Interspeech 2008 Special Session: Forensic Speaker Recognition - Traditional and Automatic Approaches.
Brisbane, Queensland, Australia. September 2008

FM Features for Automatic Forensic Speaker Recognition

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

[1]School of Electrical Engineering and Telecommunications,
The University of New South Wales, Sydney NSW 2052 Australia.
[2]National Information Communication Technology (NICTA),
Australian Technology Park, Eveleigh 1430, Australia.

thiruvaran@student.unsw.edu.au, ambi@ee.unsw.edu.au, j.epps@unsw.edu.au

Abstract

Frequency modulation (FM) information from the speech signal is herein proposed to complement the conventional amplitude based features for automatic forensic speaker recognition systems. In addition to presenting the AM-FM model of speech used to generate the proposed frequency modulation features, the significance of frequency modulation for speaker recognition is discussed. Evaluation results from an automatic forensic speaker recognition system combining FM and MFCC features are shown to out-perform those of a system employing MFCC features alone, in terms of all typical metrics, such as detection error trade-off curves, Tippett curves and applied probability of error curves.

Index Terms: frequency modulation, automatic forensic speaker recognition.

1. Introduction

Expert opinion may be required about a recording from a crime scene relating to the suspect in judicial proceedings. In that situation automatic forensic speaker recognition (FSR) system can be used by a forensic scientist to produce a meaningful estimate of the strength of the evidence (information extracted from the questioned recording) in the form of Likelihood Ratio (LR). To accomplish this task, the front-end of the automatic FSR system should utilize all possible information from the speech signal. Studies of human auditory perception suggest that frequency modulation information from the speech signal is complementary to the conventional amplitude based features such as MFCC.

Psychophysical evidence for the existence of human auditory pathways tuned to frequency modulated tones was first reported in [1]. Further experiments, supporting the claim that information pertaining to changes in amplitude and information pertaining to changes in frequency are processed in separate psychophysical channels, are reported in [2]. While these findings reveal that frequency modulation plays a significant part in human perception, several sources of evidence for the existence of frequency modulation in the speech signal are reported in [3], based on vocal tract air velocity measurements conducted in [4]. Further, a model for the speech signal, incorporating this frequency modulation information in terms of an AM-FM model, was also proposed in [3] as an AM-FM model. Later, several human perception experiments showed that an FM signal, combined with amplitude information, enhanced human perception [5, 6]. Further, the FM properties of the speech signal have been successfully exploited in cochlear implant applications [7], automatic speech recognition [8] and automatic speaker recognition [9], in each case as a complement to amplitude information. Thus, there is a significant and diverse body of research to support the notion that FM components in the speech signal provide complementary information to amplitude information in features such as MFCCs.

The primary motivation for using FM features for speaker recognition is the explanation for the modulation observed in the speech signal in [3], based on Teager's experiment. In particular, "the air jets flowing through the vocal tract during speech production is highly unstable and oscillates between its walls, attaching or detaching itself, and thereby changing the affected cross-sectional areas and air masses, which affects the frequency of the cavity resonator" [3]. It is the vocal tract walls that cause the modulation of the oscillating air flow, thus the modulation in speech can be expected to carry speaker-specific information. In addition to this, the initial volume and mass of the air flow entering through the vocal cord depends on the size and properties of the vocal cord, thus it can be assumed to contain speaker-specific information. Further, using cochlear implant subjects, a speaker recognition experiment [5] comparing AM only and AM+FM found that the performance is improved with FM. This experiment also showed that the FM contains speaker-specific information.

Based on the above motivations, we propose to use FM components from speech to improve the performance of an automatic FSR system. Evaluation of the FM features, MFCCs and the combination of both on the NIST 2001 cellular database shows that FM can be used to improve the performance of the automatic FSR system.

