A Preliminary Study on the Feature Distribution of Deceptive Speech Signals [⋆]

Xinyu Pan [a,b,*], Heming Zhao [a], Yan Zhou [a], Cheng Fan [a]
Wei Zou [a], Zhiqiang Ren [a], Xueqin Chen [a]

[a] *School of Electronics and Information Engineering, Soochow University, Suzhou 215006, China*

[b] *School of Electronics & Information Engineering, Suzhou University of Science and Technology*
Suzhou 215009, China

Abstract

A preliminary study is conducted to compare the feature distribution between normal and deceptive speech, and the results are reported in this paper. The objective of this research is to show that deceptive speech may be recognized through the acoustic parameters of general speech characteristics. Six speech parameters, i.e., Mel-frequency Cepstral Coefficients (MFCC), Relative Spectral Filter Perceptual Linear Prediction (RASTA-PLP), pitch frequency, time-domain samples, zero-crossing rate and fractal dimension are used in the statistics. The distributions of these parameters indicate clear differences between the two speech styles. The lowest average degree of difference for these features was 4.74%, and the highest degree was over 20%. Therefore, the distribution demonstrates that there is significant distinction between speech relating the truth and speech relating falsehoods. Linear Discriminant Analysis (LDA) and the Gaussian Mixture Model (GMM) are used to recognize the two psychological states of people's pronunciation, with accuracy above 50%. The results show that there is in fact deceptive information in speech signals and that it can be detected by pattern recognition. These findings provide the theoretical basis for detecting deception in speech signals.

Keywords: Deceptive Speech; Feature Distribution; MFCC; RASTA-PLP; LDA; GMM

1 Introduction

Speech signals include information regarding gender, identity, language and age. A speaker's mental status can be the subject of speculation by listeners throughout the conversation. However, mental status is very difficult to discern by means of computer identification because of limited

information. A number of successful studies show that emotions and even personality traits can be perceived by a computer through speech signals [1-3]. An instrument that might distinguish between lies and the truth would be a great contribution to many security applications, among other types of applications. Thus, it is of interest to distinguish between truth and deception through speech signals.

Traditional measurements have been used to judge the truth or falsity of voices, such as Psychological Stress Evaluators (PSE) and Voice Stress Analyzers (VSA). Additionally, Layered Voice Analysis (LVA) based on brain activity can also provide evidence regarding lying. However, the results from experiments utilizing these measures indicate that there is a great deal of room for improvement [4-7].

Only a limited amount research has evaluated the importance of different speech characteristics in detecting deceptive speech. Although the F1, F2 and F3 vowel formants cannot establish a correlation between truthfulness and deceptiveness, there may yet be a way to detect deception using acoustic parameters [8].

There is no adequate theoretical basis for lie detection using only voice signals [9], but feature selection can determine the performance of speech-processing systems [10]. Thus, it is important to establish the scientific support that digital speech signal features contain adequate information for detecting deception. It is not clear whether all or only some features of speech signals change when people tell lies, and there is no appropriate method for measuring the degree of the mutation. The characteristics of normal and deceptive speech including Mel-frequency Cepstral Coefficients (MFCC) [11, 12], Perceptual Linear Prediction (PLP) [13], pitch frequency, time-domain samples, zero crossing rate and fractal dimensions are used in the statistical analysis of this paper. The results of the experiment show that the distribution of these features in truthful and deceptive speech acts is clearly different. To a certain extent, the findings provide a theoretical basis for lie detection using only voice signals.

This paper is organized as follows: The next section describes the speech database and briefly introduces the related speech characteristics. Section 3 presents the statistical results of all of the speech features and the analyses of the differences calculated from the distribution of these features. The deception detection results are presented in section 4. Finally, the conclusions are provided in the last section.

2   Description of the Corpus and Speech Features

The speech database and six speech features (MFCC, PLP, pitch frequency, time-domain samples, zero crossing rate and fractal dimensions) are briefly described in this section.

2.1   Corpus

The speech corpus was obtained from a game. In this game, each actor (or actress) tells a complicated story, and another participant can ask the actor (or actress) any type of question. The participant then decides on the truth or falsehood of the statement based on the answers to the questions. The deceitful actors are chosen for our speech database. Each actor's voice is selected manually. We take the actor's self-introduction as the truth and the description of the story as the deception. In this way, 30 actors, 15 males and 15 females are selected. All speech is

mono sampled at 8 kHz and quantified with 16 bits. The frame length is 20 ms, and the overlap is 10 ms when the speech is under short-time analysis.

2.2   Time-domain Samples

Original speech signal $x(t)$ has not received sufficient attention in previous deception detection studies, yet all deceptive information is contained in $x(t)$. This paper studies the distribution of discrete-time speech samples $x(n)$ in the time domain and attempts to reveal differences between normal and deceptive speech.

2.3   Zero-crossing Rate

Zero-crossing rate refers to the frequency of speech signal changes in the time domain. The short-time zero-crossing rate takes the form of Eq. (1).

$$Z = \frac{1}{2N} \sum_{m=1}^{N} |\text{sgn}(x[m]) - \text{sgn}(x[m-1])| \tag{1}$$

where N is the length of the speech frame and the sign function is defined as follows:

$$\text{sgn}(x[n]) = \begin{cases} 1 & x[n] \geq 0 \\ -1 & x[n] < 0 \end{cases} \tag{2}$$

2.4   Fundamental Frequency

Pitch can be obtained with an autocorrelation function. The discrete time signal's autocorrelation function is defined by Eq. (3).

$$\phi[k] = \sum_{m=1}^{N} x[m]x[m+k] \tag{3}$$

$\phi[k]$ is a sharp impulse at the pitch period in voiced sound and is flat in unvoiced sound. Thus, the fundamental frequency can be easily obtained from the peak of the autocorrelation sequence.

