Speaker Identification using FM Features

Tharmarajah Thiruvaran[1], Eliathamby Ambikairajah[1], Julien Epps[2]

[1]School of Electrical Engineering and Telecommunications, Faculty of Engineering,
The University of New South Wales, Sydney, Australia
[2]ATP Research Laboratory, National ICT Australia,
Sydney, Australia
thiruvaran@student.unsw.edu.au ambi@ee.unsw.edu.au julien.epps@nicta.com.au

Abstract

The AM-FM modulation model of speech is a nonlinear model that has been successfully used in several branches of speech-related research. However, the significance of the AM-FM features extracted from this model has not been fully explored in applications such as speaker identification systems. This paper shows that frequency modulation (FM) features can improve speaker identification accuracy. Due to the similarity between amplitude modulation (AM) feature and the conventional Mel frequency cepstrum coefficients (MFCC), this paper mainly focuses on the FM feature. The correlation between FM feature components is shown to be very small compared with that of Mel filterbank log energies, thus reducing the need for decorrelation. FM feature components are shown to be very nearly Gaussian distributed. Further, speech synthesis using AM-FM features is performed to compare four existing AM-FM demodulation methods based on the perceptual quality of the synthesized speech. Of these, Digital Energy Separation Algorithm (DESA) gives the best synthesized speech, and is thus used as a front-end in our speaker identification system. Evaluation of speaker identification using FM features on the NIST 2001 database shows a relative improvement in speaker identification accuracy of 2% for male speakers and 9% for female speakers over the conventional MFCC-based front-end.

1. Introduction

Speech analysis using the conventional source-filter model alone appears to provide only a certain level of performance in speech processing applications. One alternative approach is the modeling of speech resonances as AM-FM signals. In the formant AM-FM model (Maragos, Kaiser & Quatieri, 1993), each formant is represented as an AM-FM signal and the total speech is the sum of all the formants, as in equation (1).

$$s(t) = \sum_{k=1}^{K} a_k(t)\cos(\omega_{ck}t + \phi_k(t)) \quad (1)$$

$$\phi_k(t) = \omega_{dk}\int_0^t q_k(\tau)d\tau \quad (2)$$

where, $a_k(t)$ is the AM component, $\cos(\omega_{ck}t + \phi_k(t))$ is the FM modulated component, $\phi_k(t)$ is the FM component, ω_{dk} is the modulation index, K is the total number of formants and f_{ck} is the center frequency of the k^{th} formant. Several examples of evidence for the existence of modulation in speech are summarized by Maragos et al., (1993). One such cause of the modulation is the cavities in the vocal tract. In other words, the modulation characteristics depend on the physical properties of the speaker. These properties can be captured in AM-FM features, which are the demodulated AM and FM components of the speech.

In speaker identification systems, the front-end must extract features that characterize the speaker. Based on the argument relating the AM-FM features to the speaker's physical properties, we can hypothesize that the speaker identification system can be improved using these AM-FM features. In work by Jankowski, Quatieri and Reynolds (1994), AM-FM features were applied to speaker identification, but the authors found that FM features gave poor performance, and focused mainly on "Teager Energy" (Maragos et al., 1993) as a feature. They used Digital Energy Separation Algorithm (DESA) with formant tracking.

The most common method of using AM-FM features in speech processing is estimating speech formants and then extracting the AM-FM components from them. Formant estimation for AM-FM extraction can be achieved using Linear Predictive Coding (LPC) analysis (Jankowski et al., 1994) or using an iterative approach as suggested in Maragos et al., (1993). The iterative approach involves repeatedly changing the center frequency of the band pass filter to the average of the extracted instantaneous frequency. All these formant estimation techniques introduce additional complexity to the AM-FM extraction methods. A method to avoid this formant estimation in AM-FM extraction is to use the band that gives maximum normalized energy (Bovik, Maragos &

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after full paper review

Quatieri, 1993). We show that using fixed band-pass filters and extracting AM-FM in all bands is an alternative method for front-end applications.