2. Frequency Modulation of Speech

2.1. AM-FM Model

Based on Teager's [4] experimental findings of the existence of modulations in speech, the resonances are each modeled as AM-FM signals in [10]. Then the total speech is taken as the sum of all the resonances as given in equation 1.

$$x[n]=\sum_{k=1}^{K}A_k[n]\cos\left[\frac{2\pi f_{c_k}n}{f_s}+\frac{2\pi}{f_s}\sum_{r=1}^{n}q_k[r]+\theta\right], \qquad (1)$$

where K is the total number of resonances, $A_k[n]$ is the time-varying amplitude component, $q_k[r]$ is the time-varying frequency component, f_{ck} is the center frequency of the resonance, f_s is the sampling frequency and θ is the initial phase. This model is the basis for FM extraction from speech.

2.2. FM Feature Extraction from Speech

In this work, long term average FM features are extracted over a 20 ms window length using the second-order all-pole method [9]. In this method, the speech is initially filtered using Bark scaled Gabor band pass filters. Then each sub band signal is approximated as the impulse response of a second order resonator, from which the pole frequency is estimated. From the pole frequency, taken as an estimate of the instantaneous frequency, the FM component is calculated by subtracting the center frequency of the sub band. The block diagram for the FM extraction is given in Figure 1. As the database used is a cellular database, 14 filters ranging from 300 Hz to 3400 Hz are used.

In principle, it is possible to decompose the speech signal into components due to each resonance, instead of using fixed bandwidth band pass filters, however in practice formant (resonance) tracking approaches pose two problems for speech front-ends: (i) formant tracking is imperfect, and inaccuracies in formant frequency estimates cause problems in the resulting FM extraction; and (ii) in most pattern recognition approaches, a fixed-dimension feature vector is required, while the number of formants (and hence the feature dimension) may vary for a fixed bandwidth. These problems were presumably also anticipated by the fixed six-band Mel-spaced Gabor filter bank proposed in the AM-FM front-end of [8] for speech processing.

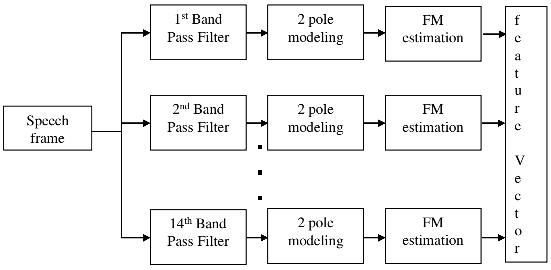

Figure 1. Block diagram of FM feature extraction

2.3. Demonstration of the Significance of FM

A speech signal was reconstructed using AM components only and using both AM and FM from a speech signal in the NIST 2001 database, to observe the effect of FM. For the AM-only reconstruction, the FM was set to zero, as in equation (2), and for the AM-FM reconstruction, the model in equation (1) was used.

$$x[n] = \sum_{k=1}^{K} A_k[n] \cos\left[\frac{2\pi f_{c_k} n}{f_s} + \theta\right] \tag{2}$$

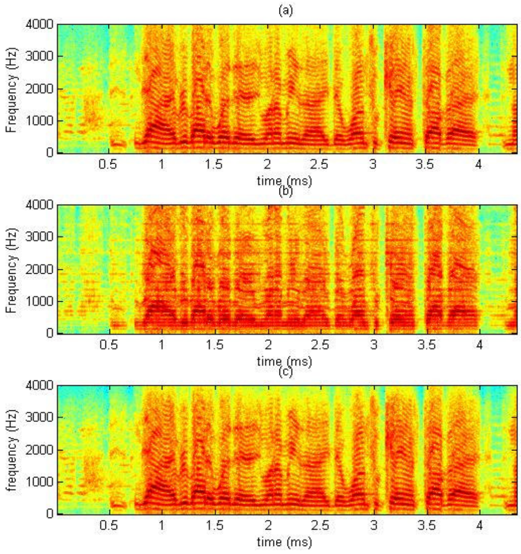

Figure 2. Spectrogram of a speech signal from NIST 2001 database (a) Original speech (b) Reconstructed speech from AM only, and (c) Reconstructed speech from AM and FM.

The formant transitions are much clearer in the spectrogram of speech reconstructed using AM and FM (Fig. 2(c)) compared with the spectrogram of speech reconstructed using AM only (Fig. 2(b)). Almost perfect reconstruction is obtained with AM and FM. This observation, for a real speech signal, is consistent with the previous analysis using synthetic syllables [7].