2.5   Fractal Dimension

A fractal dimension presents the complexity of a set under variety scales, and it shows detailed information regarding pattern changes. The box-counting dimension is used in this experiment because of the simplicity of calculation. Boxes with different side lengths may be used to comprise the speech set, in which case the fractal dimension is expressed by the correlation of the number of boxes and the side lengths. Thus, suppose the speech waveform is covered by boxes of side length ε and the number of these boxes is $N(\varepsilon)$. The box-counting dimension can be calculated as:

$$D_{box}(F) = \lim_{\varepsilon \to 0} \frac{\ln(N(\varepsilon))}{\ln(1/\varepsilon)} \tag{4}$$

2.6   Mel-frequency Cepstral Coefficients (MFCCs)

A feature that could identify deceptive speech and discard all other content would have been suitable for this experiment, but no such feature exists to date. Mel-frequency cepstral coefficients (MFCCs) have played a useful role in speech recognition and speaker identification systems. Therefore, it is necessary to evaluate the effectiveness of MFCCs in detecting deception.

The steps for calculating MFCCs are shown below:

Step 1. Divide the speech into short time frames.

Step 2. Calculate the power spectrum.

Step 3. Divide the power spectrum based on the Mel-scale filter bank and total the energy in each bank. The Mel-frequency scale formula is:

$$M(f) = 1125 \ln(1 + f/700) \tag{5}$$

Step 4. Take the discrete cosine transform of the above results $X(k)$ as Eq. (6), and let C_n denote the MFCCs.

$$C_n = \sqrt{\frac{2}{N}} \sum_{k=1}^{M} Log X(k) \cos[\pi(k - 0.5)n/M] \tag{6}$$

Here, M is the order of the MFCCs.

Step 5. If necessary, Eq. (7) can be used to calculate the delta coefficients.

$$d_t = \frac{\sum_{n=1}^{N} n(C_{t+n} - C_{t-n})}{2 \sum_{n=1}^{N} n^2} \tag{7}$$

Here, d_t is the delta coefficient of frame t, and N is typically 2. Delta-delta coefficients are also calculated in this way.

Step 6. The values for C_n and d_t comprise the MFCC vector.

2.7   Relative Spectral Perceptual Linear Predictive (RASTA-PLP)

PLP is based on the theory of psychoacoustic conversions, which is different from tradition linear prediction. It replaces the short-time speech spectrum with a relative auditory spectrum. Similar to other spectrum analysis functions, PLP may be distorted by the frequency response of the noise or the channels. Thus, the relative spectral filter (RASTA) is introduced to remove the distortion before the PLP analysis. The RASTA-PLP can be obtained by the following steps:

Step 1. Divide the speech into frames as the first step in calculating MFCCs.

Step 2. Compute the logarithm of the short-time critical-band spectrum. The conversion from angular frequency ω to bark frequency Ω is shown as Eq. (8).

$$\Omega(\omega) = 6 \ln \left\{ \omega/1200\pi + [(\omega/1200\pi)^2 + 1]^{0.5} \right\} \tag{8}$$

Step 3. Take the temporal derivative of the spectrum above based on a regression line.

Step 4. Employ an IIR system to the log critical-band temporal derivative filtering.

Step 5. Add the equal loudness curve to simulate the hearing rules.

Step 6. Obtain the relative auditory spectrums after the inverse logarithm of this log spectrum has been calculated.

Step 7. Calculate all of the spectrum's pole models, and then the PLP parameters have been met.

3 Statistical Results and Simple Analysis

3.1 Distribution Figures

Psychophysiology theory assumes that liars are more nervous, which may cause various changes in the vocal organs during pronunciation: formants may be shifted, bandwidth may be expanded, and/or the pitch track may deviate. The statistical results from this experiment show that most features were more or less able to indicate the differences between normal and deceptive speech.

All the features described in Section 2 were used in the statistical analysis, and each showed differences between normal and deceptive speech. Because of differences in pronunciation, the statistics for males and females were calculated separately. Twelve MFCC coefficients and 12 delta coefficients were used as the MFCC vector, and 13 RASTA-PLP coefficients were selected as the RASTA-PLP vector. Each dimension of the high-dimensional data, such as MFCCs and RASTA-PLPs, was calculated separately. There was no special processing with one-dimensional data, such as speech samples, pitch frequency, zero-crossing rate or fractal dimension. The distribution results are shown in Figs. 1-8.

3.2 Analysis

Suppose that the distribution data of normal speech is $f_N(x)$ (dotted line in figures) and the $f_L(x)$ (solid line in figures) presents the deceptive speech distribution. The degree of difference between normal and deceptive speech is defined by Eq. (9).

$$p = \frac{\int\limits_{-\infty}^{+\infty} |f_N(x) - f_L(x)|dx}{2} \tag{9}$$

The average degree of MFCC and RASTA-PLP differences was $\bar{P} = \sum\limits_{i=1}^{N} p_i/N$, where p_i denotes the different degree of the i-th dimension. According to Eq. (9), this degree ranges from 0 to 1 and is expressed as a percentage.

The degrees of difference of each characteristic are shown in Tables 1–5, and the comparison of each average degree of difference is shown in Fig. 9.

