International Journal of Innovative
Computing, Information and Control
Volume 11, Number 4, August 2015

 ISSN 1349-4198
pp. 1191–1204

RESEARCH ON ALGORITHM OF ROBUST SPEECH PERCEPTUAL HASHING FOR TIME-FREQUENCY DOMAIN BASED ON HILBERT TRANSFORM

Qiuyu Zhang[1], Zhongping Yang[1], Qianyun Zhang[2]
Yibo Huang[1] and Pengfei Xing[1]

[1]School of Computer and Communication
Lanzhou University of Technology
No. 287, Lan-Gong-Ping Road, Lanzhou 730050, P. R. China
zhangqylz@163.com; {zpyang90; xingpengfei0202}@126.com; huang_yibo@foxmail.com

[2]School of Communication and Information Engineering
Shanghai University
No. 98, Shang-Da Road, Shanghai 200444, P. R. China
zhangqy369@126.com

Received September 2014; revised January 2015

Abstract. *In this paper, we present a novel time-frequency domain robust speech perceptual hashing algorithm based on Hilbert transform for speech content authentication, and effectively solve a poor robustness problem of the speech content perceptual authentication under the low-pass filter and white Gaussian noise, identify tamper detection and localization of small scale. Firstly, this algorithm for speech signal pre-processing, evaluated short-time energy of each frame signal as time domain perceptual feature value. Secondly, each frame signal for Hilbert transform and two dimensional discrete cosine transform (2D-DCT) constitute a frequency domain feature matrix and evaluate its Shannon entropy as the frequency domain perceptual feature value. Finally, quantify the time domain and frequency perceptual feature value to get the authentication perceptual hashing sequences. The experiment results illustrate that when compared with the existing schemes, the proposed scheme has a good robustness and discrimination for content-preserving manipulations, the robustness is largely improved under the low-pass filter and white Gaussian noise operations, and lower computational complexity, and can realize precision tamper detection and localization.*

Keywords: Speech content authentication, Perceptual hashing, Hilbert transform, Low-pass filter, Tamper localization

1. **Introduction.** The traditional digest authentication algorithm has a poor robustness, and it cannot be effectively applied to the speech authentication of the speech mobile terminal [1,2]. Due to the particularity of the speech signal, the traditional signature authentication algorithm cannot meet the speech authentication robustness and real-time requirements. On the one hand, the speech signal will be often various external disturbed and attacked in the transmission processing, and also, the speech authentication algorithm robustness has the high request. On the other hand, the instantaneity of speech transmission and the limitation of resource of mobile terminal require higher computational efficiency of speech authentication algorithm.

Speech perceptual feature extraction is the key of speech perceptual hashing authentication algorithm [3], which mainly includes the speech feature extraction and hashing structure. The robustness of speech feature values directly influences the quality of the

structure hashing value and the authentication efficiency. The existing speech perceptual feature extraction schemes and processing algorithms are based on the human ear psychoacoustics model. The feature extraction is mainly for the logarithmic spectrum coefficient [4], linear prediction coefficient [5,6], line frequency spectrum [7], Mel-frequency cepstral coefficients (MFCC) [8,9], discrete wavelet coefficients [10], Hilbert transform spectrum estimation [11] and related derivative parameters [12], etc. Ref. [13] proposed an authentication algorithm based on auditory perceptual properties. Ref. [14] proposed a speech hashing authentication scheme based on the Hypotrochoid diagram. Ref. [15] proposed an audio fingerprint algorithm based on the spectrum energy and non-negative matrix decomposition. Ref. [16] proposed a speech perceptual algorithm based on the linear prediction coefficients, the linear prediction coefficient getting through pretreatment and linear prediction analysis of the original speech, then the non-negative matrix factorization (NMF) of linear prediction coefficient to extraction feature. The experimental results illustrate that the algorithm has a good robustness and discrimination for content preserving operations, but not very good resistance attacks to white Gaussian noise and low-pass filter, the algorithm has a poor robustness under the attacking of the white Gaussian noise and low-pass filter, and the algorithm has a large amount of authentication data, higher complexity, and lower efficiency. At the same time, it is not able to realize the detection and localization of malicious attacks or tampers.

From what has been discussed above, in this paper according to the time-frequency domain characteristics of the speech signal, in view of the existing problem of speech perceptual authentication algorithm poor robustness under such as the low-pass filter and white Gaussian noise content preserving operation, we present a novel time-frequency domain robust speech perceptual hashing algorithm based on Hilbert transform. This algorithm is to extract the speech feature value which can better reflect the original speech signal characteristics and facilitate digest processing of the perceptual hashing function, to build the perceptual hashing which can satisfy the requirements of the nature of the hashing function, and meet the speech information real-time, robustness and security requirements of mobile computing environment. The experimental results illustrate that the proposed algorithm can well resist the low-pass filter and white Gaussian white noise attack, not only having a good robustness for the general content preserving operations, but also having a good robustness for the low-pass filter and white Gaussian noise operation, in addition, a lower computational complexity; at the same time, it can realize accurate tamper localization with speech signal, which solves the problem of existing algorithm that cannot detect and locate the small scale tamper.

2. The Feature Extraction of Speech Signal.

2.1. **The feature extraction in time domain.** Speech signal is a kind of typical non-stationary signal, but in a very short period of time (30 ms), speech signal is short-time smoothly [17]. For the speech $f(t)$, the short energy can be defined as follows:

$$E_n = \sum_{t=n-(N-1)}^{n} [f(t)]^2 \tag{1}$$

where n is the current frame and N is frame length in (1), and the window function by Hamming window is used in this paper.

2.2. **The feature extraction in frequency domain.** The feature extraction contains the following three methods:

A. The Hilbert transform. In the speech signal processing, the signal often needs orthogonal decomposition. Due to the fact that the Hilbert transform can provide 90°

phase change without affecting the spectrum amplitude, the speech signal for Hilbert transform is equivalent to the speech signal for quadrature phase shift, and makes it become own orthogonal. It is advantageous to the time-frequency analysis of speech signal [18]. The speech signal $f(t)$ for Hilbert transform $H[.]$ is as follows:

$$H[f(t)] = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{f(\tau)}{t-\tau} d\tau \tag{2}$$

In fact, the (2) is convolution of the $f(t)$ and $1/\pi t$, the τ is integral variable.

