*Wireless Sensor Network*, 2009, 1, 489-495
doi:10.4236/wsn.2009.15059 Published Online December 2009 (http://www.scirp.org/journal/wsn).

# Robust Speech Endpoint Detection in Airplane Cockpit Voice Background

**Hongbing CHENG[1], Ming LEI[2], Guorong HUANG[1], Yan XIA[3]**
[1]*College of Engineering, Air Force Engineering University, Xi'an, China*
[2]*People's Liberation Army* 95340 *Unit, Tianyang, China*
[3]*Air Force Equipment Research Academy, Beijing, China*
*Email*: *newcheng*2008@*yahoo.com.cn, newcheng*2008@163.*com*
*Received July* 5, 2009; *revised June* 7, 2009; *accepted June* 24, 2009

## Abstract

A method of robust speech endpoint detection in airplane cockpit voice background is presented. Based on the analysis of background noise character, a complex Laplacian distribution model directly aiming at noisy speech is established. Then the likelihood ratio test based on binary hypothesis test is carried out. The decision criterion of conventional maximum a posterior incorporating the inter-frame correlation leads to two separate thresholds. Speech endpoint detection decision is finally made depend on the previous frame and the observed spectrum, and the speech endpoint is searched based on the decision. Compared with the typical algorithms, the proposed method operates robust in the airplane cockpit voice background.

**Keywords:** Complex Laplacian Model, Maximum A Posterior Criterion, Likelihood Ratio Test, Speech Endpoint Detection, Airplane Cockpit Voice

## 1. Introduction

The information recorded by airplane cockpit voice recorder is called cockpit voice for short. Cockpit voice background is non-human voice in cockpit voice. It will take significant effect to pick-up voice information of cockpit voice in understanding the station of pilot, investigating the fly accident and finding out causes of accident. Speech endpoint detection is the base of speech tone, and its purpose is to distinguish speech segment and non-speech segment in speech signal [1]. In the airplane communication system, voice background has many characteristics: excessive kinds, complex, non-calm, transient and broad frequency. It makes up of engine noise, air current voice when it is flying, activity voice of manipulated component, diversified switch voice, alarm voice and so on. Especially prophase of airplane wrecking, noise background energy is very strong. The signal-to-noise falls obviously [2]. How to distinguish speech signal and noise signal in cockpit voice background is still a difficulty. Many researchers put forward various algorithms, such as based on entropy [3–5], cepstral feature [6–7], higher-order statistics [8], signal recursion analysis [9] etc., which are not ideal in the circumstance.

Recently years, speech endpoint detection based on statistical model get effective evolvement [10–11], especially the method based on Gaussian mixture model (GMM) [12], which establishes models of pure speech and noise respectively, and makes use of likelihood ratio test (LRT) and maximum probability criterion to judge the station of current frame, and exhibits preferable veracity. Because cockpit voice background has traits of abnormality and complexity, and has no prior information, it is impossible to establish statistical model of noise. Goodness-of-test (GOF) in literature [13] checkout that complex Laplacian model is better than traditional Gaussian model in any noise environment.

This paper imports complex Laplacian distribute model to describe the whole speech which include noise. Aiming at the defect that traditional statistical model analysis every frame signal station distribution absolutely, it thought about interframe relativity sufficiently. Then, it gained two kinds of thresholds of speech station and non-speech station respectively. In the judge criterion, it will adjust threshold automatically depending on previous frame and the observed spectrum to judge the appear or non-appear speech station. So, it achieved cockpit voice background robust speech endpoint detection.

## 2. Speech Endpoint Detection Based on GMM and LRT

Recently years, speech endpoint detection based on GMM [12] gets effective evolvement [14], which establishes

 

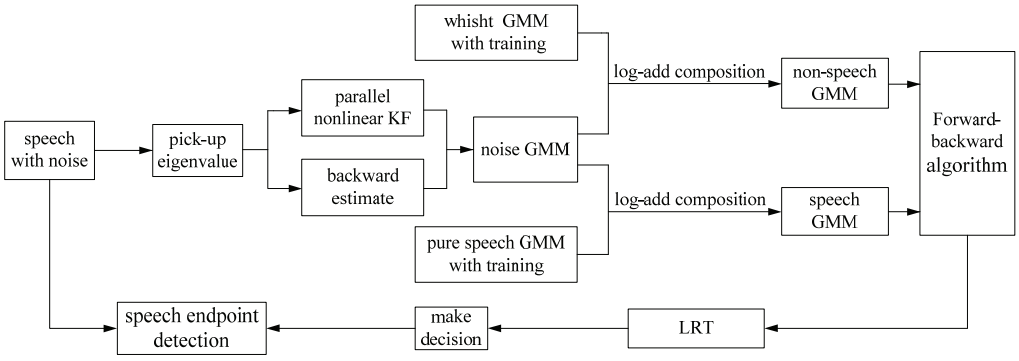


**Figure 1. The speech endpoint detection algorithm flow chart based on GMM and LRT.**

models of pure speech and noise respectively, and makes use of LRT and maximum probability criterion to judge the station of current frame, and exhibits preferable veracity. The algorithm flow chart based on GMM and LRT is showed in Figure 1.