According to Tables 1 and 2, the average degree of difference in males' MFCCs was 18.22%, and that of females' MFCCs was 12.89%. The degrees of difference for MFCC9, MFCC10, MFCC11, Delta MFCC2, and Delta MFCC7–12 were greater than the average degree in males' speech, and
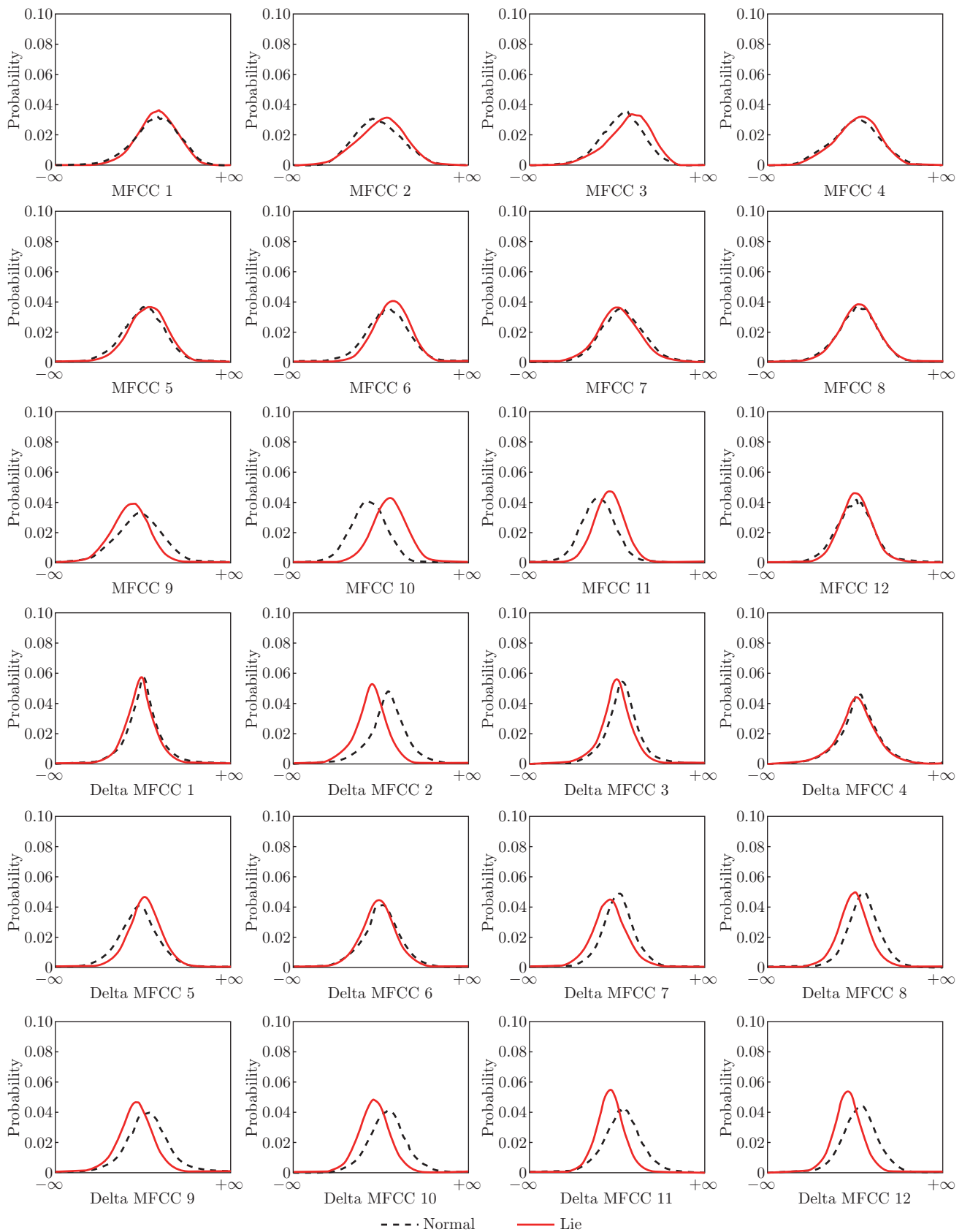
