Speaker variability on a continuum of spectral sub-bands from 297-speakers' non-contemporaneous cepstra of Japanese vowels

Mehrdad Khodai-Joopari
School of Inf. Tech. & Elec. Eng.,
University of N.S.W. (ADFA),
Canberra, Australia
m.khodaijoopari@adfa.edu.au

Frantz Clermont
Dept of Computer Science & Maths
American University of Paris,
Paris, France
frantz.clermont@aup.fr

Michael Barlow
School of Inf. Tech. & Elec. Eng.,
University of N.S.W. (ADFA),
Canberra, Australia
spike@cs.adfa.edu.au

Abstract

In this paper we describe the results of a preliminary investigation of inter- and intra-speaker variability in the vowel cepstra extracted from a forensically realistic database of non-contemporaneous telephone recordings of 297 adult, male speakers of Japanese. The methodology adopted to conduct this investigation is based on one-way ANOVA analysis, with the novelty of replacing the standard Euclidean distance with a more flexible band-selective formulation of the cepstral distance (Clermont and Mokhtari, 1994). The results of our investigation show a strong similarity between traditional formant-based F-ratios, and the cepstrally-based F-ratios obtained over consecutive sub-bands. The similarity reinforces the possibility to explore formant-related regions directly from the cepstrum. By virtue of the statistical depth of the data used, the results also provide convincing evidence that the whole-spectrum (i.e., full-band) approach based on the cepstrum, often overvalued in many studies, does not provide complete insights into the spectral regions of speaker-specific importance.

1. Introduction

It is well known that acoustic speech signals convey both phonetic characteristics and speaker-specific information, where the consequences of phonetic and speaker variability have been studied in terms of both discrete (Formant Frequencies) and a more complete (LP-Cepstrum) representations of the frequency spectrum (Sambur, 1975; Kitamura and Akagi, 1994; Mokhtari and Clermont, 1994-1996; Hayakawa and Itakura, 1994). The common motivation shared by all these studies is to improve the performance of Automatic Speech or Speaker Recognition (ASR) systems – firstly, by finding specific parameters, which are more directly related to either phonetic-quality or speaker-specific properties of the acoustic speech signal and, secondly, by exploring the frequency spectrum in search for the part(s), that best account(s) for the observed variations.

In this vein, the cepstrum coefficients have long been regarded as the most powerful parameter-sets in ASR, where it is assumed that cepstrally-based methods are superior to formant-based methods. Apart from actual performance, the cepstral coefficients are computationally less expensive, more robust, and easily extracted from the acoustic signal. They also provide a more complete representation of the spectral continuum with important correlates in both acoustic-phonetic, (Furui, 1986; Mokhtari and Clermont, 1996), and acoustic-auditory and perceptual domains (Zahorian and Jagharghi, 1993).

The dominance of the cepstrum is not found in the filed of forensic phonetics. This lack of prominence may be, in part, due to the difficulty of explaining to a jury the meaning of cepstral coefficients and, in part, due to their indirect correlates in auditory- and articulatory-phonetic (Rose, 1999: 7) terms. A counter argument to the latter could however be raised in the light of perceptual results reported by Zahorian et al. (1993: 1980), *"...perceptual confusions were more similar to confusions resulting from classification based on spectral shape features than to those resulting from classification based on formants, thus supporting the claim that overall spectral shape is more important to perception than are the precise locations of spectral peaks"*.

Indeed, it is only recently that a few daring phoneticians have questioned the performance of the traditional methods based on formant frequencies against the more technology-oriented cepstral coefficients (Rose and Clermont, 2001; Rose, Osanai and Kinoshita, 2003). Not so surprisingly, these studies have also reported that the cepstrum outperformed the

Proceedings of the 10th Australian International Conference on Speech Science & Technology
Macquarie University, Sydney, December 8 to 10, 2004.

Accepted after full paper review

more traditional formant parameter in forensic speaker identification tasks.