## 2.1. Mathematical Describing of Statistical Models

Hidden Markov models (HMM), as a statistical model of speech signal, can describe the produce process of speech signal accurately. The method of speech endpoint detection based on statical models makes use of LRT to differentiate the speech frame and non-speech frame. Figure 2 shows the analysis platform of speech endpoint detection based on speech or non-speech transfer model [10] of every station.

where,

$H_0$: non-speech station in cockpit voice;

$H_1$: speech station in cockpit voice;

$a_{i,j}$: transfer probability from i to j,

$a_{i,j} = p(q_t = H_j \mid q_{t-1} = H_i)$, i,j=0 or 1;

$b_j(\boldsymbol{O}_t)$: the probability when the output of t frame cockpit voice is j station, $b_j(\boldsymbol{O}_t) = p(\boldsymbol{O}_t \mid q_t = H_j)$;

$\boldsymbol{O}_t$: the L dimension station vector of the t short time amplitude.

The way of distinguish speech frame and non-speech is to estimate the station $q_t$ of t frame short time amplitude on the condition of $\boldsymbol{O}_{0:t} = \{\boldsymbol{O}_0, \cdots, \boldsymbol{O}_t\}$. The compute formula of conditional probability density $p(q_t \mid \boldsymbol{O}_{0:t})$ is:

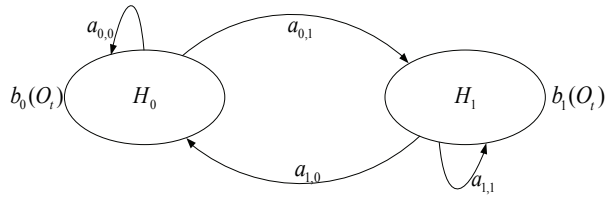


**Figure 2. Speech/non-speech transfer model.**

$$p(q_t \mid \boldsymbol{O}_{0:t}) = p(\boldsymbol{O}_{0:t}, q_t) / p(\boldsymbol{O}_{0:t}) \propto p(\boldsymbol{O}_{0:t}, q_t) \quad (1)$$

Applying one rank Markov chain recursion formula, the combine probability $p(\boldsymbol{O}_{0:t}, q_t)$ of Formula (1) can be showed as:

$$p(\boldsymbol{O}_{0:t}, q_t) = \sum_{q_{t-1}} p(q_t \mid q_{t-1}) p(\boldsymbol{O}_t \mid q_t) p(\boldsymbol{O}_{0:t-1}, q_{t-1}) \quad (2)$$

$p(\boldsymbol{O}_{0:t}, q_t)$ usually called as forward probability $a_{j,t}$, combining $a_{i,j}$ with $b_j(\boldsymbol{O}_t)$:

$$\alpha_{j,t} = a_{0,j} b_j(\boldsymbol{O}_t) \alpha_{0,t-1} + a_{1,j} b_j(\boldsymbol{O}_t) \alpha_{1,t-1} \quad (3)$$

Finally, we can get station $q_t$ through likelihood ratio threshold $R_t = \alpha_{1,t} / \alpha_{0,t}$:

$$q_t = \begin{cases} H_0 & R_t < Threshold \\ H_1 & R_t \geq Threshold \end{cases} \quad (4)$$

For example, if we can ascertain observed that signal $q_t$ is in station $H_1$, comparatively, $q_t$ is speech frame.

## 2.2. The Computation of Probability Density Function Based on GMM

In Formula (3), The computation of $b_j(O_t)$ take significant effect in the precision of endpoint detection. It is more flexible and more applicable to use the method based on GMM of log-mail spectrum than to use the method based on prior and posterior signal-to-noise, so that the precision of estimate of $b_j(O_t)$ will be higher.

$$b_j(\boldsymbol{O}_t) = \sum_{k=1}^{k} \omega_{j,k} \prod_{l=0}^{L-1} \frac{1}{\sqrt{2\pi}\sigma_{j,k,l}} \exp\left\{ -\frac{(O_{t,l} - \mu_{j,k,l})^2}{2\sigma_{j,k,l}^2} \right\} \quad (5)$$

where, $\omega_{j,k}$ is the k mixture weight of gauss distribution of GMM; $O_{t,l}$ is the Lth element of $\boldsymbol{O}_t$; $\mu_{j,k,l}$ is the average of $O_{t,l}$; $\sigma_{j,k,l}^2$ is the variance of $O_{t,l}$. In this method, if we know the average vector of whisht speech GMM, pure speech GMM and noise, we can figure

 

out real time noise GMM and GMM with noise through log-add composition [15] (LAC), so we can gain $b_j(\boldsymbol{O}_t)$.