Fig. 1: Distribution of the 12 MFCC and 12 Delta MFCC coefficients between the normal and deceptive speech of males
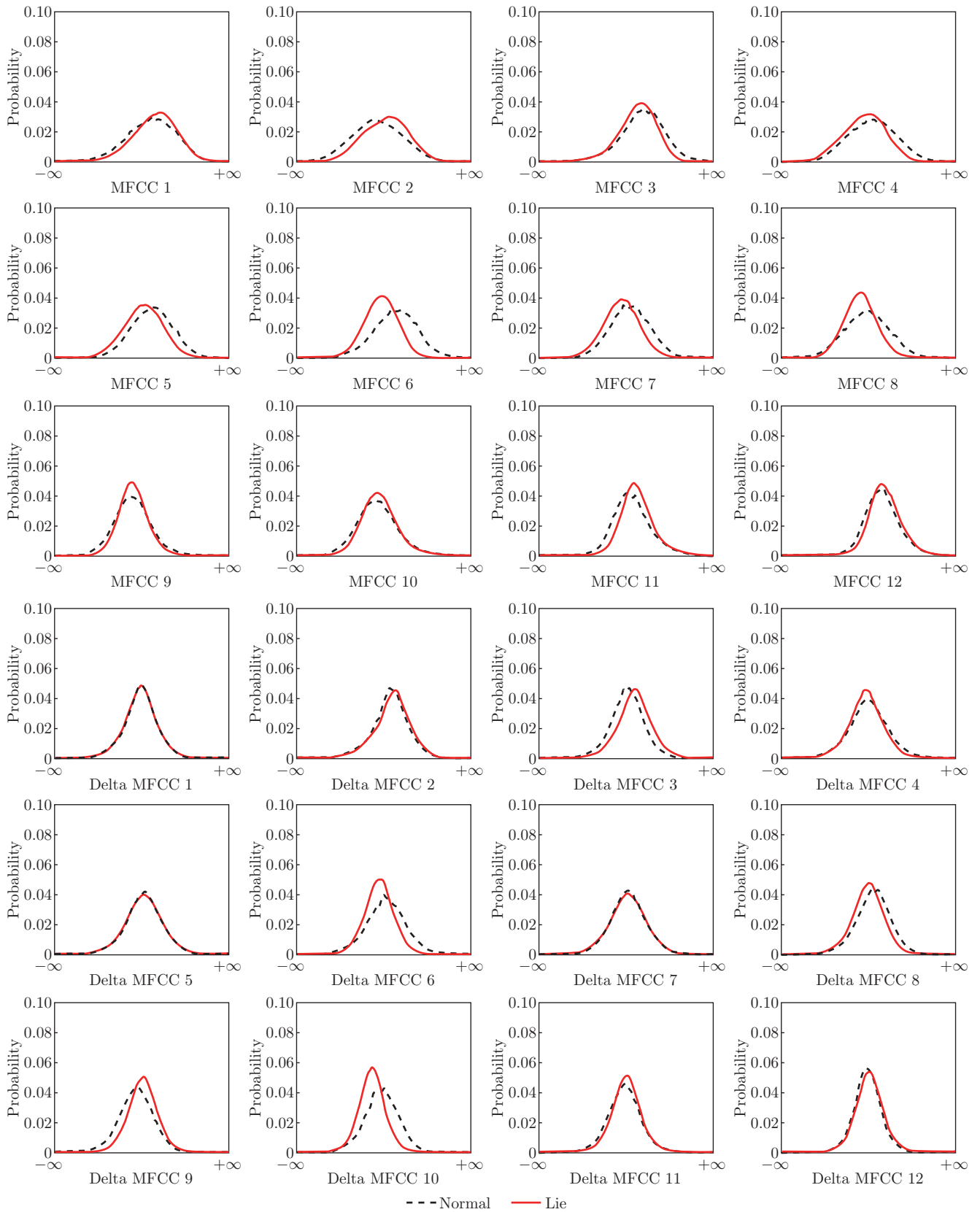
Fig. 2: Distribution of the 12 MFCC and 12 Delta MFCC coefficients between the normal and deceptive speech of females
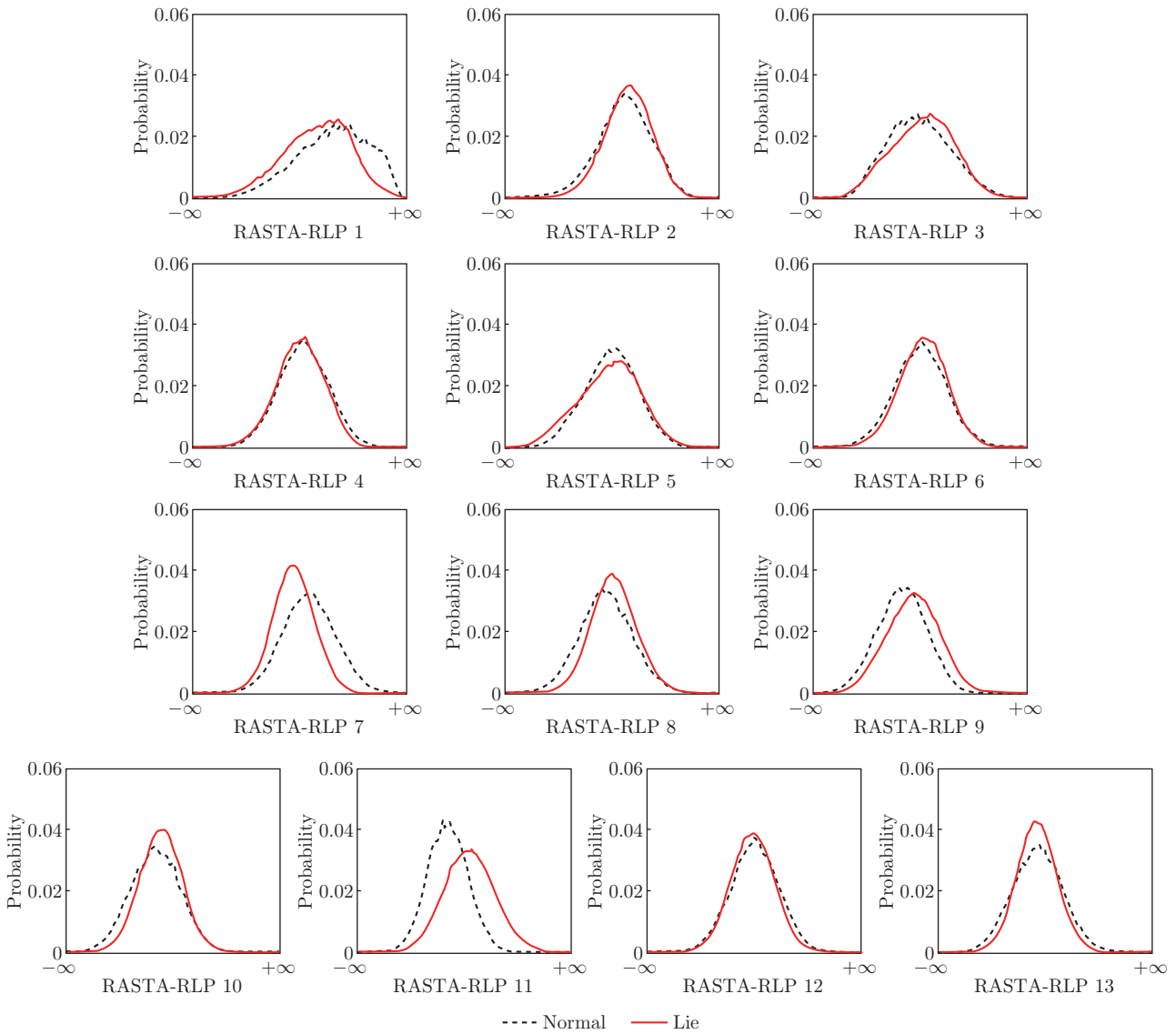
Fig. 3: Distribution of the 13 RASTA-PLP coefficients between the normal and deceptive of males