Several methods of demodulation of AM-FM signals are available in the literature. DESA (Maragos et al., 1993) is a popular method. A good comparison between the demodulation method based on Hilbert transform and DESA is given by Potamianos and Maragos (1994). Some other demodulation methods are comprehensive modulation spectra (CMS) method (Wang, Greenberg, Swaminathan, Kumaresan & Poeppel, 2005) and methods based on: (i) differentiation (Ziemer & Tranter, 2002); (ii) quadrature filtering (Nie, Stickney & Zeng, 2005); (iii) statistics (Wan-Chieh & Doerschuk, 2000); (iv) curve fitting (Sekhar & Sreenivas, 2004); (v) short time Fourier transform (Nelson, 2001); (vi) auditory filters (Quatieri, Hanna & O'Leary, 1997).

In this paper, we show that FM features extracted without formant tracking can improve MFCC-based speaker identification performance. Due to the similarity between the AM feature and the Mel filterbank energies of the conventional MFCC, we focus on investigating the significance of FM features in speaker identification and the properties of the FM feature.

Four of the abovementioned demodulation methods were implemented, and the performance was compared based on the perceptual quality of the synthesized speech using the AM-FM features. The four methods implemented herein were:

1. Digital Energy Separation Algorithm (DESA)
2. Demodulation method based on quadrature filtering (also known as 'FAME')
3. Demodulation based on differentiation
4. Comprehensive modulation spectra (CMS)

In section 2 we provide the details of the speech synthesis, the comparison of the above four demodulation methods and the significance of the constant phase, in section 3 we provide the details of FM feature extraction and the properties of the feature and in section 4 we provide the details of speaker identification experiment, before concluding in section 5.

2. Speech synthesis using AM-FM features to compare demodulation methods

2.1. Speech synthesis

Speech synthesis is a perceptual means of testing features extracted from speech that gives a good subjective impression of how well a feature set can represent the speech signal, and are used here to test the quality of the feature extraction methods. Speech synthesis was performed by Potamianos and Maragos (1997) with AM and FM features extracted using DESA.

Let $q_k(n)$ be the extracted FM component and $a_k(n)$ be the AM component of the k^{th} filter. Then the synthesized speech signal can be expressed as in equation (3),

$$s(n) = \sum_{k=1}^{K} a_k(n) \cos\left(\frac{2\pi f_{ck} n}{f_s} + \frac{2\pi}{f_s} \sum_{r=1}^{n} q_k(r)\right) \tag{3}$$

where f_{ck} is the center frequency of the k^{th} filter, f_s is the sampling frequency and K is the total number of filters in the filter bank. For speech synthesis, a uniform Gabor filter bank with center frequency spacing from 50Hz to 8000Hz with K=160 was used, similarly to Potamianos et al., (1997).

2.2. Comparison of the demodulation methods

In this section, the synthesized speech is used to compare the four AM-FM demodulation methods listed in section 1. All methods initially pass the speech signal through a band pass filterbank and then use different techniques to extract FM from the band passed signal. The first method, DESA uses a non-linear energy operator, the Teager Energy to extract FM. The Teager Energy (Ψ) is defined as in equation (4).

$$\Psi(s(n)) = s^2(n) - s(n-1)s(n+1) \tag{4}$$

where $s(n)$ is the output of the band pass filter. Then the FM is extracted as in equation (5).

$$FM \approx \arccos\left(1 - \frac{\Psi[y(n)] + \Psi[y(n+1)]}{4\Psi[s(n)]}\right) \tag{5}$$

$$y(n) = s(n) - s(n-1) \tag{6}$$

The second method, FAME extracts the in-phase component (a) and the out-of-phase component (b) of the signal by modulating with sines and cosines of the center frequency of the band. Then FM is calculated using those components as in equation (7).

$$FM = \frac{b(da/dt) - a(db/dt)}{2\pi(a^2 + b^2)} \tag{7}$$

The third method differentiates the signal and passes the results through an envelope detector to extract FM. If the band pass signal is defined as in equation (8) then the differentiator output is given by equation (9), with the assumption that the amplitude is slowly varying.

$$s(n) = a(n)\cos(\Omega_c n + \phi(n)) \tag{8}$$

$$\dot{s}(n) \approx -a(n)\sin(\Omega_c n + \phi(n))(\Omega_c + \dot{\phi}(n)) \tag{9}$$

Where $a(n)$ is the amplitude modulated signal, Ω_c is the center frequency of the band pass filter and $\dot{\phi}(n)$ is the FM. Then the FM component can be obtain by taking the envelope of $\dot{s}(n)$, however it is affected by $a(n)$.