B. The two dimensional discrete cosine transform. The discrete cosine transform (DCT) is similar to the fast Fourier transforms orthogonal transform, whose function is very close to the Karhunen-loeve transform (KTL). Nowadays, the compression algorithm of some very popular audio signal using DCT [19,20], two dimensional discrete cosine transform is defined as follows:

$$F(u,v) = c(u)c(v) \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x,y) \cos\left[\frac{\pi(2x+1)u}{2M}\right] \cos\left[\frac{\pi(2y+1)v}{2N}\right] \tag{3}$$

where the M, N represent the points of discrete cosine transform, $0 \le u \le M-1$, $0 \le v \le N-1$, $c(u) = \begin{cases} \sqrt{1/M}, u = 0 \\ \sqrt{2/M}, 0 \le u \le M-1 \end{cases}$, $c(v) = \begin{cases} \sqrt{1/N}, v = 0 \\ \sqrt{2/N}, 0 \le v \le N-1 \end{cases}$.

C. The Shannon entropy. In the information theory, the information entropy is also called Shannon entropy [21]. The entropy is closely related to speech contents, and the entropy based on speech content has the ability to accurately represent features of speech signal [22]. For the speech signal $f(t)$ after the Hilbert transform and two dimensional discrete cosine transform in the frequency domain, the speech signal $f(t)$ can be represented as follows:

$$F_i = \{F_1, F_2, \cdots, F_n\} \tag{4}$$

To normalized F_i, it is as follows:

$$f_i = \frac{F_i}{\sum_{i=1}^{n} F_i}, \quad 1 \le i \le n \tag{5}$$

Similarly, the Shannon entropy of the speech signal can be defined as follows:

$$H(F_i) = \sum_{i=1}^{n} f_i \log_2 \frac{1}{f_i} = -a \sum_{i=1}^{n} f_i \ln f_i, \quad i = 1, 2, \cdots, n \tag{6}$$

where $a = \log_2 e$ is constant in (6).

3. **Proposed Scheme.** Speech signal is a kind of typical non-stationary signal, but in a very short period time (30 ms), it can be regarded as a stationary signal. In this paper, the speech signal for short-time energy is the time domain feature value. The speech signals for Hilbert transform and two dimensional discrete cosine transform calculate the Shannon entropy as the frequency domain feature value. So it can better reflect the characteristics of the speech signal in the time-frequency domain. Combining with the characteristics of speech signal short-time smoothly, this paper proposed a time-frequency domain robust speech perceptual hashing authentication algorithm based on Hilbert transform. The proposed algorithm block diagram is shown in Figure 1.

The proposed algorithm detailed steps as follows:

Step 1: Pre-processing. The speech library speech signal is for pre-emphasis, and improve the high frequency spectrum, reduce the edge effect, eliminate noise.

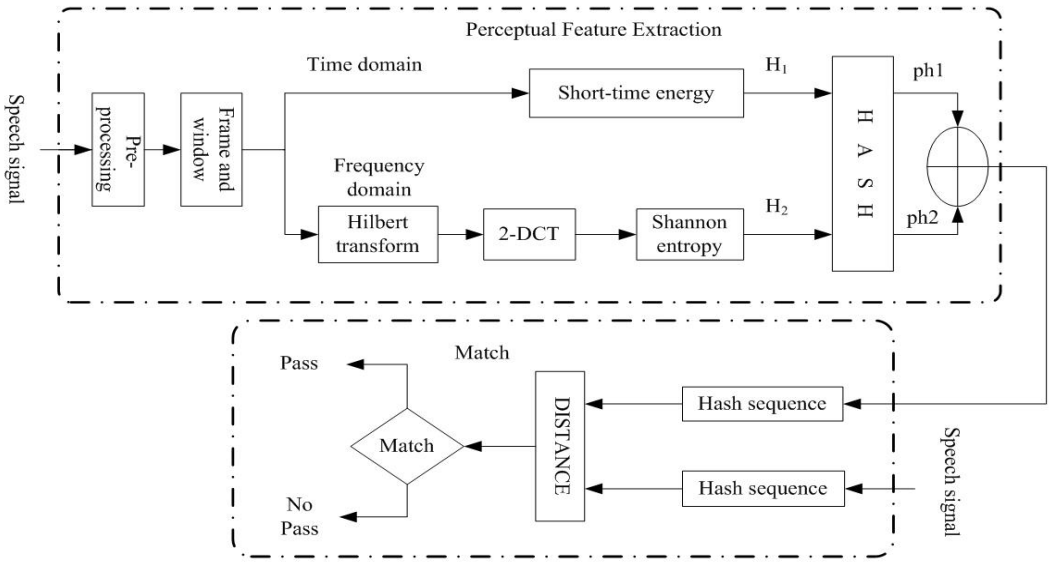


FIGURE 1. The block diagram of the proposed algorithm

Step 2: Adding window and framing. In order to eliminate the inter-frame loss, adding hamming window that is used to smooth the frame edge of speech signal, the original speech signal denoted as S, is split into n equal and overlapping, denoted as $T_i = \{T_i(k)|i = 1, 2, \cdots, n, k = 1, 2, \cdots, L\}$, where L is the frame length, $L/2$ is frame shift.

Step 3: The feature extraction in time domain. According to Section 2.1, the short-time energy of frame signal T_i in time domain is computed to obtain feature sequence H_1.

Step 4: The feature extraction in frequency domain. According to Section 2.2, the frame signal T_i is processing by Hilbert transform and two dimensional discrete cosine transform to assurance perceptual integrity of feature information, and then Shannon entropy is computed to obtain feature sequence H_2.

Step 5: Quantization. The feature sequence e for median quantized is as shown in (7). The feature sequence H_1 and H_2 are similarly quantized via (7), get perceptual sequences $ph1$ and $ph2$ and obtain this algorithm perceptual sequences $H = [ph1, ph2]$.

$$h(i) = \begin{cases} 1 & e(i) \geq \hat{e} \quad 1 \leq i \leq n \\ 0 & e(i) < \hat{e} \quad 1 \leq i \leq n \end{cases} \tag{7}$$

where $\hat{e}$ is median of sequences e.