Table 1: Degrees of difference for the males' MFCCs

Character	MFCC 1~12											
Degree of difference (%)	4.62	9.59	14.52	5.58	7.51	11.07	7.27	2.70	19.14	45.63	28.03	7.01
Character	Delta MFCC 1~12											
Degree of difference (%)	10.49	43.05	16.46	6.63	16.13	8.42	23.60	26.89	25.01	32.04	33.30	32.66

the degrees of difference for MFCC2, MFCC4–8, MFCC11, Delta MFCC3, Delta MFCC6, and Delta MFCC 8–10 were greater than the average degree in females' speech. According to Tables 3 and 4, the average degree of difference in males' RASTA-PLPs was 11.95%, and that for females' RASTA-PLPs was 12.31%. The degrees of difference for RASTA-PLP1, RASTA-PLP7, RASTA-PLP9, and RASTA-PLP11 were greater than the average degree in males' speech, and the degrees of difference for RASTA-PLP2–3, RASTA-PLP8, and RASTA-PLP10–11 were greater than the
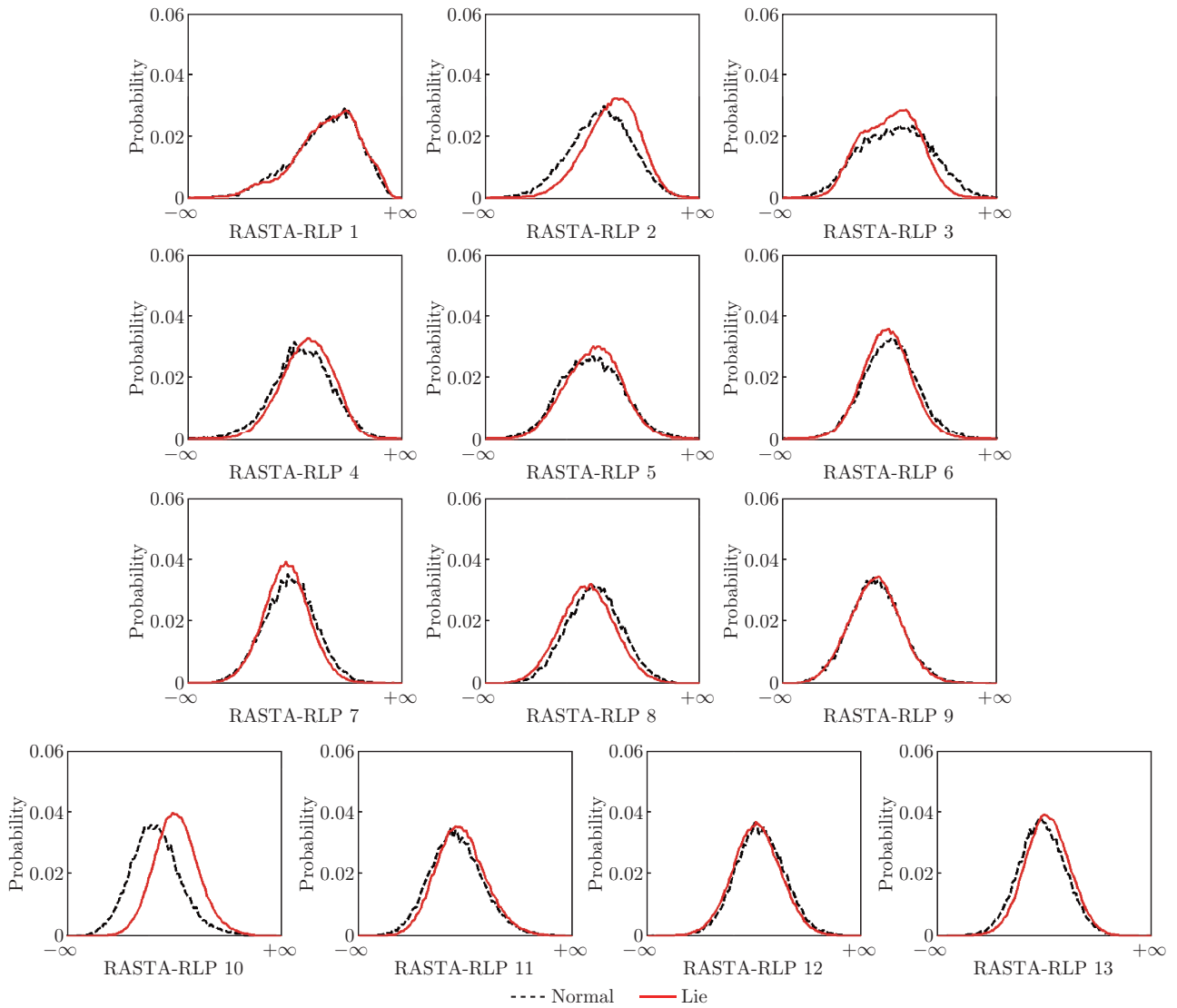
Fig. 4: Distribution of the 13 RASTA-PLP coefficients between the normal and deceptive of females
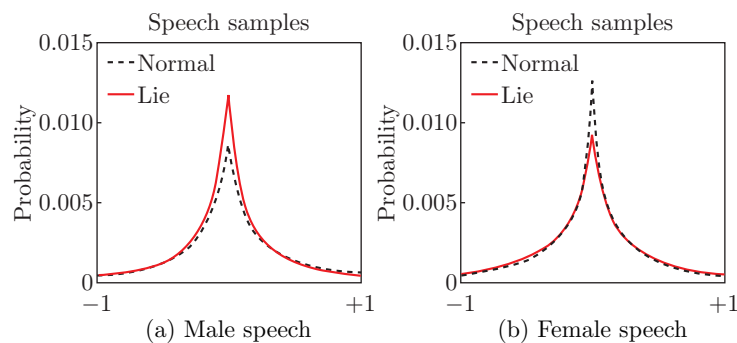