The final method, CMS, extracts the angle from the analytical version of the signal as in equation (10) and then differentiates the angle to get FM.

$$x(n) = s(n) + j\hat{s}(n) \tag{10}$$

The first two methods (DESA and FAME) gave very good perceptual quality synthesized speech signals in informal tests. The advantage of the DESA method is the easy implementation, while FAME requires less computational complexity and executes quickly. The perceptual quality of the speech synthesized using the differentiation method and CMS did not reach a reasonable quality. The reason for this and a remedy are explained in the next subsection. The synthesized speech for the first two methods are given in figure 1.

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after full paper review

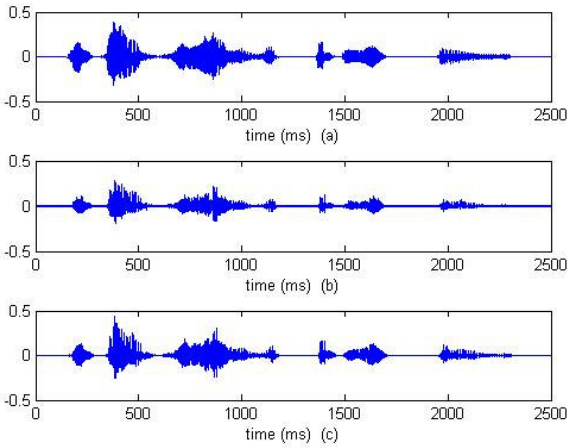

Figure 1: Synthesized speech from AM-FM features (a) Original speech (b) DESA method (c) FAME method

2.3. Importance of constant phase

The complete AM-FM model for speech signal is,

$$s(n) = \sum_{k=1}^{K} a_k(n) \cos\left(\frac{2\pi f_{ck} n}{f_s} + \frac{2\pi}{f_s} \sum_{r=1}^{n} q_k(r) + \theta \right) \qquad (11)$$

where θ is the constant phase angle, and the other parameters are the same as in equation (3). Usually, the constant phase in the AM-FM model is not captured in the demodulation. We identified the significance and the usefulness of the constant phase in the model. We propose to include the knowledge of change of sign or zero crossing in the speech by a constant phase of π in the FM part of the model. Although the model explicitly contains a constant phase angle as in equation (1), in several demodulation algorithms this has not been emphasized for the following reasons.

- AM is extracted as the absolute envelope of the speech
- In FM extraction, the phase of the FM component is differentiated. In the differentiation the constant phase vanishes.

We introduced a constant phase π for each zero-crossing in the speech synthesis, and observed a significant improvement in the perceptual quality of the synthesized speech. The synthesized speech with the constant phase is shown in Fig. 2.

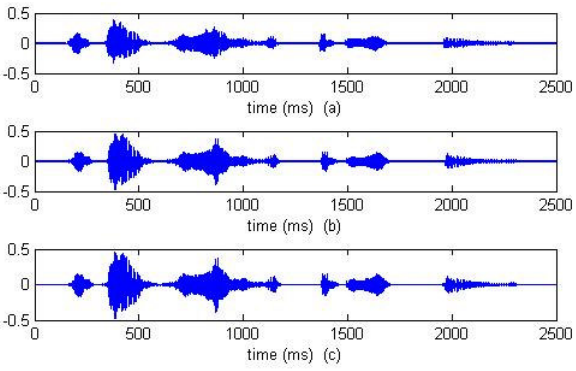

Figure 2: Synthesized speech from AM-FM features with constant phase. (a) Original speech (b) Differentiation method (c) CMS method

3. FM feature extraction

3.1. Overview

Based on the preliminary experiments in section 2, we used the DESA method to extract FM features for our speaker identification system. Initially, the speech signal is passed through a bank of band pass filters. In these experiments, instead of identifying formants, we extracted FM feature from each band of the speech signal as in Fig. 3.