Further, the FM variation for two speakers for a vowel sound /o/ is shown in Figure 3 for seven sub bands, where the FM frequency has been offset from the band center frequencies. This figure shows the ability of FM for speaker discrimination, particularly the top band (centered at 3400 Hz) and the bottom band (centered at 450 Hz) produce more discrimination. This figure, together with the observation from Figure 2 that the formant transitions are carried by FM, suggests that FM carries mostly the same information as the traditional F-patterns used in [11]. However, further research is required to support this claim.

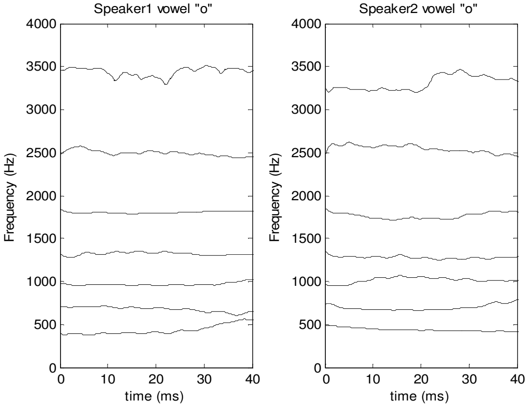

Figure 3. FM variation for a vowel sound /o/ for two speakers for seven bands.

3. Automatic Forensic Speaker Recognition System

The automatic FSR system used in this paper is based on a Bayesian interpretation, having a two-stage modeling approach [12, 13]. In the first stage, the acoustic features are modeled using Gaussian mixture models (GMMs) similar to the conventional automatic speaker recognition used in NIST evaluations. In the second stage, the scores of the within-source (intra-speaker variation of suspect) and scores of the between-source (inter-speaker variation of the potential population) are used to model the within-source and between-source distributions. In this paper, the within-source distribution is modeled by a Gaussian distribution and the between-source distribution is modeled by a kernel density estimation approach similar to [14]. Finally, the *LR* estimate of the automatic FSR system is calculated as in equation (3)

$$LR = \frac{\Pr(E/H_0,I)}{\Pr(E/H_1,I)} \tag{3}$$

where E is the evidence, H_0 is the hypothesis that the evidence was spoken by the suspect, H_1 is the hypothesis that the evidence was not spoken by the suspect and I is the related background information.

For this experiment, the cellular NIST 2001 database was used. It consists of 174 target speakers with 1038 test segments recorded under different environmental conditions. Each test segment is considered to be the questioned recording, and the corresponding claimed speaker is considered to be the suspect. All speakers in the training data other than the suspect are considered as the potential population (P), and all the test segments from the suspect other than the test segment considered are taken as the suspect control database (C). The training data of the claimed speaker is considered to be the suspect reference database (R). The block diagram of the automatic FSR system is given Figure 4.

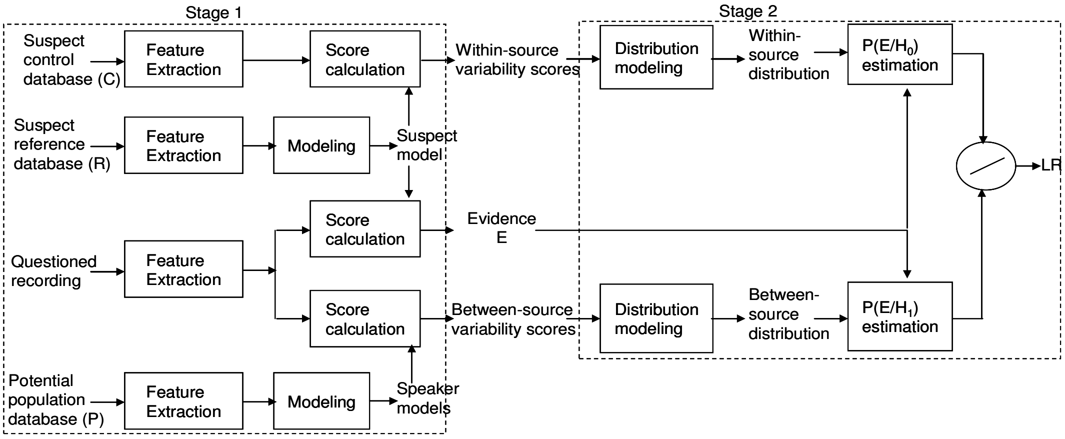

Figure 4. Block diagram of automatic forensic speaker recognition system.