Fig. 5: Distribution of speech samples between normal and deceptive speech

average degree in females' speech. According to Table 5, the one-dimensional characteristics also performed differently in normal vs. deceptive speech.

Information theory may be a simple way to interpret the above data. Thirty people participated
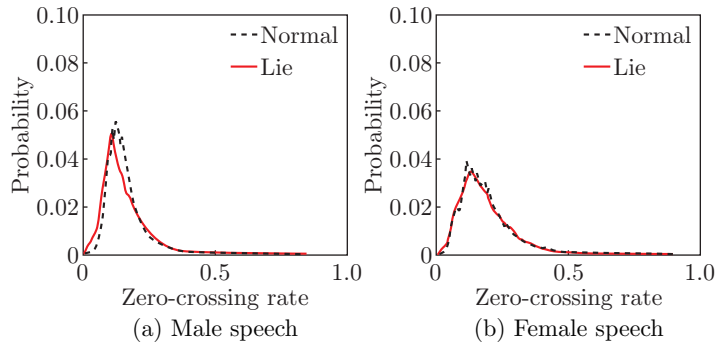
Fig. 6: Distribution of zero-crossing rates between normal and deceptive speech
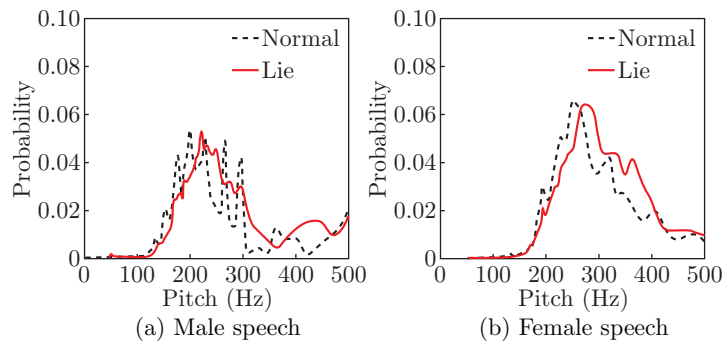

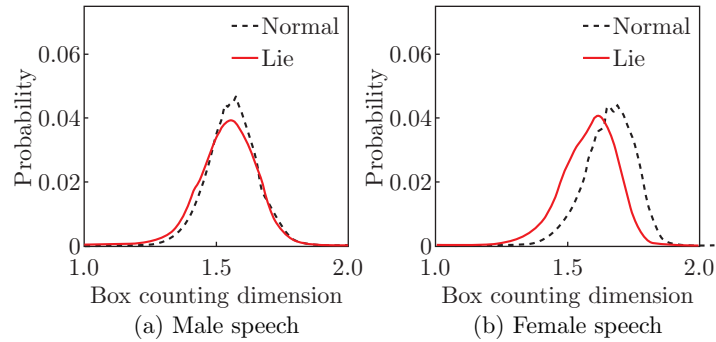Fig. 7: Distribution of pitch frequencies between normal and deceptive speech

Fig. 8: Distribution of the box counting dimension between normal and deceptive speech

Table 2: Degrees of difference for the females' MFCCs

Character	MFCC 1~12											
Degree of difference (%)	8.20	15.34	12.21	14.23	16.36	31.67	17.84	19.85	9.69	7.87	14.54	8.01
Character	Delta MFCC 1~12											
Degree of difference (%)	1.98	6.22	18.22	8.64	2.06	20.64	2.76	14.55	17.04	28.02	7.56	5.73

in this experiment. There are 1332 syllables in Mandarin Chinese, and that number reduces to 407 when tone is disregarded. Normal speech and deceptive speech can be regarded as products from a speaker's two different statuses. The total information from each word is $I_{speech} = \log_2(407 \times 30 \times 2) = 14.58 \; bit$, and the status information is $I_{status} = \log_2(2) = 1 \; bit$. Therefore, the proportion
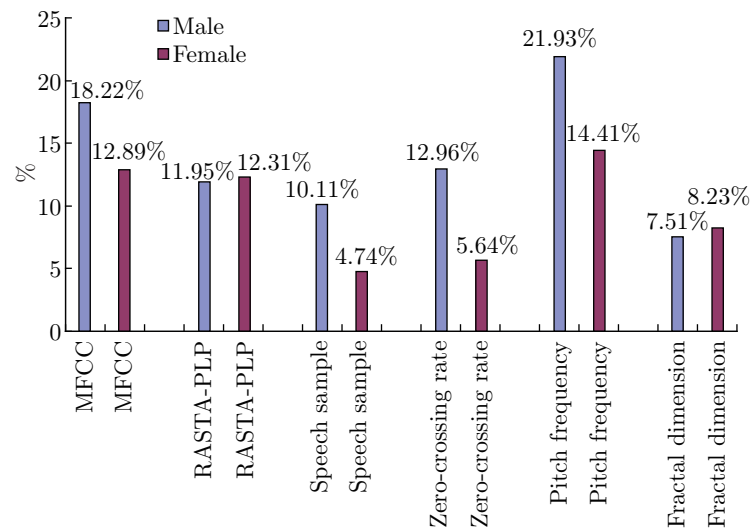
Fig. 9: Average degrees of difference