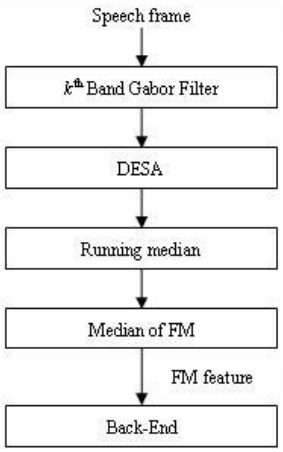

Figure 3: FM feature extraction

3.2. Filter bank

For AM-FM extraction a Gabor filter bank is preferred because of the optimum time and frequency sensitivity and the absence of large side lobes (Maragos et al., 1993). The impulse response of the Gabor filter is given in equation (12) and the transfer function is given in equation (13), where ω_c is the center frequency.

$$h(t) = \exp(-\alpha^2 t^2)\cos(\omega_c t) \qquad (12)$$

$$H(\omega) = \frac{\sqrt{\pi}}{2\alpha}\exp\left[-\frac{(\omega-\omega_c)^2}{4\alpha^2}\right] + \exp\left[-\frac{(\omega+\omega_c)^2}{4\alpha^2}\right] \qquad (13)$$

Here, α is used to control the bandwidth of the filter, which is $\frac{\alpha}{\sqrt{2\pi}}$. The impulse response in discrete time domain is

$$h(n) = \exp(-b^2 n^2)\cos(\Omega_c n), \qquad (14)$$

where $\Omega_c = 2\pi f_c T$, $b = \alpha T$, $-N \le n \le N$ and T is the sampling period. N is selected to satisfy the constraint of $\exp(-b^2 N^2) = 10^{-5}$, to truncate the response. In our system, the center frequencies are spaced according to the critical band specification.

3.3. FM features for front-end processing

In order to use FM as a feature for the speaker identification front end, the FM feature in each band should be represented in a compact form. Taking the maximum in each band is intuitively meaningful, as this value will represent the maximum bandwidth or deviation from the center frequency in

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after full paper review

each band. However, the FM extracted from DESA contains some spurious peaks and dips, as in Fig. 4, so taking a maximum is problematic in this case. We also ruled out the mean value as a feature, since it is also affected by spurious peaks and dips. As a compromise, we choose the median in each band to represent FM in a compact form. The block diagram of FM feature extraction for the speaker identification front-end is given in Fig. 3.

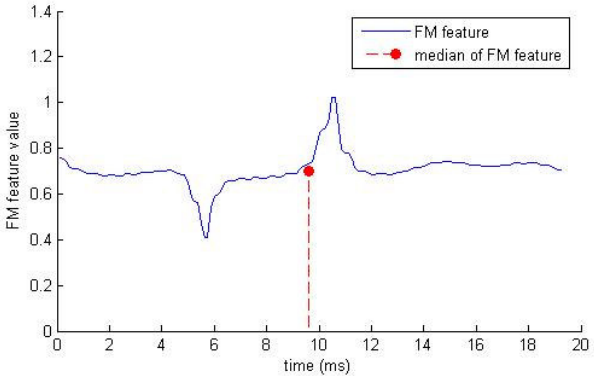

Figure 4: FM feature with spurious peaks and dips superimposed with the median value.

3.4. Correlation of FM features

The correlation of a feature set gives a very good indication of how compactly it characterizes the speech information. Higher correlation implies that the feature set has more redundancy and can be compressed by decorrelating the feature set. We observed the correlation of FM features extracted according to the process given in Fig. 3. We took the covariance of the FM feature component in each band with other bands, and obtained the correlation as below. Let X_i and X_j be the FM feature values in bands i and j. Considering X_i and X_j as random variables, the covariance and correlation of them are given in equations (15) and (16) respectively.