3.1. Performance Measures and Calibration

We used conventional performance measures, such as Tippett plots [13] and Detection Error Trade-off (DET) curves as performance measures. In addition, we calculated C_{llr}, the application-independent evaluation metric proposed by Brümmer [15]. A brief summary of the performance measure proposed in [15] is given below.

The speaker recognition system can be decomposed into two stages: (i) the extraction stage and (ii) the presentation stage. The extraction stage is completed with the likelihood obtained using the block diagram in Figure 4. Then the presentation stage makes the likelihood ratio directly interpretable as a log likelihood ratio, and is achieved by calibration. In our experiment, logistic regression is used for calibration as used in [16].

The quality of the extraction stage is measured by the discrimination loss (C_{llr}^{min}) and the quality of the presentation stage is measured by the calibration loss (C_{llr}^{cal}), while the sum of these two losses is C_{llr} [15]. C_{llr} is calculated using equation (4) [15], where N_0 is the number of trials corresponding to hypothesis H_0 and N_1 is the number of trials corresponding to hypothesis H_1.

$$C_{llr} = \frac{1}{2*\log 2}\left[\frac{1}{N_0}\sum_{i=1}^{i=N_0}\left(1+\frac{1}{LR_i}\right)+\frac{1}{N_1}\sum_{i=1}^{i=N_1}\left(1+LR_i\right)\right] \quad (4)$$

Then (C_{llr}^{min}) is calculated using LR values calibrated using the algorithm of pair-adjacent violators (PAV). This PAV algorithm is a non parametric perfect calibration (with zero calibration loss), thus C_{llr}^{min} measures the quality of the extraction stage. For an ideal system C_{llr}^{min}, C_{llr}^{cal}, and C_{llr} are zero. In real systems, smaller values indicate a better system. Further, a graphical way of presenting these performance measures as applied probability of error (APE) curves [15] is also used in this experiment. The FoCal toolbox is used in this experiment to calculate these performance measures (http://www.dsp.sun.ac.za/~nbrummer).

4. Experiment

In this experiment, 12-dimensional MFCCs, 14-dimensional FM features [9] and the concatenation of both were used. Feature warping and TNorm were performed in all three cases. Feature-level concatenation of MFCC and FM was preferred to score-level fusion of the individual sub-systems, since it can use the joint distribution between MFCC and FM. Although the feature dimensions are assumed independent, the correlation among them is not strictly zero in practice. Usually diagonal covariance is used in GMM-based speaker modeling (although in such cases the correlation information modeled by a full covariance matrix can be equally captured using a higher number of mixtures together with diagonal covariance matrices [17]).

The DET curves for the automatic FSR system are shown in Figure 5, where the combined MFCC and FM improve the error rate. Tippett plots of the un-calibrated and calibrated systems are shown in Figure 6 and 7 respectively, where the combined MFCC and FM produce more separation compared with MFCC in both cases. It is worth noting that the calibration will not change the DET curves, as calibration only improves the presentation of the information extracted by the system. This system was calibrated using an affine transformation with a logistic regression objective as per [16], using the FoCal toolbox. For calibration training, 10% of the database was used. The C_{llr} related performance is given in Table 1. Not only the discrimination loss but also the calibration loss of the calibrated automatic FSR system is improved by combining FM with MFCC. In particular, the calibration loss of FM is less than that of MFCC, showing that the FM is better in the presentation stage. Again, the discrimination loss will not be changed by calibration. An APE curve for the calibrated system is shown in Figure 8. The height of the light portion of the bar chart is the area under the solid curve (error-rate of the PAV optimized score) and the height of the complete bar chart is the area under the dotted curve (error rate of the calibrated system). The dashed curve is the error-rate of the neutral system [15].

