwa sɯrɯ, a[**ɕ**]ita no asa, kane o yooi[**ɕ**]iro.
oo (10)	giɴk**oo**, k**oo**sokɯd**oo**ro, daij**oo**bɯ, ky**oo**, an**oo**, kodomo wa daij**oo**bɯda, ky**oo**jɯɯni, giɴk**oo**ni hɯrikome, m**oo**ičido deɴwa sɯrɯ.

As long a spectrally homogeneous portion as possible - up to aproximately 100 ms - was selected in the middle of a typical token of each of the three segments for each speaker, and dynamic programming was used to automatically identify a comparable portion in all their other tokens. The identification was checked auditorily in each case. Mean values for the first five formants, and a mean 12th order linear prediction cepstrum were then extracted from these portions. No channel normalisation was used. Mean values were then derived from all the tokens in each of the four samples (i.e. recording 1-repeat 1; rec1-rep2; rec2-rep1; rec2-rep2) of each speaker.

$$V \cong \frac{\tau}{a\sigma} \times \exp\left\{-\frac{(\bar{x}-\bar{y})^2}{2a^2\sigma^2}\right\} \times \exp\left\{-\frac{(w-\mu)^2}{2\tau^2} + \frac{(z-\mu)^2}{\tau^2}\right\} \quad (2)$$

similarity term typicality term

$\bar{x}$, $\bar{y}$ = means of offender and suspect samples; μ = mean of reference sample; σ = pooled standard deviations of offender and suspect samples; τ = standard deviation of reference sample; $z = (\bar{x} + \bar{y})/2$; $w = (m\bar{x} + n\bar{y})/(m + n)$; m, n = number in offender, suspect samples; $a = \sqrt{(1/m + 1/n)}$.

It was obviously important to use a forensically appropriate formula to calculate the LR, rather than one of the conventional speech science classification metrics. Thus the same formula was implemented as that derived for normally distributed continuous data by Lindley for refractive indices of glass fragments (Aitken 1995: 177-181), and used by Kinoshita (2001) in her forensic discrimination experiment with Japanese formants. This formula is reproduced at (2). Basically, it compares the suspect and offender samples against the background data for the relevant population, by quantifying the *similarity* between two sets of data relative to their *typicality* in the relevant population. Values of the LR greater than unity were taken to indicate same-subject data; values less than unity to indicate different subject data. The formula is not totally appropriate for speech acoustics (Rose 2001: 320ff.) because of unmet assumptions of normality and inadequate variance modelling. Nevertheless, as already mentioned, it has been found to perform promisingly on speech data (Kinoshita 2001). LR values were calculated using the formula at (2) for all 240 same-speaker and 28,320 different-speaker non-contemporaneous pairs. A cross-validation (leave-one-out) method was used, whereby every pair of speakers was evaluated against a reference population that did not include them.

A LR was calculated for the formants F1 through F5, both separately and combined, and for the combined 14 cepstral coefficients. The LR for combined data is simply the product of the LR for the individual data, if it can be assumed that the individual data are not correlated (Robertson and Vignaux 1995: 227ff.). There will of course be correlations both between certain of the formants in the three different segments; and also between the individual CCs. However, as their incorporation in the Lindley LR formula is still not clear, it was decided to treat the individual formants, and CCs, as effectively independent. LR values were also calculated for the three individual segments separately, and for all three combined (again by simply taking the product of the LRs for the individual segments). In addition to the discrimination with the LR, a conventional discrimination was carried out using EER thresholds derived from simple difference magnitudes (for individual formants) and Euclidean distances (for combined data).