Step 6: The perceptual hashing digital distance and match. With regard to two speech clips α and β, the hashing digital distance $D(:,:)$ is computed as follows:

$$D(H(\alpha), H(\beta)) = \sum_m |ph1(m) - ph2(m)|; \quad m = 1, 2, \cdots, n \tag{8}$$

According to hashing digital distance $D(:,:)$ and hashing sequence $H(:)$ hypothesis test for matching is as follows:

K_1: Two audio clips α and β are from the perceptual content same clips, if:

$$D(ph(\alpha), ph(\beta)) \leq \tau \tag{9}$$

K_2: Two audio clips α and β are from the perceptual content different clips, if:

$$D(ph(\alpha), ph(\beta)) > \tau \tag{10}$$

where τ is match threshold, through setting match threshold to judge speech signal perceptual content whether or not. Therefore, realize speech signal perceptual authentication.

4. Experimental Results and Analysis.

4.1. **Experimental environment.** In the experiment, use the TIMIT speech library and studio recorded speech. The speech library is a total of 1,280 speech clips that consist of 600 English speech clips and 680 Chinese speech clips. Every speech clip is converted to a general format with the same length 4 s. The speech parameters are as follows: the sampling rate 16 kHz, the bit rate is 256 kbps, the channels is mono, sampling precision is 16 bits, format is WAV, frame length is 20 ms, and frame shifts is 10 ms. The experimental hardware platform is Inter Core i3, 2450M, 2 G, 2.27 GHz, and software environment is the MATLAB R2012b under Win 7.

In speech signal analysis, binary data of simple structure, small volume, easy to statistical analysis and operation, therefore, this paper takes the binary perceptual hashing sequence to calculate the experiment. The perceptual hashing distance evaluation parameter is bit error rate (BER), BER is pointed out error bits percentage in the total number of bits, and the calculation formula is as follows:

$$BER = \frac{\sum_{i=1}^{N} (ph_{new} \oplus ph_{orig})}{N} \tag{11}$$

where N is number of speech, the ph_{new} and ph_{orig} are perceptual hashing values.

The 1280 clips speech in the library is for the following operations, and the computer simulations get the BER of the various operations.

Increase volume: The original speech volume increased by 50 %.

Decrease volume: The original speech volume decreased by 50 %.

Resampling: Speech signal sampling frequency reduced to 8 kHz, and up to 16 kHz.

Echo: Stack attenuation was 60 %, the time delay for 300 ms, the echoes of the initial intensity are 20 % and 10 % respectively.

Narrowband noise: The speech signal with the center frequency distribution in $0 \sim 4$ kHz narrowband Gaussian noise.

Cut: Randomly cut off more than one place or greater than 0.3 s of speech.

FIR filter: The speech signal filtering with 12-order FIR low-pass filter for 3.4 kHz.

Butterworth filter: The speech signal filtering with 12-order Butterworth low-pass filter for 3.4 kHz.

MP3 (32kbps): The speech signal coding for MP3 and decoding for WAV with MP3 rate of 32 kbps.

MP3 (128kbps): The speech signal coding for MP3 and decoding for WAV with MP3 rate of 128 kbps.

4.2. **Discrimination analysis.** This paper totally gets 818,560 BER data by conducted pairwise comparison between perceptual hashing values from 1,280 different speech clips. In this paper, the perceptual hashing value is binary bits sequences and subject to independent identically distributed ideally, which can be denoted $ph = h_1, h_2, \cdots, h_N$. The distance between perceptual hashing value in this paper can be recognized as Hamming distance approximately obeying normal distribution ($\mu = Np, \sigma = \sqrt{Np(1-p)}$) from De Moivre-Laplace central limit theorem [2]. In this paper, the algorithm considers the BER which obeys normal distribution ($\mu = p, \sigma = \sqrt{p(1-p)/N}$), where N is the length of the perceptual hashing sequences. The distance value $x = D(:,:)$ obtained by (8) has a

normal distribution if N is large enough. Then, the probability density function of x can be written as follows:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \frac{-(x-\mu)^2}{2\sigma^2} \tag{12}$$

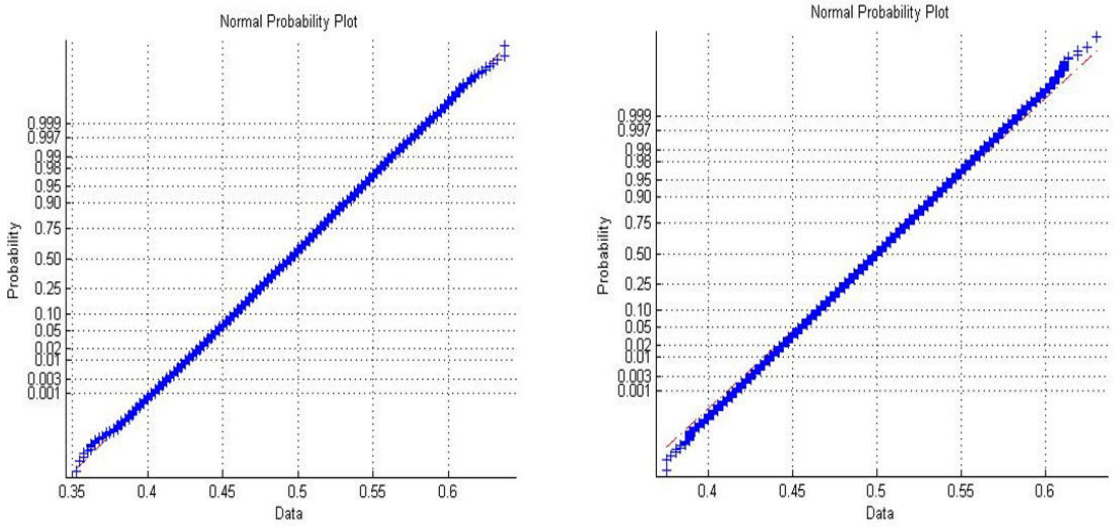
where μ and σ are the expected value and the standard deviation of the x, then, for any $\tau > 0$, the false accept rate (FAR) is denoted as follows:

$$FAR(\tau) = \int_{-\infty}^{\tau} f(x\,|\mu,\sigma)dx = \int_{-\infty}^{\tau} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} dx \tag{13}$$

Accordingly, the false reject rate (FRR) is denoted as follows:

$$FRR(\tau) = 1 - \int_{-\infty}^{\tau} f(x\,|\mu,\sigma)dx = 1 - \int_{-\infty}^{\tau} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} dx \tag{14}$$

The oblique line expected and the curve basically obeying normal distribution whose expected value $\mu = 0.4947$ and standard deviation $\sigma = 0.0292$, and the normal probability plot of the above BER data is shown in Figure 2.