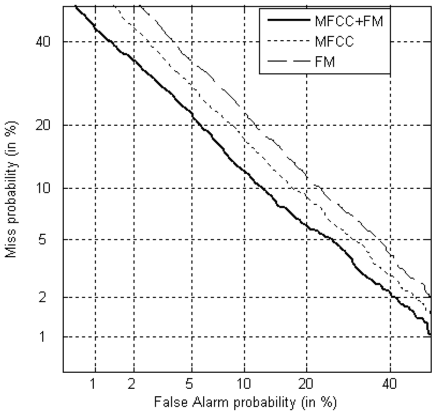

Figure 5. DET curves for the automatic FSR system

5. Conclusion

We have successfully utilized frequency modulation for automatic FSR system to complement MFCC, supporting the psychophysical evidence that FM is a significant component in human perception. The combination of MFCC and FM as a feature gave reduced error rates in terms of DET curves, more separation in terms of Tippett plots and produced smaller discrimination and calibration loss in terms of C_{llr}^{min} and C_{llr}^{cal} respectively. The finding that the calibration loss of FM is better than that of MFCC shows the capability of FM in automatic FSR. The importance of using FM for speaker recognition is also demonstrated. As future work we are investigating the performance of automatic FSR system using recent NIST databases such as NIST 2006 with channel compensation such as nuisance attribute projection (NAP) in the stage one and recently proposed suspect adapted [16] within source modeling in the second stage.

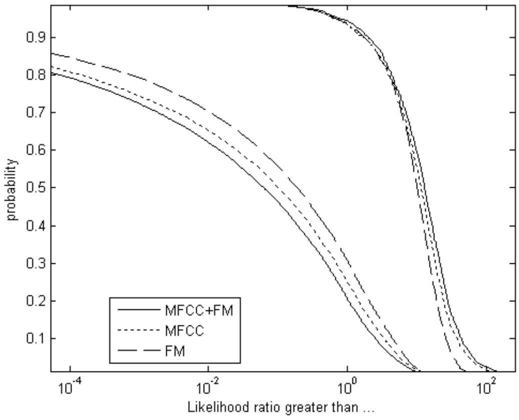

Figure 6. Tippett plots of the un-calibrated automatic FSR system.

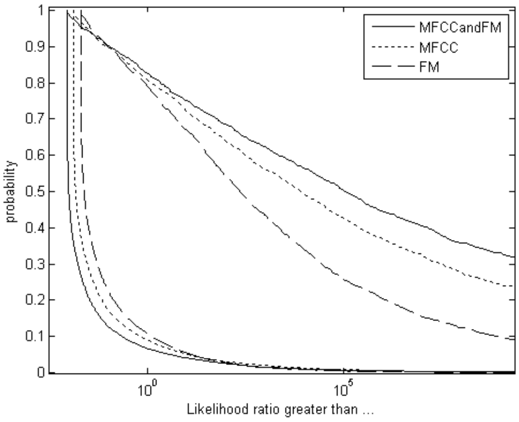

Figure 7. Tippett plots of the calibrated automatic FSR system.

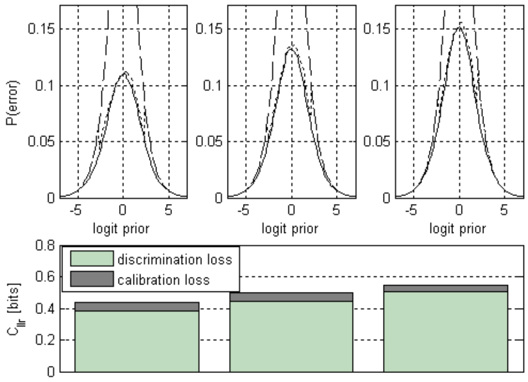

Figure 8. Applied probability of error curves of the calibrated automatic FSR system.

Table 1. Calibration loss and discrimination loss for the calibrated automatic FSR system.