Unfortunately, the cepstral coefficients which are extracted from the speech signal at full spectral range, i.e., the whole-spectrum approach, have been too often overvalued without consistent concerns for the relative effectiveness of different spectral regions. Therefore, a central objective of this study is to investigate the potency of cepstral ratios of inter- to intra-speaker variability of the data, using both cumulative and consecutive sub-band approaches. This provides the opportunity to compare the amount of speaker-specific information extant in sub-bands, both as a function of incremental inclusion of the upper spectral band and also in individual spectral bands. To this end, we have used the statistical F-ratio from ANOVA analysis as a direct method of observation to quantify separately the phonetic or speaker component of spectral variability (Wolf, 1972; Nolan, 1983).

In Section 2 we describe the speech material used for this study. In Section 3 the method for sub-band analysis is outlined. In Section 4, the whole-spectrum and sub-band approaches are compared from different perspectives. Finally, we conclude in Section 5.

2. Speech materials and speaker set

The speech materials for this study were taken from the "Speaker Database of the Japanese Research Institute for Police Science" -NRIPS, (Osanai, Tanimoto, Kido and Suzuki, 1995). The database comprises two non-contemporaneous recordings (separated by 3 to 4 months) of 297 adult-male native speakers of Japanese, aged between 20 and 60 years. All speakers were members of the Japanese police force and were recorded through the landline telephone circuit. Each recording comprises two repeats (this gives a total of 4 tokens per person) of the 5 Japanese vowels (pronounced in isolation), 37 words and 17 sentences. Recordings were sampled at the rate of 10 KHz and quantized to 12 bits.

2.1. Cepstral Parameterisation

For the purpose of our current research aimed at cepstrally-based Forensic Speaker Identification (FSI), we specifically sought the steady-state vocalic nuclei /i, e, a, o, u/ from utterances of sustained vowels pronounced in isolation. Sustained isolated vowels presumably are most steady and least co-articulated. Therefore, not only the external influences to vowel and speaker variations are minimised but also more speaker-specific information is preserved. By contrast with the discrete representation of the spectrum (based on formants), we decided to use the LP-cepstrum in order to have a more complete representation of the whole spectrum. Details of the analysis procedures are described in our recent paper (Khodai, Clermont and Barlow, 2004).

Non-contemporaneity of recording sessions with time interval of at least 3 months, together with the fact that recordings are from telephone conversations, contributes to the forensically realistic nature of our database. It is worth noting that, for estimates of intra-speaker variance to be forensically adequate, samples must have come from the non-contemporaneous recordings. This is because (Rose, 1999: 1): first, the forensic samples are inevitably non-contemporaneous (the criminal would be known if the data were from a single recording session). Second, a more complete representation of intra-speaker variability is manifested in non-contemporaneous recordings which otherwise will be underestimated using only contemporaneous recordings. From a statistical point of view, the results reported here can be considered particularly significant since our database has a large population of speakers.

3. Method

In order to uncover which frequency sub-band of the entire spectrum of a particular vowel provides the greatest amount of speaker-specific information, hence has a greater speaker discrimination power, the statistical F-ratio of inter- to intra-speaker variability was calculated for both cumulative and consecutive frequency sub-bands.

The inter-speaker variability, which is the variance of the speaker means weighted by the number of tokens per speaker, was calculated via:

$$\sigma^2_{\text{inter}} = \frac{\sum_{i=1}^{N} n_i (\overline{X}_i - \overline{\overline{X}})^2}{N-1} \qquad (1)$$

The intra-speaker variability, which is the mean of speakers' variances, was calculated via:

$$\sigma^2_{\text{intra}} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{n_i} (X_{ij} - \overline{X}_i)^2}{\sum_{i=1}^{N} n_i - N} \qquad (2)$$

where n_i is the number of tokens (4) per speaker, N is the total number of speakers (297), X_i is a vector of LPCC coefficients of order 14;

$\overline{X}_i = \frac{1}{n_i}\sum_{j=1}^{n_j} X_{ij}$ is the mean for jth speaker;

$\overline{\overline{X}} = \frac{1}{N}\sum_{i=1}^{N} \overline{X}_i$ is the grand mean.