$$\mathrm{cov}(X_i X_j) = E\{(X_i - E\{X_i\})(Y_i - E\{Y_i\})\} \quad (15)$$

$$\mathrm{cor}(X_i X_j) = \frac{\mathrm{cov}(X_i X_j)}{\sigma_i \sigma_j} \quad (16)$$

where $E\{.\}$ denotes expectation and σ denotes standard deviation. The feature values of all bands were concatenated into matrix form and covariance and correlation matrix were calculated. In the correlation matrix the off-diagonal terms refer to the correlation between FM features of different bands. All these off-diagonal terms are used to compare with the corresponding off-diagonal terms of the correlation matrix of MFCC and Mel filterbank log energies to judge the relative amount of correlation. In MFCCs, the discrete cosine transform (DCT) is used for decorrelation. For this analysis the training data of NIST 2001 Evaluation database was used. Confidence intervals for the correlation values were calculated for each speaker and averaged over all speakers. These confidence intervals are given in Tables 1 and 2 for male speakers and female speakers respectively.

Table 1: 95% confidence interval averaged over all female speech data in the training database

	95 % confidence interval	
FM without DCT	0.093	0.094
Mel filterbank log energies	0.800	0.801
MFCC	0.092	0.093

Table 2: 95% confidence interval averaged over all male speech data in the training database

	95 % confidence interval	
FM without DCT	0.083	0.084
Mel filterbank log energies	0.749	0.751
MFCC	0.090	0.091

The analysis shows that the correlation of FM features is considerably less than that of Mel filterbank log energies, thus reducing the need to decorrelate the FM feature. One reason for this observation is the compactness of Gabor filters that reduces the overlap with adjacent bands (particularly relative to the highly overlapping triangular filter bank conventionally used in MFCC extraction).

3.5. Distribution of FM features

We found that the distribution of the FM feature extracted using the algorithm given in Fig. 3 can be well approximated by a Normal distribution. The normplot of a sample speech signal from the NIST 2001 database for different bands is given in Fig. 5. The normplot provides a linear plot if the data follows a Normal distribution, and deviates from this line otherwise. The normplot is superimposed on a line plot joining the first and third quartiles of the data in Fig. 5. The normplot of the FM feature centered at 50Hz has negligible deviation, demonstrating how well it is approximated by a Normal distribution. For the other bands only a small portion (around 1%) of the data deviates from linearity on the normplot.

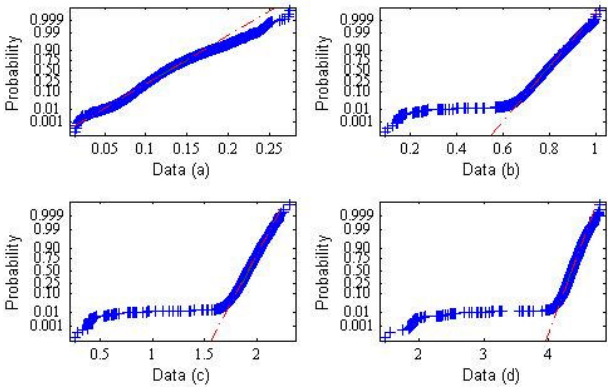

Figure 5: Norm plot of FM feature data centered at (a) 50Hz (b) 570Hz (c) 1370Hz (d) 3400Hz, superimposed on a dashed line joining the first and third quartiles.

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after full paper review

4. Speaker identification

4.1. Database

Speaker identification experiments were conducted using the NIST 2001 Evaluation database. This is a cellular database, where all speech segments are recorded over different types of cellular phones. The environment, channel and handset of the test segments are different from that of the training segments. The training data files are about 2 minutes long and the testing data files are varying length below 60 seconds. The total training data is 145 hours for male speakers and 197 hours for female speakers. The total testing data is 432 hours for male speakers and 648 hours for female speakers.

4.2. Speaker identification system

Speech signal is passed through a Gabor filter bank and FM is extracted at each bands. These FM features were used to train Gaussian mixture models (GMMs). A detailed explanation of the use of GMMs in speaker identification has been given in (Reynolds & Rose, 1995). Our system was trained using 74 male speakers and 100 female speakers, and tested with 300 male utterances and 300 female utterances from a total of 850 male utterances and 1188 female utterances. The test database is designed for speaker verification, such that each test segment is to be checked with 11 target speaker models. For speaker identification experiments, each test segment is checked against all speaker models in the database. In our experiment for male test segments the test segment was checked against 74 male speaker models. Likewise, for female test segments the segment was checked against 100 female speaker models. Out of 17 FM feature terms, only the first 12 terms were used in our experiment, in order to compare the performance with 12 coefficients of MFCCs.