	Calibration loss (C_{llr}^{cal})	Discrimination loss (C_{llr}^{min})
FM	0.0514	0.5027
MFCC	0.0707	0.4452
MFCC+FM	0.0665	0.3848

6. References

[1] R. H. Kay and D. R. Matthews, "On the existence in human auditory pathways of channels selectively tuned to the modulation present in frequency-modulated tones," *J. of Physiology,* vol. 225, pp. 657-677, 1972.

[2] D. Regan and B. W. Tansley, "Selective adaptation to frequency-modulated tones: Evidence for an information-processing channel selectively sensitive to frequency changes," *JASA,* v. 65, pp. 1249-1257, 1979.

[3] P. Maragos, J. F. Kaiser, and T. F. Quatieri, "Energy separation in signal modulations with application to speech analysis," *IEEE Transactions on Signal Processing,* vol. 41, pp. 3024-3051, 1993.

[4] H. Teager, "Some observations on oral air flow during phonation," *IEEE Transactions on Acoustics, Speech, and Signal Processing,* vol. 28, pp. 599-601, 1980.

[5] F. G. Zeng, K. Nie, G. S. Stickney, Y. Y. Kong, M. Vongphoe, A. Bhargave, W. Chaogang, and C. Keli, "Speech recognition with amplitude and frequency modulations," *Proc. Nat. Acad. of Sciences of the USA,* vol. 102, pp. 2293-8, 2005.

[6] L. D. Alsteris and K. K. Paliwal, "Short-time phase spectrum in speech processing: A review and some experimental results," *Digital Signal Processing: A Review Journal,* vol. 17, pp. 578-616, 2007.

[7] N. Kaibao, G. Stickney, and Z. Fan-Gang, "Encoding frequency Modulation to improve cochlear implant performance in noise," *IEEE Transactions on Biomedical Engineering,* vol. 52, pp. 64-73, 2005.

[8] D. V. Dimitriadis, P. Maragos, and A. Potamianos, "Robust AM-FM Features for Speech Recognition," *IEEE Sig. Proc. Letters,* vol. 12, pp. 621-624, 2005.

[9] T. Thiruvaran, E. Ambikairajah, and J. Epps, "Extraction of FM components from speech signals using all-pole model," *Electronics Letters,* vol. 44, pp. 449-450, 2008.

[10] P. Maragos, T. F. Quatieri, and J. F. Kaiser, "Speech nonlinearities, modulations, and energy operators," in *Proc. IEEE ICASSP*, 1991, pp. 421-424.

[11] Phil Rose, Y. Kinoshita, and T. Alderman, "Realistic Extrinsic Forensic Speaker Discrimination with the Diphthong /ai/," in *Proc. SST*, 2006, pp. 329-334.

[12] A. Drygajlo, D. Meuwly, and A. Alexander, "Statistical Methods and Bayesian Interpretation of Evidence in Forensic Automatic Speaker Recognition," in *Proc. Eurospeech*, 2003.

[13] J. Gonzalez-Rodriguez, A. Drygajlo, D. Ramos-Castro, M. Garcia-Gomar, and J. Ortega-Garcia, "Robust estimation, interpretation and assessment of likelihood ratios in forensic speaker recognition," *Computer Speech & Language,* vol. 20, pp. 331-355, 2006.

[14] J. Gonzalez-Rodriguez, J. Fierrez-Aguilar, and J. Ortega-Garcia, "Forensic identification reporting using automatic speaker recognition systems," in *Proc. IEEE ICASSP*, Hong Kong, China, 2003, pp. 93-6.

[15] N. Brummer and J. du Preez, "Application-independent evaluation of speaker detection," *Computer Speech & Language,* vol. 20, pp. 230-275, 2006.

[16] J. Gonzalez-Rodriguez, P. Rose, D. Ramos, D. T. Toledano, and J. Ortega-Garcia, "Emulating DNA: Rigorous Quantification of Evidential Weight in Transparent and Testable Forensic Speaker Recognition," *IEEE Transactions on Audio, Speech, and Language Processing, ,* vol. 15, pp. 2104-2115, 2007.

[17] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Sig. Proc.,* vol. 10, pp. 19-41, 2000.