(a) The proposed algorithm (b) The LPC algorithm

FIGURE 2. Normal distribution curves of the algorithm

In this paper, the perceptual hashing sequences length is 400, namely $N = 400$, according to the theoretical value, obtained by parameter $\mu = 0.5$, $\sigma = 0.0250$. The parameters of the experimental measuring $\mu = 0.4947$, $\sigma = 0.0292$, which are very close to the theoretical value calculated parameter values. So the algorithm has a good discrimination performance.

According to the above analysis it shows that the bit error rate of different content of speech perceptual hashing value basically obeys normal distribution, the probability distribution of parameters for the $\mu = 0.4947$, $\sigma = 0.0292$. The LPC algorithm [16] probability distribution parameters for $\mu = 0.5$, $\sigma = 0.0272$, according to the (13) get $FARs$ of the algorithm.

The FAR of the proposed algorithm and LPC algorithm is shown in Table 1.

As can be seen from Table 1, the FAR values are very small, when $\tau = 0.3$, the FAR = 1.3874e-11. It means that when $\tau = 0.3$, there is approximately one which is wrong in judging 10^{11} speech clips. It so far can satisfy the requirements from people to speech perceptual authentication.

TABLE 1. The FAR of the proposed algorithm and LPC algorithm

Threshold	Proposed algorithm	LPC algorithm
τ	FAR	FAR
0.35	3.7514e-07	1.7905e-08
0.30	1.3874e-11	1.0104e-13
0.25	2.9244e-17	2.0677e-20
0.20	3.4414e-24	1.5045e-28
0.15	2.2347e-32	3.8486e-38

TABLE 2. The BER of proposed algorithm

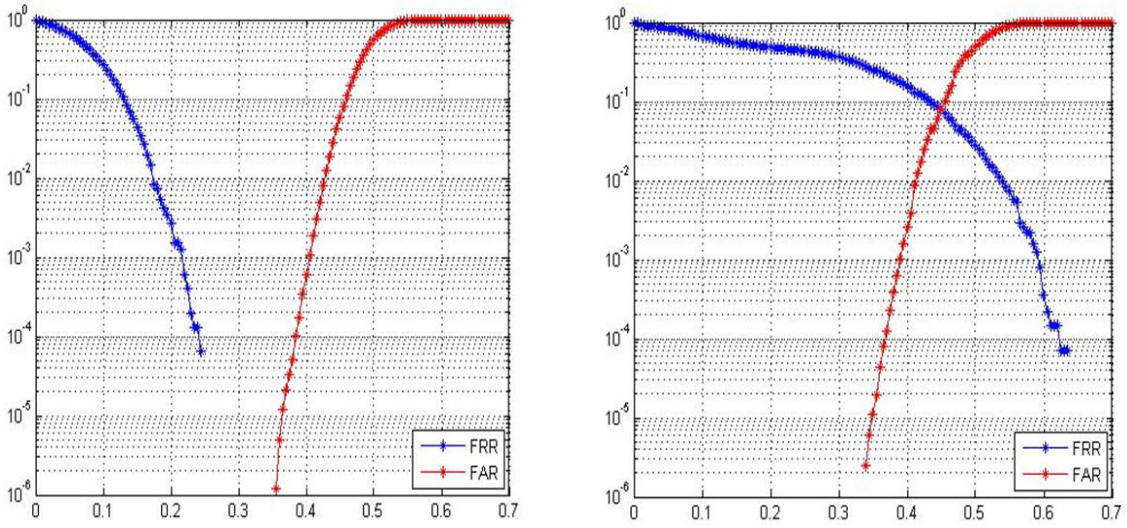
Parameters	Average BER	Standard deviation	Width BER
Increase volume	0.0495	0.0199	0.1175
Decrease volume	0.0455	0.0150	0.1075
Echo	0.1294	0.0317	0.2450
Resampling	0.0053	0.0046	0.0300
Narrowband noise	0.0702	0.0274	0.0187
Cut	0.0642	0.0109	0.0975
FIR filter	0.1023	0.0270	0.2025
Butterworth filter	0.1021	0.0284	0.2150
MP3 (32 kbps)	0.1003	0.0269	0.1791
MP3 (128 kbps)	0.0372	0.0161	0.0945

4.3. **Robustness analysis.** According to the operations in Section 4.1 to get BER as shown in Table 2, drawing the $FRR - FAR$ curves as shown in Figure 3.

As can be seen from Table 2, the above several attacks average BER are below decision threshold $\tau = 0.25$. This paper algorithm is based on the channel model for the human ear, due to the fact that the increased volume and decreased volume do not change channel model of human ear, increased volume and decreased volume do not have a big change that the perceptual feature value obtained by this paper algorithm, accounting for adjust volume does not have a big change of average BER. Due to the fact that MP3 compression coding is based on the human ear psychoacoustics model coding method, and does not change the human ear channel model, the BER has less effect of this paper algorithm of the MP3 compression operation. The algorithm BER has little effect for speech clips by resampling operation in this paper. Also, this paper algorithm has a good robustness to resampling. In all of the above operations, the maximum BER is no more than 0.25; hence the proposed algorithm has a good robustness.

The LPC algorithm $FRR - FAR$ curves are as shown in Figure 3(b). The perceptual hashing value extraction is from the content of the same speech, in which BER is below threshold $\tau = 0.25$ in Figure 3(a). Experimental results show that in the curves of FRR and FAR no intersection intersects, FRR curve has obvious convergence, and has a relatively broad decision interval, when the decision threshold between 0.25 to 0.35 it can accurately authenticate from the same speech clips and different clips of speech and can be authenticated by content preserving operation and content malicious attacks of speech. The proposed algorithm also has good discrimination and robustness.