Accepted after full paper review

The standard Euclidean distances in the above formulae were replaced with the parametric formulation of the cepstral distance (Clermont and Mokhtari, 1994). This was done in order to allow for parametric specification of any sub-band $[\theta_1, \theta_2]$ Hz within the Nyquist interval [0, Fs/2] Hz, thereby obviating the need to re-generate cepstral coefficients for every frequency sub-band.

In these experiments we have employed both cumulative and consecutive approaches to select the spectral sub-bands. In the cumulative approach, the lower spectral limit is fixed to a constant θ_1=0 Hz, and the upper limit θ_2 (starting at 300-Hz) is incremented up to the full range (Fs/2 Hz) by steps of 100-Hz. Therefore, the spectral region available is gradually incremented as the upper spectral limit is extended. In the consecutive approach, a window of length 300-Hz with a frame advance of length 10-Hz was selected to partition the entire available spectral region into M (471) number of sub-bands. By contrast with the cumulative approach, the spectral region available at each band is therefore limited to within the window's boundaries. The overlapping (rather than strictly contiguous) sub-bands are selected to preserve both the sense of spectral continuum and the smooth transition between the sub-bands. For both approaches, then, we investigated the inter- and intra-speaker variability across the spectral continuum.

4. Results and discussion

The amount of speaker-specific information conveyed by the cepstral coefficients across the vowels is examined in Section 4.1 as a function of incremental inclusion of the upper spectral band and also in individual spectral sub-bands. In order to emphasise the differences between the two approaches, the F-ratio profiles are superimposed on each other. Section 4.2 serves the purpose of completing the assessment with respect to previous studies. Section 4.3 focuses on the vowel-speaker variance obtained from the cepstrum in different spectral regions, wherein the contrasting behavior is found to confirm observations reported in recent literature. An assessment of our methodology for selection of the steady-states representatives (Khodai et al., 2004) is presented in Section 4.4.

4.1. Profiles of F-ratios

The graphs in Figure 1 provide a vowel-by-vowel perspective of the cepstrally-based F-ratio, where both inter- and intra-speaker variations were captured across the spectral continuum by using both cumulative and consecutive sub-bands. For the sake of comparison, the expected regions for formant frequencies are also marked. The expected formant frequency regions were carefully estimated using LPC poles and plots of Log Magnitude Spectra.