Table 3: Degrees of difference for the males' RASTA-PLPs

Character	PLP 1~13												
Degree of difference (%)	18.76	5.65	6.60	3.30	5.62	4.44	25.96	7.93	15.94	7.89	33.24	9.27	10.74

Table 4: Degrees of difference for the females' RASTA-PLPs

Character	PLP 1~13												
Degree of difference (%)	3.55	16.55	12.40	9.41	6.57	10.95	8.40	18.97	2.36	44.74	17.71	3.61	4.85

Table 5: Degrees of difference for the remaining 4 characteristics

Character	Speech sample		Zero-crossing rate		Pitch frequency		Fractal dimension	
Degree of difference (%)	males	females	males	females	males	females	males	females
	10.11	4.74	12.96	5.64	21.93	14.41	7.51	8.23

of the status information that accounts for all of the information is $\eta = \dfrac{I_{status}}{I_{speech}} = \dfrac{1}{14.58} = 6.86\%$. Most of the degrees of difference were greater than 6.86%, and thus, it may be speculated that deception information is in fact hidden in speech signals, based on these statistical results.

4   Recognition Result

Linear Discriminant Analysis (LDA) and the Gaussian Mixture Model (GMM) are introduced to identify deceptive speech. These two models are introduced below.

4.1 LDA

There is a training set $X = \{x_1, x_2, \cdots, x_n\}$, and each n-dimensional x_i belongs to a class ω_1 or ω_2. A projection operator w called the best vector may map x_i to the one-dimensional y.

$$y = w^T x \tag{10}$$

Then, the classification is easily selected using a simple comparison,

$$\begin{cases} x_i \in \omega_1 & y \geq t \\ x_i \in \omega_2 & y < t \end{cases} \tag{11}$$

where t is the threshold.

The center of each class is shown in Eq. (12).

$$\mu_i = \frac{1}{N_i} \sum_{x \in \omega_i} x \tag{12}$$

The new center after mapping is shown as Eq. (13).

$$\tilde{\mu}_i = \frac{1}{N_i} \sum_{y \in \omega_i} y = \frac{1}{N_i} \sum_{x \in \omega_i} w^T x = w^T \mu_i \tag{13}$$

Eq. (14) shows the distance between the centers of the two classes.

$$D(w) = |\tilde{\mu}_1 - \tilde{\mu}_2| = |w^T(\mu_1 - \mu_2)| \tag{14}$$

A scatter plot similar to the variance can be calculated in Eq. (15).

$$\tilde{s}_i^2 = \sum_{y \in \omega_i} (y - \tilde{\mu}_i)^2 \tag{15}$$

Then, the final measurement formula can be Eq. (16).

$$J(w) = \frac{|\tilde{\mu}_1 - \tilde{\mu}_2|^2}{\tilde{s}_1^2 + \tilde{s}_2^2} \tag{16}$$

The next task is to identify the suitable w that makes $J(w)$ a maximum.

Eq. (15) can be expanded as Eq. (17).

$$\tilde{s}_i^2 = \sum_{y \in \omega_i} (y - \tilde{\mu}_i)^2 = \sum_{x \in \omega_i} (w^T x - w^T \mu_i)^2 = \sum_{x \in \omega_i} w^T (x - \mu_i)(x - \mu_i)^T w \tag{17}$$

We define

$$s_i = \sum_{x \in \omega_i} (x - \mu_i)(x - \mu_i)^T \tag{18}$$

Here,

$$s_w = s_1 + s_2 \tag{19}$$

And then

$$\tilde{s}_1^2 + \tilde{s}_2^2 = w^T s_w w \tag{20}$$

Such that we have

$$(\tilde{\mu}_1 - \tilde{\mu}_2)^2 = (w^T \mu_1 - w^T \mu_2)^2 = w^T (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T w = w^T s_k w \tag{21}$$

And

$$J(w) = \frac{w^T s_k w}{w^T s_w w} \tag{22}$$

The projection operator w can be calculated in Eq. (23).

$$w = s_w^{-1}(\mu_1 - \mu_2) \tag{23}$$

This result means that the input x is located on one side of a hyperplane by a map factor w.

4.2   GMM

The probability density function $P(X/\lambda)$ of an M-order GMM is constructed by M sum of the weighted Gaussian probability density function.

$$P(X/\lambda) = \sum_{i=1}^{M} w_i b_i(X) \tag{24}$$

Here, X is the speech feature vector, $b_i(X)$ is the Gaussian probability density function, and w_i is the weight.