As can be seen from Table 3, the proposed algorithm average BER is far less than LPC algorithm of adding white Gaussian noise and low-pass filter, and average BER is much



(a) The proposed algorithm (b) The LPC algorithm

FIGURE 3. The $FRR - FAR$ curve of different algorithms

TABLE 3. The average BER of different algorithms

Operating means	The proposed algorithm	The LPC algorithm
Parameters	Average BER	Average BER
Increase volume	0.0495	0.1042
Decrease volume	0.0455	0.0457
Echo	0.1294	0.2418
Resampling	0.0053	0.0512
Narrowband noise	0.0702	0.3818
Cut	0.0642	0.0699
FIR filter	0.1023	0.4129
Butterworth filter	0.1021	0.3939
MP3 (32 kbps)	0.1003	0.2447
MP3 (128 kbps)	0.0372	0.1741

smaller than LPC algorithm for a variety of malicious attacks. So the proposed algorithm has a good robustness of the content preserving operation in this paper.

Figure 4 shows the average BER comparison graphs with the proposed algorithm and LPC algorithm.

The robustness strength of the speech perceptual authentication algorithm, in addition to related to perceptual feature extraction, also largely depends on that determined by the threshold setting. Authentication pass rates that under different matching threshold of different operating are as shown in Table 4.

As can be seen from Table 4, when the matching threshold is set at between 0.25 and 0.35, authentication pass rates reached percent for all operations; when the matching threshold $\tau = 0.25$, in addition to the echo all authentication pass rates also reached percent; when the matching threshold is set at between 0.12 and 0.35, authentication pass rates reached percent for resampling, adjust volume, cut and MP3 compression (128 kbps) operations. The proposed algorithm has good robustness with the operations of the low-pass filter, noise and echo. Therefore, the proposed algorithm has a good robustness.

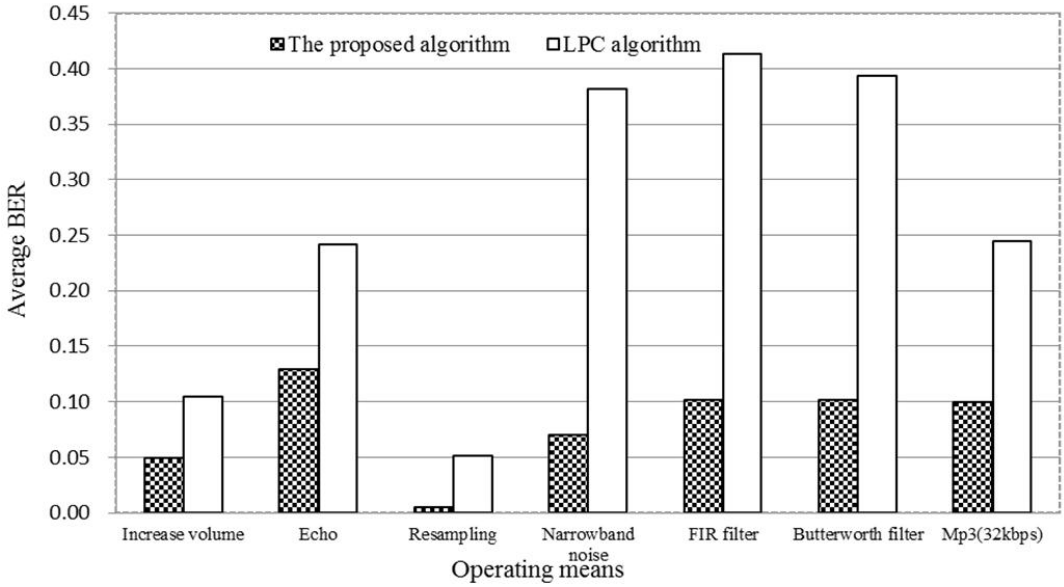


FIGURE 4. The average BER comparison graphs

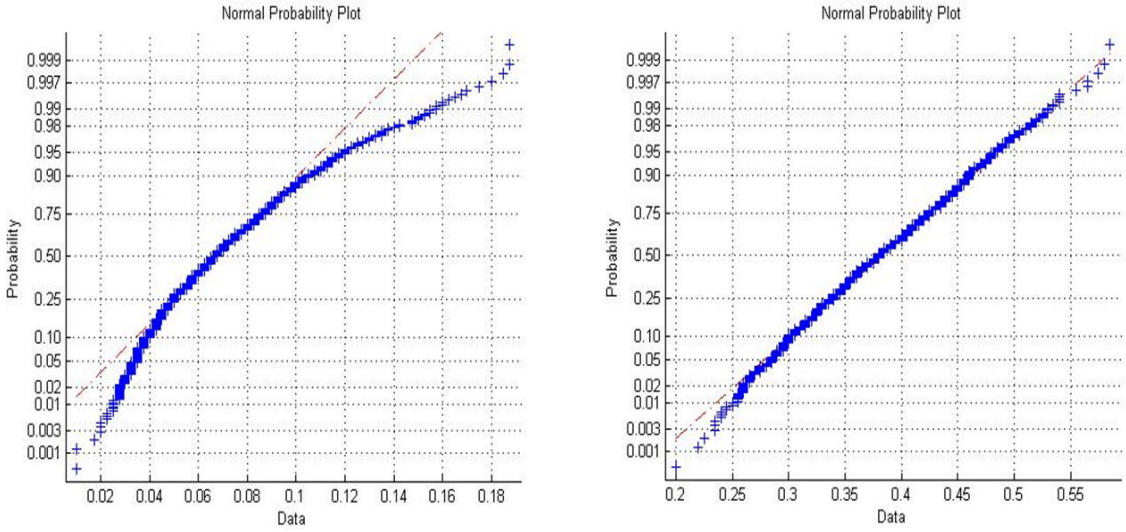
TABLE 4. Authentication pass rate of proposed algorithm (%)

Threshold	0.35	0.25	0.22	0.18	0.15	0.12
Increase volume	100.00	100.00	100.00	100.00	100.00	100.00
Decrease volume	100.00	100.00	100.00	100.00	100.00	100.00
Echo	100.00	100.00	99.30	93.20	73.52	35.94
Resampling	100.00	100.00	100.00	100.00	100.00	100.00
Narrowband noise	100.00	100.00	100.00	99.69	98.44	94.77
Cut	100.00	100.00	100.00	100.00	100.00	100.00
FIR filter	100.00	100.00	100.00	99.53	93.59	75.16
Butterworth filter	100.00	100.00	100.00	99.22	92.50	73.36
MP3(32 kbps)	100.00	100.00	100.00	100.00	96.95	78.59
MP3(128 kbps)	100.00	100.00	100.00	100.00	100.00	100.00

The algorithms normal distribution curves in adding white Gaussian noise is shown in Figure 5.