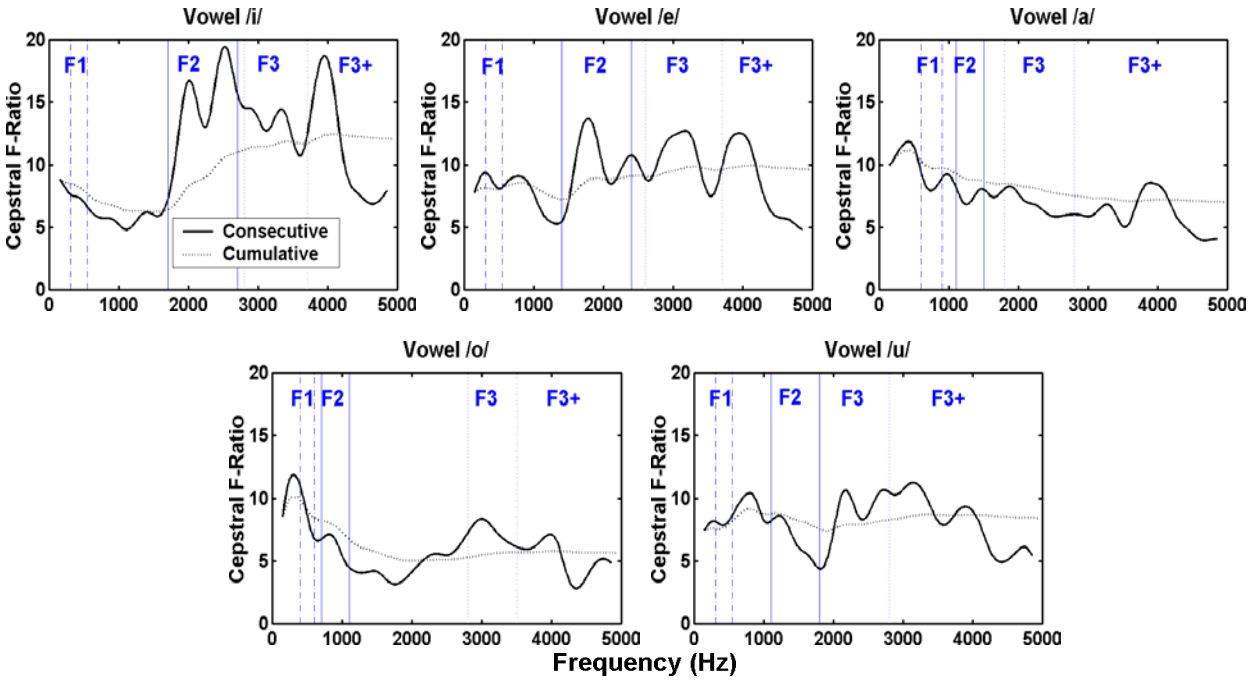

Figure 1: Profile of cepstral F-ratios of Min-variance based frames across the spectral continuum (1) as a function of incremental upper sub-band (dotted curve), and (2) with a window of length 300-Hz and a frame advance of length 10-Hz (solid curve). Dataset: 4 repetitions of 5 sustained Japanese vowels spoken by 297adult, male speakers of Japanese.

The dotted curves illustrate the behavior of the cepstrally-based F-ratio profile as a function of the upper spectral limit θ_2 (i.e., cumulative approach). It appears that, as the upper spectral limit is extended to the higher regions and more information becomes available, the F-ratio curve tends to exhibit a behavior consistent with a spectral averaging process; hence its smooth nature and less apparent sensitivity to individual spectral regions.

By contrast with the gradual increment of the upper limit approach, a more enlightening perspective on the behavior of inter- to intra-speaker variability is gained by employing a consecutive sub-band approach, where the only information available at a time is contained within the limits of the specified frame boundaries. The close proximity of the behavior of the second curve (solid line) to what is already known based on the traditional notion of formant frequencies, confirms the effectiveness and the importance of differing spectral regions when they are exploited in cepstrally-based methods.

Indeed, one can easily distinguish spectral regions where the greater amount of inter- to intra-speaker variability is present. In particular, spectral regions 1700-2700 Hz (spanning F2), and 3700-4500 Hz (spanning $F3^+$) of /i/, spectral regions 1400-2400 Hz (spanning F2), 2600-3700 Hz (spanning F3) and 3800-4500 (spanning $F3^+$) of /e/, and mainly the higher spectral regions (F3 and beyond) in case of back vowels /a, o, u/ are powerful candidates for speaker discrimination. For the back vowels /a/ and /o/ there is a peak at low frequency regions (below F1), which may be due to glottal-source rather than supralaryngeal characteristics. Nevertheless, the high F-ratio values at these peaks indicate their potency for speaker discrimination based on the data at hand. It also appears that /i/ and /e/ are the vowels with higher F-ratios than the back vowels, hence greater discrimination power.

4.2. Comparisons with other studies

In her PhD study of spoken Japanese vowels (based on formant frequencies), Kinoshita (2001) found that F2 of /i/ and F2 and F3 of /e/ are the most promising candidates for speaker identification parameters, which confirms our findings for the same vowel set. The reason for the greater significance of higher spectral regions ($F3^+$) in our study might be due to the fact that Kinoshita's measurements were limited to 4 kHz.

This study also agrees with findings by other researchers on different languages, (e.g., Sambur, 1975-American English; Mokhtari and Clermont, 1996-Australian English) in that F2 of front vowels is the most speaker discriminating parameter. These authors also list F3 of back vowels as another promising parameter in speaker discrimination.

4.3. Speaker-Vowel variance

The contribution of the lower and higher spectral regions (in terms of formant frequencies) to phonetic-quality and speaker-specificity has been the subject of many research studies for many decades. The early findings in this area appear to date back to Lewis' study of resonance frequencies of sung vowels in 1936. Lewis put forward the conjecture that the quality of a vowel may be determined by the two lowest resonance frequencies, and differences in individual voices may be related to other resonances.