RESULTS

Segmental acoustics The upper limit of the frequency range passed by the telephone was quite high at ca. 4.5 kHz, with a lower limit of about 300 Hz. Truncated, narrow bandwidth lowest resonances typical of phone transmission were evident for N and oo. The spectra of the individual segments require some comment. (1) **oo** With the exception of an extra peak at ca 2.0 kHz, there was good agreement between the values for the spectral peaks in the oo, and means for F-pattern in 11 male speakers' /o/ given by Kinoshita (2001: 146, table 4.5), viz: F1 = 462 Hz, F2 = 1125 Hz, F3 = 2508 Hz, F4 = 3553 Hz. (2) **sh** The spectrum of sh showed a fairly typical compact-acute profile expected for palatal/high front segments. According to Stevens (2000: 405-408) one expects the spectrum of a male palatoalveolar fricative [ʃ] to be dominated below 4 kHz by three peaks: the lowest, F2, at ca. 1.9 kHz, is the $\lambda/2$ resonance of the back cavity; the next highest, F3, at ca. 2.5 kHz, is the natural frequency of the palatal channel, and the highest, F4, at 3.5 kHz, is the resonance associated with the sublingual cavity. Although the sound in question is not [ʃ] but [ɕ], it was clear that there was a good deal of similarity in their values. With the exception of the lowest peak, there was also good agreement between the values for the spectral peaks in the sh, and the means for F-pattern in 11 male speakers' sh in *moshimoshi* given by Kinoshita (2001: 187, table 5.2) viz: 1.9 kHz (F2); 2.4 - 2.6 kHz (F3); 3.15 - 3.25 kHz (F4). Kinoshita measured the F-pattern in [ɕ] direct from spectrograms, on the basis of continuity with the surrounding vowel formants. It looks from a comparison with her data that we missed F3, and that our F3 is actually F4 (and our F4 actually F5). It is possible that the lowest spectral peak for sh at ca. 600 Hz, which we have labelled as "F1", is the expected helmholz resonance at somewhat less than 400 Hz (Stevens 2000: 385, 406) shifted up by the telephone transmission. Its non-truncated, symmetrical shape may also indicate a subglottal resonance typical in voiceless fricatives, although its amplitude is rather high for that. (3) **N** The mean N spectrum showed a low peak at ca. 300 Hz labelled as "F1", with a clear second "F2" peak at ca. 1.2 kHz. A third spectral peak occurred at ca. 2.2 kHz labelled "F3", above which the spectrum is fairly undifferentiated, but from which two peaks, labelled "F4" at ca. 3.1 kHz, and "F5" at ca. 4.1 kHz, were picked out. These values agree very well with those for a male velar nasal in Stevens (2000: 507ff.) which suggests a predominant realisation with little or no oral shunt, as in [ŋ] or [N].

Discrimination Generally, with the LR threshold of 1, different-speaker pairs were discriminated much better than same-speaker pairs. Typically, about 40% - 50% of same-speaker pairs were correctly discriminated whatever the condition, so about half of their probability density functions was typically distributed below the threshold of LR = 1.About 70% of different-speaker pairs were correctly discriminated for individual formants, with discrimination rates substantially improving on this with differing conditions. Thus, rates of ca. 90% and 98% were obtained for different-speaker pairs with combined formants and cepstrum respectively; and rates of ca. 99% and 99.98% were obtained for the combined three segments using formant and cepstrum respectively. This meant that just about all of the probability density function for different-speaker pairs lay below the LR = 1 threshold for the cepstral comparison based on three segments.

Specific performance results are given in table 2. They are arranged as follows. To the left of the table are shown the results for the individual and combined ("All") formant data; to the right are the results for the cepstral comparisons. Results for the three segments are given above, then for all three segments combined below ("All 3"). For each segment are the percent of the same-speaker and different-speaker pairs (SS_{LR} and DS_{LR} respectively) that were evaluated as same-speaker. So for example it can be seen that using "F1" in /N/, 70% of the 240 same-speaker pairs had LR values greater than 1 and were thus correctly discriminated, and 33.7% of the 28,000 odd different-speaker pairs had LR values greater than 1 and were therefore incorrectly discriminated.

The results in table 2 are clear. Firstly, from the point of view of the individual formant results, there are no large or consistent differences, with the best performance at ca. 69% being separated from the worst at ca. 55% by only 14%. Neither is there any superiority either for a particular segment (i.e. N vs sh vs oo.) or for the analysis approach (i.e. LR vs. conventional distance). (A two-way ANOVA on the individual formant performances gave p = 0.7 for segment; p = 0.65 for analysis; and p = 0.97 for interaction.) It is also clear that the individual formant approaches are inferior to both the cepstrum and combined formant approaches, which we now address. When the results for the combined formants are compared to those for the cepstrum, clear superiorities can be demonstrated with two-way ANOVA. Firstly, the cepstrum significantly outperforms the formant analysis ($p \leq 0.0001$) by ca. 10%, and secondly the LR-based distance significantly outperforms the conventional distance by ca. 3% (p = 0.053).