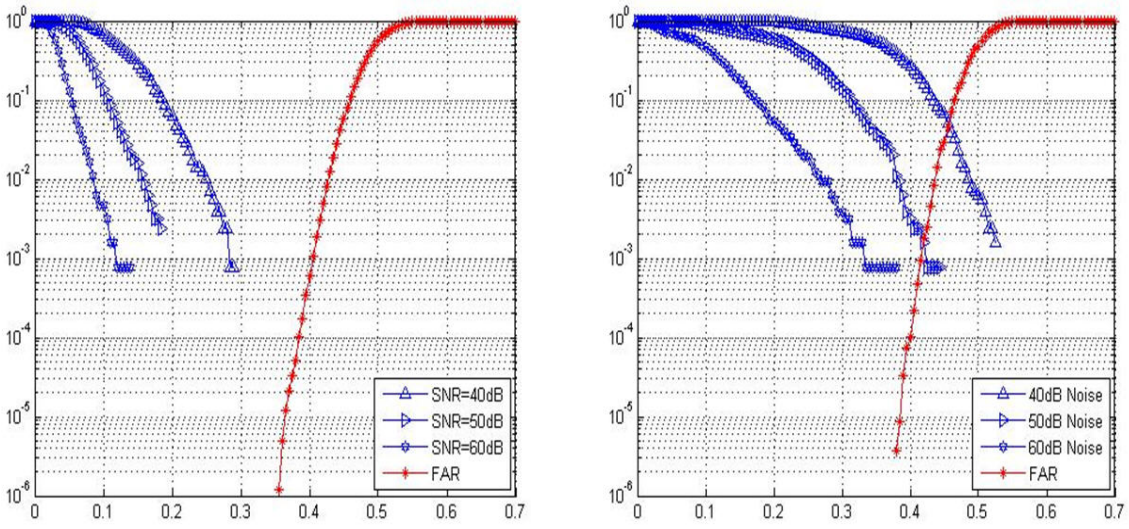
As can be seen from Figure 5(a) obviously, the proposed algorithm authentication normal distribution curves are mainly distributed in between 0.03 and 0.17 under adding white Gaussian noise, below decision threshold $\tau = 0.25$, while the LPC algorithm authentication normal distribution curves are mainly distributed in between 0.25 and 0.55 under adding white Gaussian noise from Figure 5(b). Therefore, the proposed algorithm could be good resistance to white Gaussian noise on the influence of the authentication. The proposed algorithm $FRR-FAR$ curves under adding different white Gaussian noise are shown in Figure 6(a) clearly. The proposed algorithm has a good robustness under adding different white Gaussian noise operations. Hence, the proposed algorithm robustness is obviously improved to resist white Gaussian noise.

Existing authentication algorithm for speech perceptual, the robustness of the low-pass filter operation generally poor, the proposed algorithm normal distribution curves under adding low-pass filter are shown in Figure 7.



(a) The proposed algorithm (b) The LPC algorithm

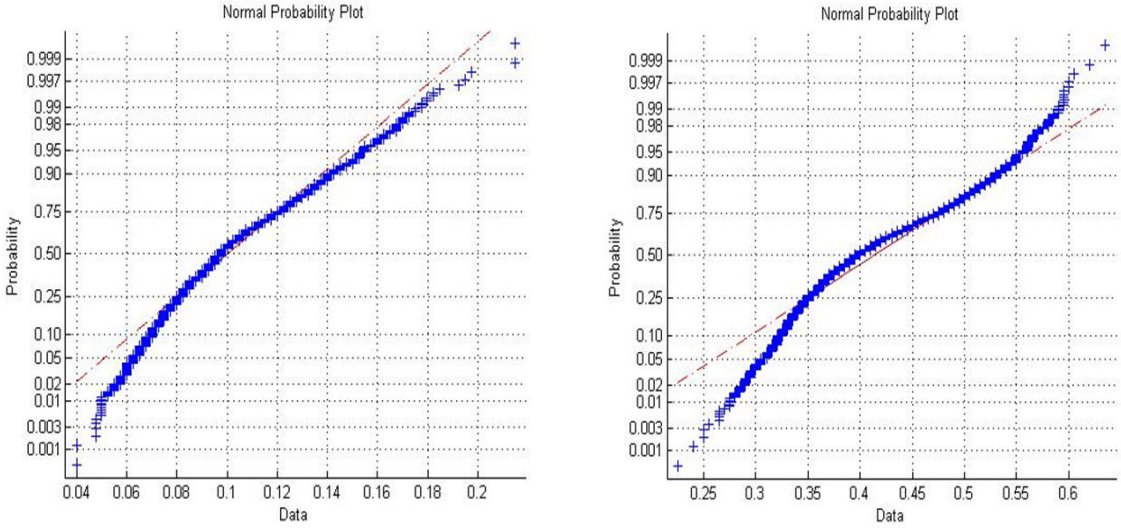
FIGURE 5. Normal distribution curves under adding white Gaussian noise



(a) The proposed algorithm (b) The LPC algorithm

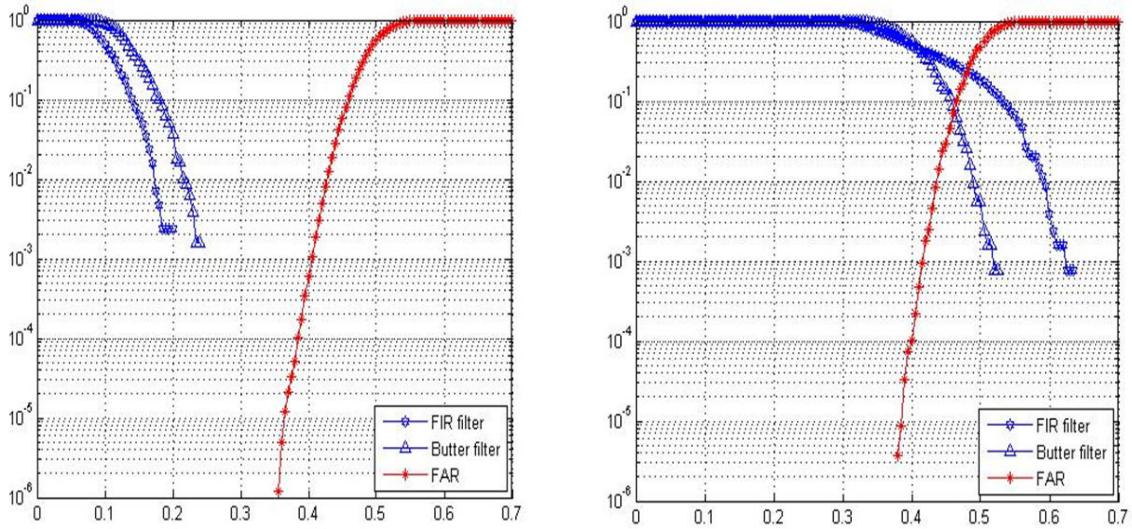
FIGURE 6. The algorithm $FRR - FAR$ curves under adding white Gaussian noise