Recent studies based on the profiles of cepstral variances, (Kitamura and Akagi, 1994; Hayakawa and Itakura, 1994; Mokhtari and Clermont, 1994), have also confirmed findings of the *traditional approach* based on the formant frequencies of the vocal tract. That is, higher spectral regions contain much more speaker-discriminating information, while vowel characteristics are concentrated mainly in the lower spectral regions.

Figure 2 illustrates well the contrast between speaker and vowel variance profiles, as they unfold band by band through the entire spectral range. It is quite evident that the largest amount of vowel variance is concentrated in the lowest spectral region, which spans the F1 and low F2 ranges. By contrast, speaker variance reaches its maximum at high spectral regions starting from high F2 up to low F4. This behavior is also in agreement with findings reported in recent literature.

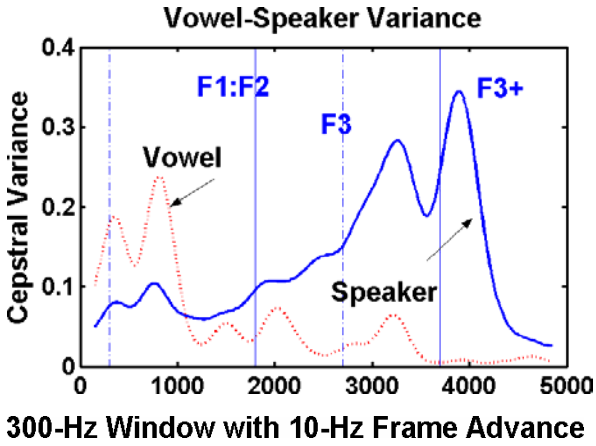

Figure 2: Profile of speaker and vowel variance of Min-variance based frames, based on consecutive sub-band approach with a window of length 300-Hz and a frame advance of length 10-Hz. Dataset: 4 repetitions of 5 sustained Japanese vowels spoken by 297 adult, male speakers of Japanese.