Expectation Maximization (EM) based on maximum likelihood estimation is a primary method for calculating GMM parameters. The related formulas are shown below.

$$w_i = \frac{1}{T} \sum_{t=1}^{T} P(i/X_t, \lambda) \tag{25}$$

$$\mu_i = \frac{\sum_{t=1}^{T} P(i/X_t, \lambda) X_t}{\sum_{t=1}^{T} P(i/X_t, \lambda)} \tag{26}$$

$$\sigma_i^2 = \frac{\sum_{t=1}^{T} P(i/X_t, \lambda)(X_t - \mu_i)^2}{\sum_{t=1}^{T} P(i/X_t, \lambda)} \tag{27}$$

$$P(i/X_t, \lambda) = \frac{w_i b_i(X_t)}{\sum\limits_{k=1}^{M} w_k b_k(X_t)} \tag{28}$$

Here, μ_i is the mean value, σ_i^2 is the variance, and $P(i/X_t, \lambda)$ is a posteriori probability. The GMM can be established with the steps described above.

4.3   Experimental Result Analysis

The features introduced in Section 2 are all used in this experiment. The feature set is divided into two parts, i.e., the training set and the test set. The training set contains 30% of the full data, and the remaining data are regarded as the test set. Both LDA and GMM are used in the experiment, and the results are shown in Tables 6 and 7, respectively.

Table 6: LDA recognition accuracy

Index	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Male (%)	43.3	50.2	51.3	61.0	62.2	61.0	63.0	54.6	54.5	62.4	44.0	32.8	49.7	46.5	69.2
Female (%)	63.3	61.2	47.0	28.2	40.6	49.1	64.3	58.0	69.0	55.0	49.0	62.4	59.0	78.4	36.3

Table 7: GMM recognition accuracy

Index	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Male (%)	52.6	58.1	48.7	69.2	70.5	63.3	55.9	63.1	59.6	66.6	41.8	40.7	56.8	52.5	78.0
Female (%)	71.4	69.0	53.3	33.9	45.5	54.2	68.1	52.3	75.6	59.1	53.2	58.0	64.5	79.9	38.6

The average accuracy of LDA was 53.7% and 54.7% for male and female speech, respectively. The average GMM accuracy was 58.5% and 58.4% for male speech and female speech, respectively.

Some simple conclusions can be deduced from these recognition results.

(1) Average GMM accuracy is slightly higher than average LDA accuracy. Most of the recognition accuracy for all actors increased with the use of the GMM. The recognition effect of the probability model was better than that of the linear mapping model.

(2) Speech features had the same effect on detecting deception with both male and female speech, and average accuracy was nearly the same for both.

(3) GMM accuracy decreased only when the speech data were not adequate, such as for males 3, 7 and 11 and females 8 and 12. When the speech set contained sufficient syllables, accuracy was sometimes greater than 70%, as with males 5 and 15 and females 1, 9 and 14. Thus, the recognition results correlate with the syllables contained in the speech set.

(4) Information regarding deception can be detected by pattern recognition in speech signals.

5   Conclusions

A number of characteristics that are extracted from normal and deceptive speech are used as statistics in this paper. The differences in normal and deceptive speech are assessed. All these characteristics reflect differences to certain degrees. The speaker's status can be roughly distinguished using LDA and the GMM. Even if the deceptive information retained in the speech signal is very weak, it can nonetheless be detected using the main acoustic features to some extent. Therefore, the results from this experiment also show the possibility of detecting deception using only discrete-time speech signals. The next step is to further explore the correlation of context speech and develop a time series model to improve accuracy.

References

[1] C. Hu, Q. Wang, L. A. Short, G. Fu, Speech spectrum's correlation with speakers' Eysenck personality traits, PLoS ONE, 7(3), 2012, e33906

[2] E. Hill, D. Han, P. Dumouchel, N. Dehak, T. Quatieri et al., Long term SuboxoneTM emotional reactivity as measured by automatic detection in speech, PLoS ONE, 8(7), 2013, e69043

[3] R. Schwartz, M. D. Pell, Emotional speech processing at the intersection of prosody and semantics, PLoS ONE, 7(10), 2012, e47279

[4] A. Eriksson, F. Lacerda, Charlatanry in forensic speech science: A problem to be taken seriously, International Journal of Speech Language and the Law, 14(2), 2007, 169-193

[5] J. D. Harnsberger, H. Hollien, C. A. Martin, K. A. Hollien, Stress and deception in speech: Evaluating layered voice analysis, Journal of Forensic Science, 54(3), 2009, 642-650

[6] L. Anolli, R. Ciceri, The voice of deception: Vocal strategies of Naïve and Able liars, Journal of Nonverbal Behavior, 21(4), 1997, 259-284

[7] K. Gopalan, S. Wenndt, Speech analysis using modulation-based features for detection deception, Proc. of the 15th Intl. Conf. on Digital Signal Processing, 2007, 619-622

[8] K. Christin, M. David, Detecting suspicious behaviour using speech: Acoustic correlates of deceptive speech - An exploratory investigation, Applied Ergonomics, 2012, 1-9

[9] M. Christie, Fuller David P., Biros Dursun Delen, Exploration of feature selection and advanced classification models for high-stakes deception detection, Proceedings of the 41st Hawaii International Conference on System Sciences, 2008, 1-8

[10] Y. Zhou, H. M. Zhao, L. Shang, T. Liu, Immune K-SVD algorithm for dictionary learning in speech denoising, Neurocomputing, 137(5), 2014, 223-233

[11] S. Davis, P. Mermelstein, Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences, IEEE Transactions on Acoustics, Speech, and Signal Processing, 28(4), 1980, 357-366

[12] X. Huang, A. Acero, H. Hon, Spoken Language Processing: A Guide to Theory, Algorithm, and System Development, Prentice Hall, 2001

[13] H. Hermansky, N. Morgan, A. Bayya, P. Kohn, RASTA-PLP speech analysis technique, IEEE International Conference on Acoustics, Speech, and Signal Processing, 1, 1992, 121-124