4.3. Results

The performance of FM is considerably improved for female speakers and slightly degraded for male speakers compared with the performance of MFCCs. When combining MFCCs with FM components, the performance increased over the individual performance. When averaging the performance of male and female speakers, the overall absolute improvement in accuracy is 5%. The results are given in Table 3.

Table 3. Speaker identification accuracy (%)

	Male	Female
12 FM	39.3	36.3
12 MFCC	41.3	28.3
12FM+12MFCC	43	37.3

5. Conclusions

FM features have been successfully used to augment MFCC features in a speaker identification system, with improved performance. The distributions of FM features have been shown to closely approximate a Gaussian distribution, a highly desirable property of features used in conjunction with a GMM back-end. The correlation of FM with adjacent bands was found to be very small compared with that of Mel filterbank log energies, which reduces the need to take DCT for decorrelation. Speech synthesis was used to compare the qualities of different AM-FM demodulation methods, leading to the selection of the DESA method of FM feature extraction. Finally, when FM features were concatenated with MFCCs, an overall 5% relative improvement in accuracy over the conventional MFCC-based front-end was obtained.

6. References

Bovik, A. C., Maragos, P., & Quatieri, T. F. (1993). AM-FM energy detection and separation in noise using multiband energy operators. *IEEE Transactions on Signal Processing,* 41(12): 3245-3265.

Jankowski, C. R., Quatieri, T. F. & Reynolds, D. A. (1994). Formant AM-FM for speaker identification. *Proceedings of IEEE International Symposium on Time-Frequency and Time-Scale Analysis,* pp. 608-611.

Maragos, P., Kaiser, J. F. & Quatieri, T. F. (1993). Energy separation in signal modulations with application to speech analysis. *IEEE Transactions on Signal Processing,* 41(10): 3024-3051.

Nelson, D. J. (2001). Cross-spectral methods for processing speech. *Journal of the Acoustical Society of America* 110(5): 2575-92.

Nie, K., Stickney, G. & Zeng, F. G. (2005). Encoding frequency modulation to improve cochlear implant performance in noise. *IEEE Transactions on Biomedical Engineering,* 52(1): 64-73.

Potamianos, A. & Maragos, P. (1994). A comparison of the energy operator and the Hilbert transform approach to signal and speech demodulation. *Signal Processing* 37(1): 95-120.

Potamianos, A. & Maragos, P. (1997). Speech analysis and synthesis using an AM-FM modulation model. *Proceedings of EUROSPEECH-1997,* pp. 1355-1358.

Quatieri, T. F., Hanna, T. E. & O'Leary, G. C. (1997). AM-FM separation using auditory-motivated filters. *IEEE Transactions on Speech and Audio Processing,* 5(5): 465-480.

Reynolds, D. A. & Rose, R. C. (1995). Robust text-independent speaker identification using Gaussian mixture speaker models. *IEEE Transactions on Speech and Audio Processing,* 3(1): 72-83.

Sekhar, S. C. & Sreenivas, T. V. (2004). Novel approach to AM-FM decomposition with applications to speech and music analysis. *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 753-756.

Wan-Chieh, P. & Doerschuk, P. C. (2000). Statistical AM-FM models, extended Kalman filter demodulation, Cramer-Rao bounds, and speech analysis. *IEEE Transactions on Signal Processing,* 48(8): 2300-2313.

Wang, Y., Greenberg, S., Swaminathan, J., Kumaresan, R. & Poeppel, D. (2005). Comprehensive modulation representation for automatic speech recognition. *Proceedings of INTERSPEECH-2005,* pp. 3025-3028.

Ziemer, R. E. & Tranter, W. H. (2002). *Principles of communications* (5th ed.), John Wiley & Sons, Inc. U.S.A., chap. 3, pp. 142-147.

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after full paper review