Table 2. Results of discriminant tests (%). SS_{LR} / DS_{LR} = percent of same-speaker / different-speaker pairs evaluated as same speaker with LR-based distance. VR_{LR} / VR_{Con}= verification rate with LR-based distance / conventional distance

	Formants						Cepstrum	
N	F1	F2	F3	F4	F5	All		
SS_{LR}	70.0	48.3	55.4	50.0	38.8	50.0	SS_{LR}	47.1
DS_{LR}	33.7	29.4	32.5	31.1	28.0	8.4	DS_{LR}	1.7
VR_{LR}	**67.6**	**59.6**	**62.3**	**58.8**	**56.2**	**76.3**	VR_{LR}	**79.0**
VR_{Con}	**68.8**	**57.5**	**62.1**	**59.2**	**56.2**	**70.8**	VR_{Con}	**80.0**
sh	F1	F2	F3	F4	F5	All		
SS_{LR}	46.4	51.7	64.2	37.9	60.0	47.9	SS_{LR}	52.5
DS_{LR}	32.5	32.9	35.5	27.4	34.0	9.2	DS_{LR}	2.1
VR_{LR}	**58.5**	**58.5**	**64.1**	**55.9**	**64.1**	**74.6**	VR_{LR}	**84.7**
VR_{Con}	**57.1**	**59.6**	**62.9**	**55.4**	**60.8**	**70.8**	VR_{Con}	**83.3**
oo	F1	F2	F3	F4	F5	All		
SS_{LR}	67.9	52.5	46.7	55.8	59.6	49.2	SS_{LR}	37.9
DS_{LR}	36.9	37.1	34.4	34.4	36.6	8.1	DS_{LR}	1.1
VR_{LR}	**65.0**	**56.9**	**55.8**	**60.9**	**60.6**	**72.7**	VR_{LR}	**82.7**
VR_{Con}	**64.6**	**57.5**	**55.0**	**58.3**	**60.0**	**67.5**	VR_{Con}	**80.8**

Next, the performance can be examined when data from all three segments are combined ("All 3"). It is clear that this constitutes an improvement over results from individual segments: there is a mean increase of ca. 7% in performance over the mean performance from individual segments. Again, the LR-based approach is better than the conventional distance, and the cepstrum is better than the formants. So the best overall verification rate - 89.6% - is obtained with a cepstrum using a LR based distance on the combined data from three segments.

All 3	Form	Cep
SS_{LR}	41.3	38.3
DS_{LR}	0.8	0.04
VR_{LR}	**81.0**	**89.6**
VR_{Con}	**79.2**	**87.5**

Finally, what is the strength of forensic-phonetic evidence associated with this kind of approach? As explained above, this is found by calculating the LR for the test using the formula at (1). It is clear that the strength of the evidence based on a test using a single formant will be low. For example, for “F1” in N and the LR threshold of 1, the LR_{test} is 70/33.1 = 2.07. One would be on average only about twice as likely to observe a LR greater than 1 for “F1” in N assuming same rather than different speakers. The strength of evidence clearly improves for a test based on a combination of all three segments and all five formants, where LR_{test} = 41.3/0.8 = ca 50. This means that one would be 50 times more likely, on average, to observe a LR greater than 1 from a combination of all three segments quantified by all five formants if the pair were from the same rather than different speakers. In terms of the verbal scale equivalents for the LR developed by the British Forensic Science Service (Rose 2002: 61), this would count as 'moderate' evidence in support of the prosecution hypothesis. For a test using combined cepstral data, however, the strength of evidence is actually quite good. With this approach, one would, on average, be ca. 900 times (i.e. 38.3/0.04, and allowing for rounding error) more likely to get a LR bigger than one assuming same rather than different speaker data. This - an LR between 100 and 1000 – would rate as 'moderately strong' evidence in support of the prosecution hypothesis. It is clear that the magnitude of the LR for the test is primarily a result of the powerful discrimination of different-speaker pairs: as already pointed out, the discrimination of same-speaker pairs (38.3%) is very poor on its own, but it is considerably bigger than the (0.04%) incorrectly identified different-speaker pairs. This is a good demonstration of the power of an approach which takes into account, as it should, not just the similarity between samples - the probability of getting the difference assuming same-sample provenance - but also their typicality - the probability of getting the difference at random from different speakers in the population (Aitken 1995: 180).