As can be seen from Figure 7(b) obviously, the LPC algorithm authentication normal distribution curves are mainly distributed in between 0.26 and 0.60 under adding low-pass filter. However, the proposed algorithm authentication normal distribution curves are mainly distributed in between 0.05 and 0.18 under low-pass filter from Figure 7(a), below decision threshold $\tau = 0.25$. Compared with the LPC algorithm, the proposed algorithm BER is smaller in low-pass filter operation, and the robustness is better than LPC algorithm for the low-pass filter operation. The proposed algorithm $FRR - FAR$ curves under different low-pass filters are shown in Figure 8(a) clearly. The proposed algorithm has a good robustness under different low-pass filter operations. Therefore, the proposed algorithm robustness is obviously enhanced to resistance to low-pass filter.



(a) The proposed algorithm (b) The LPC algorithm

FIGURE 7. Normal distribution curves under adding low-pass filter



(a) The proposed algorithm (b) The LPC algorithm

FIGURE 8. The algorithm $FRR - FAR$ curves under adding low-pass filter

From what has been discussed above, compared with the LPC algorithm, the proposed algorithm not only largely improves the robustness under white Gaussian noise and low-pass filter operating, but also has a good robustness to other operations. As a result, the proposed algorithm has a good robustness.

4.4. **Tamper detection and localization.** Mobile terminal real-time speech communication is vulnerable to malicious tampering attack outlaws; in order to achieve reliable security speech content authentication, speech perceptual hashing algorithm should have to be sensitive and have accurate ability of tamper detection. Therefore, the speech perceptual authentication algorithm has certain capacity of tamper detection.

Due to the binary perceptual hashing used in this algorithm, we can determine whether the information has been tampered or not by comparing the perceptual hashing. Assuming that the perceptual hashing of the original speech signal is ph_{orig} and the perceptual

hashing of the tamper speech signal is ph_{tam}, the determination of tamper localization according to the perceptual hashing is shown in the (15) and (16):

$$ph_{orig}(i) = ph_{tam}(i), \quad 1 \leq i \leq l \tag{15}$$

$$ph_{orig}(i) \neq ph_{tam}(i), \quad 1 \leq i \leq l \tag{16}$$

where, l is the length of the perceptual hashing values.

If the perceptual hashing ph_{orig} and ph_{tam} meet the (15), it is not tampered, if the perceptual hashing ph_{orig} and ph_{tam} satisfy type (16), the ph_{tam} is tampered, and the tamper location is $ph_{tam}(i)$, the tamper location can be determined according to (16).

To test proposed algorithm's sensitivity against content manipulations, we replace the parts of original speech clip with another different speech data. In the experiment we randomly select a speech clip, and then randomly replace three places greater than 10 frames. The tamper detection result is shown in Figure 9. In this algorithm by comparing the original speech and tampered speech of hashing value to determine whether the little speech clip is tampered, it can achieve single or several speech clips tamper detection and localization. The tampered areas are red elliptic curves inclusion areas in Figure 9.

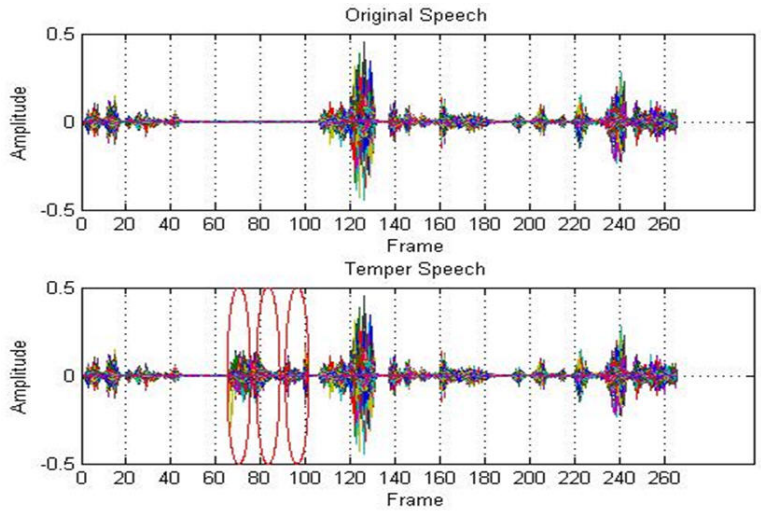


FIGURE 9. Several local tamper localization schematic diagrams

As a result, this algorithm has certain tamper detection ability, but the Ref. [16] without some sort of tamper detection and localization capabilities, the proposed algorithm can achieve single or several local tamper detection and localization.

4.5. **Efficiency analysis.** In this paper, the hash algorithm must meet the requirements of small authentication data and high computational efficiency. In terms of perceptual hashing, the proposed algorithm operational methods are simple accumulative operations or modularized DCT, Hilbert transform and Shannon entropy operations.

Randomly extract 100 speech clips from the speech library and statistical algorithm running time. A total of run 100 times and the average run time is shown in Table 5. Compared with the running time of the LPC algorithm, the proposed algorithm improved robustness and discrimination, and the computing speed is not very big loss, the average authentication time $t = 3.4988\text{e-}04$ s. It means that the authentication time is short. It can satisfy the real-time application requirement.