Accepted after full paper review

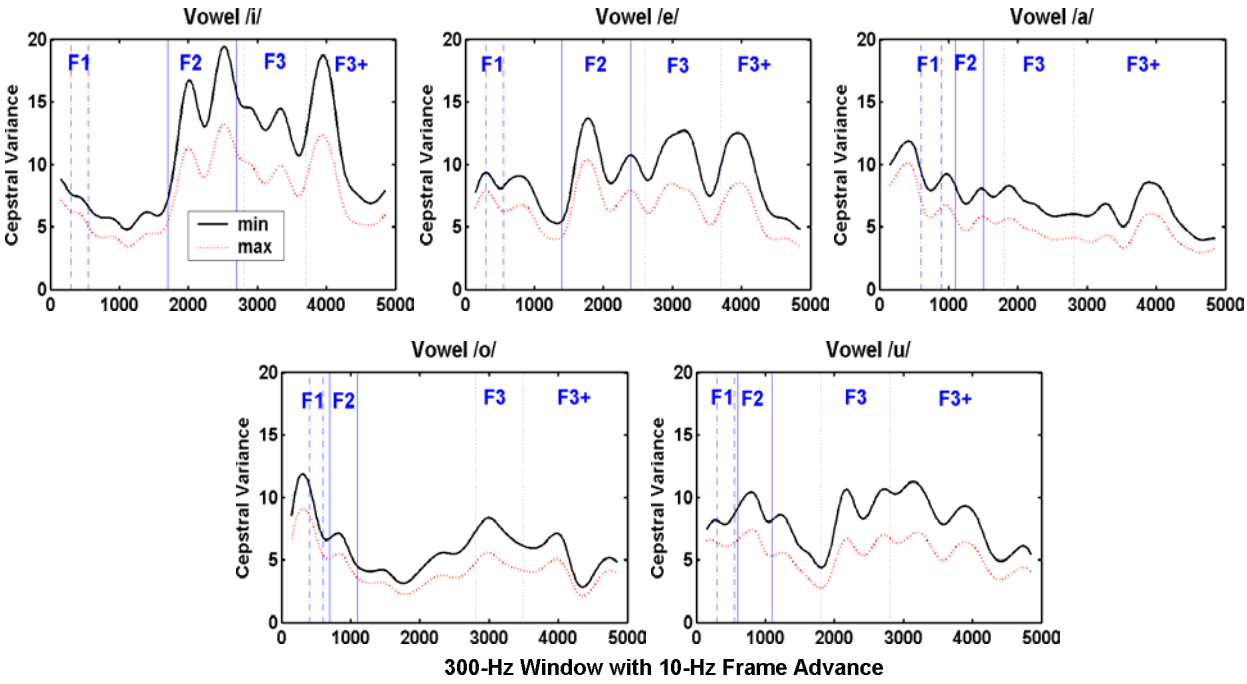

Figure 3: Profile of cepstral F-ratios of Min- and Max-variance based frames with a window of length 300-Hz and a frame advance of length 10-Hz. Dataset: 4 repetitions of 5 sustained Japanese vowels spoken by 297adult, male speakers of Japanese.

4.4. Assessment of Min- and Max-variance based frames

For a parameter to be considered efficient and robust, it needs to concomitantly demonstrate a high degree of inter-speaker variability, while being consistent and less sensitive to intra-speaker variations. Indeed, one of the most crucial problems in FSI or ASR systems is the long-term intra-speaker variability of a parameter and its effect on system's accuracy, which is known to have adverse effects on accuracy (Nolan, 1983: 12; Furui, 1986: 194). These authors also reported that differences in recording occasions could cause larger intra-speaker variations. Therefore, the efficiency of a parameter in the long-term may or may not be the same as that in short-term.

As an attempt to minimize the effect of long-term intra-speaker variability in our analysis procedure, (Khodai et al. 2004), we selected 4 frames for each token such that they are the closest set to represent the whole duration of the signal for that particular token. Using 4 tokens and 4 frames per token, we constructed a matrix of all possible combinations amongst these candidates, such that no two candidates in a single combination belong to the same token. This gave us a total of 256 possible combinations. We then selected the combination with the least spectral variance amongst its members, that is, with minimum long-term intra-speaker variability. To satisfy our curiosity and to facilitate later comparisons, we also saved the frames with maximum intra-speaker spectral variability.

The F-ratio results for both min- and max-variance based frames are illustrated in the graphs of Figure 3. One can observe that, the F-ratio profile is always larger for min-variance based frames than that of max-variance based frames. Careful inspection of the individual inter- and intra-speaker variability (the results of which are not shown here), indicates that we achieved minimisation of long-term intra-speaker variability as well as maximisation of long-term inter-speaker variability, both of which are most desirable in any FSI or ASR systems.

5. Conclusion and future directions

Using a cumulative- and a sub-band approach, a cepstrally-based database of 5 Japanese vowels, uttered in isolation by 297 adult, male speakers of Japanese was examined across the spectral continuum, on an intra- and an inter-speaker basis.

In regard to the relative importance of individual spectral regions, the consecutive sub-band approach has yielded cepstral F-ratio profiles, which are in agreement with results already established from the traditional use of formant frequencies. By contrast, the whole-spectrum approach based on cumulative sub-bands tends to blend spectral information in a manner consistent with a spectral averaging process; hence its smooth nature and less apparent sensitivity to individual spectral regions.

Accepted after full paper review

In the light of our findings, it would seem highly desirable for researchers to confront the whole-spectrum against the sub-band approach when utilising cepstrally-based methods. The results reported here clearly show that certain sub-bands are likely to yield greater speaker discriminating power than the whole spectrum. It would be also important for forensic phoneticians 1) to consider the conjunctional use of cepstrally-based methods, where a much more complete and detailed description of the whole spectral shape is provided, and 2) to reconsider the issue of whole-spectrum versus formants from the point of view of interpretability. Perhaps it is the performance that should be of greater importance to forensic science, rather than the interpretability to the jury (Rose et al. 2001: 34).