DISCUSSION

This paper has shown that, by comparing samples using combinations of just three segments – though with quite a few parameters! - this test is capable of giving probative evidence. It is also clear that, although probative forensic-phonetic evidence should be possible with formants, very much stronger evidence is possible with the cepstrum. This is possibly because the latter involves about three times as many parameters (another reason might be the difficulty of tracking formants in nasals.) This is clearly an indication that forensic speaker identification should also avail itself of the power of the cepstrum for segmental comparison. Many questions remain, the most important of which is: why does deriving a LR from the combination of segments have the desired result of causing the different-speaker probability density function to migrate lower than unity, but leave the same-speaker distribution unaffected and straddling unity? Perhaps the fact that unity was not a good threshold for same-speaker data is partly due to deviations from normalcy in the same-speaker distributions caused by between-speaker differences in non-contemporaneous within-speaker variation that are not being resolved well by the LR formula. It also needs to be determined to what extent the proper incorporation of interparametric correlations affects the value for the Lr_{test}. It would also be interesting to see if cepstral performance could be improved by the use of a band-selective cepstral distance rather than the whole Nyquist interval (Rose and Clermont 2001).

Finally, caveats are in order with respect to the application of the test in real world case-work. Firstly, the experiment described in this paper is an example of what has been called an 'average probability' approach (Aitken 1991: 70-77). It tells us how the LR test would evaluate pairs of speakers *on the average*, and is therefore a possible test of the LR approach. This, however, is clearly not the ideal in a proper LR-based forensic-phonetic identification, where an appropriate LR has to be estimated *for a particular case* (Rose 2002: 324 - 325). In other words, it is clear that if the segmental-cepstral comparison of two speech samples gives a LR bigger than one, this will, *on the average*, constitute strong support for the prosecution. However, the LR for an *actual* pair of samples may be very much bigger, or smaller than one, and be associated with very much stronger, or weaker evidence than suggested by the test. Ideally, then, one needs to know the actual LR for the data. Secondly, although we have conformed in this experiment to the realism criteria of: non-contemporanetiy, a large number of speakers, and use of telephone speech (cf Rose 2002: 96,97), the speech was still unnatural in being elicited and read-out. It is to be expected that performance would drop were we to use natural, unelicited speech.

ACKNOWLEDGEMENTS

This research was carried out whilst the second author was a visiting fellow in the Linguistics Department at the Australian National University. This fellowship was made possible by an International Training and Research Award from the Training Institute of Forensic Science of the Japanese National Police Agency, which the second author gratefully acknowledges. We also thank our two reviewers for their helpful comments, some of which we have tried to incorporate.

REFERENCES

Aitken, C.G.G. (1995) Statistics and the Evaluation of Evidence for Forensic Scientists, Wiley: Chichester.

Brown, K. (1996) Evidential value of elemental analysis of glass fragments, unpublished First Class Honours Thesis, University of Edinburgh.

Evett, I.W. Scrange, J. and Pinchin, R. (1993) "An Illustration of the Advantages of Efficient Statistical Methods for RFLP Analysis in Forensic Science", American Journal of Human Genetics 52: 498-505.

Gonzalez-Rodriguez, J. Ortega-Garcia, J. and Lucena-Molina, J.J. (2001) "On the Application of the Bayesian Framework to Real Forensic Conditions with GMM-based Systems", Proc. of the 2001 Speaker Odyssey Speaker Recognition Workshop.

Kinoshita, Y. (2001) Testing Realistic Forensic Speaker Identification in Japanese: A Likelihood Ratio Based Approach Using Formants, unpublished Ph.D. Thesis, the Australian National University.

Meuwly, D. and Drygajlo, A. (2001) "Forensic Speaker Recognition Based on a Bayesian Framework and Gaussian Mixture Modelling (GMM)", Proc. of the 2001 Speaker Odyssey Speaker Recognition Workshop.

Nakasone and Beck (2001) "Forensic Automatic Speaker Recognition", Proc. of the 2001 Speaker Odyssey Speaker Recognition Workshop.

Nolan, F. (1983) The Phonetic Bases of Speaker Recognition, Cambridge Studies in Speech Science and Communication. CUP: Cambridge.

Robertson, B. and Vignaux, G.A. (1995) Interpreting Evidence, Wiley: Chichester.

Rose, P. (2002) Forensic Speaker Identification, Taylor and Francis: London.

Rose, P. & Clermont F. (2001) A Comparison of Two Acoustic Methods for Forensic Speaker Identification, Acoustics Australia 29/1: 31-35.

Stevens, K. N. (2000) Acoustic Phonetics, MIT press: Cambridge, Mass.

Vance, T.J.(1987) An Introduction to Japanese Phonology, State University of New York Press: Albany.