TABLE 5. Run time

Algorithm	Proposed algorithm	LPC algorithm
Operating means	Average run time (s)	Average run time (s)
File length	4 s	4 s
Platform working frequency	2.7 GHz	2.7 GHz
Feature extraction	0.229672	0.151794
Hashing structure	0.008004	0.008515
Total	0.237676	0.160309
Authentication time	0.000349	0.000352

5. **Conclusions.** In order to improve speech perceptual hashing authentication algorithm robustness under low-pass filter and white Gaussian noise of speech content authentication, this paper proposed a time-frequency domain robust speech perceptual hashing authentication algorithm based on Hilbert transform. In this paper algorithm respectively gets speech perceptual feature value in the time domain and frequency domain. So it can better show the inherent characteristics of speech signal. Firstly, obtain short-time energy as the time domain perceptual feature value of each speech frames after pre-processing of speech signal. Then, the each speech frame for Hilbert transform and two dimensional discrete cosine transform constitute a frequency domain feature matrix and evaluate its Shannon entropy as the frequency domain perceptual feature value. This algorithm introduced into the Shannon entropy can reduce the influence of white Gaussian noise for speech feature parameters, and improve the noise robustness, at the same time, reduce the amount of processing data. Experimental results show that when compared with the existing method, the proposed algorithm not only has better robustness and discrimination, but also has good robustness under the white Gaussian noise and low-pass filter attack, and lower computational complexity, small perceptual hashing data rate, short time authentication. It can meet the real-time performance and robustness requirements of the mobile speech communication.

Acknowledgment. This work is partially supported by the National Natural Science Foundation of China (No. 61363078), the Natural Science Foundation of Gansu Province of China (No. 1212RJZA006, No. 1310RJYA004). The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

REFERENCES

[1] J. Gu, *Research on Key Technologies of Speech Perceptual Authentication*, Ph.D. Thesis, University of Science and Technology of China, Hefei, China, 2009.

[2] Y. H. Jiao, *Research on Perceptual Audio Hashing*, Ph.D. Thesis, Harbin Institute of Technology, Harbin, China, 2010.

[3] X. M. Niu and Y. H. Jiao, An overview of perceptual hashing, *Acta Electronica Sinica*, vol.36, no.7, pp.1405-1411, 2008.

[4] H. Özer, B. Sankur, N. Memon and E. Anarim, Perceptual audio hashing functions, *EURASIP Journal on Applied Signal Processing*, vol.2005, no.12, pp.1780-1793, 2005.

[5] P. Lotia and D. M. R. Khan, Significance of complementary spectral features for speaker recognition, *International Journal of Research in Computer and Communication Technology*, vol.2, no.8, pp.579-588, 2013.

[6] Y. B. Huang, Q. Y. Zhang and Z. T. Yuan, Perceptual speech hashing authentication algorithm based on linear prediction analysis, *TELKOMNIKA Indonesian Journal of Electrical Engineering*, vol.12, no.4, pp.3214-3223, 2014.

[7] Y. H. Jiao, L. P. Ji and X. M. Niu, Robust speech hashing for content authentication, *IEEE Signal Processing Letters*, vol.16, no.9, pp.818-821, 2009.

[8] V. Panagiotou and N. Mitianoudis, PCA summarization for audio song identification using Gaussian mixture models, *Proc. of the 18th IEEE International Conf. on Digital Signal Processing (DSP)*, Fira, Greece, pp.1-6, 2013.

[9] N. Chen, H. D. Xiao and W. G. Wan, Audio hash function based on non-negative matrix factorisation of mel-frequency cepstral coefficients, *Information Security, IET*, vol.5, no.1, pp.19-25, 2011.

[10] N. Chen, W. G. Wan and H. D. Xiao, Robust audio hashing based on discrete wavelet transform and non-negative matrix factorization, *Communications, IET*, vol.4, no.14, pp.1722-1731, 2010.

[11] H. Zhao, H. Liu and K. Zhao, Robust speech feature extraction using the Hilbert transform spectrum estimation method, *International Journal of Digital Content Technology and Its Applications*, vol.5, no.12, pp.85-95, 2011.

[12] Q. Y. Zhang, Z. P. Yang and Y. B. Huang, Efficient robust speech authentication algorithm for perceptual hashing based on Hilbert-Huang transform, *Journal of Information and Computational Science*, vol.11, no.18, pp.6537-6547, 2014.

[13] J. Gu, L. Guo, H. Liang and L. Cheng, Effective robust speech authentication algorithm based on perceptual characteristics, *Journal of Chinese Computer Systems*, vol.31, no.7, pp.1461-1465, 2010.

[14] M. Nouri, N. Farhangian, Z. Zeinolabedini and M. Safarinia, Conceptual authentication speech hashing base upon hypotrochoid graph, *Proc. of the 6th IEEE International Conf. on Symposium Telecommunications (IST)*, Tehran, Iran, pp.1136-1141, 2012.

[15] J. J. Deng, W. G. Wan, X. Q. Yu and W. Yang, Audio fingerprinting based on spectral energy structure and NMF, *Proc. of the 13th IEEE International Conf. on Communication Technology (ICCT)*, Jinan, China, pp.1103-1106, 2011.

[16] N. Chen and W. G. Wan, Robust speech hash function, *ETRI Journal*, vol.32, no.2, pp.345-347, 2010.

[17] Z. F. Wang and B. Wang, Research on a speech signal feature extraction method based on the short-time energy-LPCC, *Computer and Digital Engineering*, vol.40, no.11, pp.79-80,127, 2012.

[18] C. Yin and S. Yuan, A novel algorithm for embedding watermarks into audio signal based on DCT, *Proc. of the International Conf. on Information Engineering and Applications (IEA'2012)*, Springer London, pp.683-688, 2013.

[19] P. F. Yu, P. C. Yu and D. Xu, Palmprint authentication based on DCT-based watermarking, *Applied Mechanics and Materials*, vol.457, pp.893-898, 2014.

[20] S. G. Sathyanarayana, A. Gargava and S. M. Venkatesan, Parameterized transform domain computation of the Hilbert transform applied to separation of channels in Doppler spectra, *Proc. of the 3rd IEEE International Advance Computing Conference (IACC)*, Ghaziabad, India, pp.1189-1194, 2013.

[21] H. W. Liu, *A Study on Feature Selection Algorithms Using Information Entropy*, Ph.D. Thesis, Jilin University, Changchun, China, 2010.

[22] H. Misra, S. Ikbal, H. Bourlard and H. Hermansky, Spectral entropy based feature for robust ASR, *Proc. of the IEEE International Conf. on Acoustics, Speech, and Signal Processing (ICASSP'04)*, Quebec, Canada, vol.1, pp.193-196, 2004.