The above recommendations do not however imply that the formants are not useful parameters. The advantage of the formants is that a large amount of information is contained in few features (Zahorian, 1993: 1978-9). Unfortunately, the formants do suffer the serious limitation of determinacy, that is, the persistent difficulty in achieving robust determination of their location for many conditions. Therefore, forensic scientists should perhaps take advantage of this rather fortunate coincidence that computationally inexpensive, more robust, and easily extractable parameters such as the LP-cepstrum, do give access to inter- and intra-speaker variability in greater depths and more details across the entire spectral region.

Further research in forensic science is clearly desirable in order to investigate the full potential of cepstrally-based methods, which acknowledge the importance of individual spectral sub-bands.

6. Acknowledgements

The authors extend grateful thanks to Takashi Osanai for his permission to use the recorded speech from NRIPS database.

7. References

Clermont, F. & Mokhtari, P. (1994). *Frequency-band specification in cepstral distance computation.* Proc. of the 5th Australian Int. Conf. on Speech, Science, and Technology, 1: 354-359.

Furui, S. (1986). *Research on individuality features in speech waves and automatic speaker recognition techniques.* Speech Communications 5: 183-197.

Hayakawa, S. & Itakura, F. (1994). *Text-dependent speaker recognition using the information in the higher frequency bands.* Proc. of the Int. Conf. on Acoustics, Speech and Signal Processing I: 137-140.

Khodai-Joopari, M., Clermont, F. & Barlow, M. (2004). *A forensically motivated tool for selecting cepstrally-consistent steady-states from non-contemporaneous vowel utterances.* Proc. of the 8th Int. Conf. on Spoken Language Processing, Jeju Island, Korea, 1047-1050.

Kinoshita, Y. (2001). *Testing Realistic Forensic Speaker Identification in Japanese: A Likelihood Ratio Based Approach Using Formants.* Unpublished PhD thesis, The Australian National University.

Kitamura, T. & Akagi, M. (1994). *Speaker individualities in speech spectral envelopes.* Proc. of the 3rd Int. Conf. on Spoken Language Processing, 1183-1186.

Lewis, D. (1936). *Vocal Resonance.* The Journal of the Acoustical Society of America 8: 91-99.

Mokhtari, P. & Clermont, F. (1994). *Contributions of selected spectral regions to vowel classification accuracy.* Proc. of the 3rd Int. Conf. on Spoken Language Processing, Yokohama, Japan, 1923-1926.

Mokhtari, P. & Clermont, F. (1996). *A methodology for investigating vowel-speaker interactions in acoustic-phonetic domain.* Proc. of the 6th Australian Int. Conf. on Speech Science & Technology, 127-132.

Nolan, F. (1983). *The phonetic Bases of Speaker Recognition.* Cambridge University Press, Cambridge.

Osanai, T. & Tanimoto, M. & Kido, H. & Suzuki, T. (1995). *Text-dependent speaker verification using isolated word utterances based on dynamic programming.* [In Japanese], National Research Institute for Police Science Report, 48(1): 15-19.

Rose, P. (1999). *Long- and short-term within-speaker differences in the formants of hello.* Journal of International Phonetic Association 29: 1-31.

Rose, P. & Clermont, F. (2001). *A comparison of two acoustic methods for forensic speaker discrimination.* Acoustics Australia 29: 1-31.

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.* Forensic Linguistics 10 (2): 179-202.

Sambur, R. (1975). *Selection of Acoustic Features for Speaker Identification.* IEEE Transactions on Acoustic, Speech and Signal Processing, ASSP-23 (2), 176-182.

Wolf, J.J. (1972). *Efficient acoustic parameters for speaker recognition.* The Journal of the Acoustical Society of America 51: 2044-2056.

Zahorian, S. A. & Jalali Jagharghi A. (1993). *Spectral-shape features versus formants as acoustic correlates for vowels.* JASA: 94 (4) 1966-82.

Accepted after full paper review