LAC showed as:

$$\mu_{j,k,l} = \mu_{S,j,k,l} + \log(1+\exp(\mu_{N,l} - \mu_{S,j,k,l})) \quad (6)$$

where $\mu_{S,j,k,l}$ is the average of whisht (j=0) or speech (j=1) GMM in log-mail spectrum, $\mu_{N,l}$ is the average of noise.

In the method, we can establish whisht and pure speech GMM by training pure speech. The average of noise ( $\mu_{N,l}$ ) can be estimated one by one frame by using parallel nonlinear KF. The noise GMM and GMM with noise will update timely with $\mu_{N,l}$.

The traditional likelihood estimation is gained by forward estimating with present and past parameter. The value of t+1,…,T is still the important factor of time sequence estimate. Processing likelihood estimate with the future frame is backward estimate. The definition of backward estimate is:

$$p(\boldsymbol{O}_{0:T}, q_t) = p(\boldsymbol{O}_{0:t}, q_t)\, p(\boldsymbol{O}_{t+1:T} \mid q_t) \quad (7)$$

Similar with Formula (2), conditional probability is showed as:

$$p(\boldsymbol{O}_{t+1:T} \mid q_t) = \sum_{q_{t+1}} p(q_{t+1} \mid q_t)\, p(\boldsymbol{O}_{t+1} \mid q_{t+1})\, p(\boldsymbol{O}_{t+2:T} \mid q_{t+1}) \quad (8)$$

$p(\boldsymbol{O}_{t+1:T} \mid q_t)$ has usually called forward probability $\beta_{j,t}$, combining $a_{i,j}$ with $b_j(\boldsymbol{O}_t)$:

$$\beta_{j,t} = a_{i,0} b_0(\boldsymbol{O}_{t+1}) \beta_{0,t+1} + a_{i,1} b_1(\boldsymbol{O}_{t+1}) \beta_{1,t+1} \quad (9)$$

Usually, backward estimate begin from terminal of tested signal, but in the test of endpoint, the terminal is unknown. So we introduce back modularize estimation. It is begin from T=t+b, where b is a constant. When b=0, backward estimate equal to does not process.

We can conclude from the definition of the Forward-Backward (F-B) algorithm that: $p(\boldsymbol{O}_{0:t}, q_t = H_t) = \alpha_{j,t}\beta_{j,t}$. We can gain likelihood ratio $R_t$ by applying likelihood ratio test.

$$R_t = \frac{p(O_{0:T}, q_t = H_1)}{p(O_{0:T}, q_t = H_0)} = \frac{\alpha_{1,t}\beta_{1,t}}{\alpha_{0,t}\beta_{0,t}} \quad (10)$$

Finally, substituting $R_t$ in Formula (4), we get the station value $q_t$ of speech endpoint detection.

## 3. The Establishing of Complex Laplacian Distribution Model

Speech endpoint detection is processed one by one frame. Every frame includes M sampling. In generally, speech signal is thought as windless signal in short period (10~30ms). We can suppose that speech signal with noise is statistical irrelated complex Laplacian random course. We denote coefficient vector of discrete fourier transform (DFT) of M dimension noise speech with $\mathbf{X}(t)$:

$$\mathbf{X}(t)=[X_1(t), X_2(t), \cdots X_k(t) \cdots X_M(t)]^T$$

If $X_{\mathrm{k(R)}}$ and $X_{\mathrm{k(I)}}$ denote real part and imaginary part of $X_{\mathrm{k}}$ respectively, the probability density distribution of $X_{\mathrm{k(R)}}$ and $X_{\mathrm{k(I)}}$, according to the Laplacian probability distribution, can be written as:

$$p(X_{\mathrm{k(R)}}) = \frac{1}{\sigma_{\mathrm{x}}} \exp\{-\frac{2\left|X_{\mathrm{k(R)}}\right|}{\sigma_{\mathrm{x}}}\} \quad (11)$$

$$p(X_{\mathrm{k(I)}}) = \frac{1}{\sigma_{\mathrm{x}}} \exp\{-\frac{2\left|X_{\mathrm{k(I)}}\right|}{\sigma_{\mathrm{x}}}\} \quad (12)$$

where, $\sigma_{\mathrm{x}}^2$ is the variance of $X_{\mathrm{k}}$. If the real part and imaginary part of $X_{\mathrm{k}}$ are uncorrelated, the distribution density of $X_{\mathrm{k}}$ can be written as:

$$\begin{aligned} p(X_{\mathrm{k}}) &= p(X_{\mathrm{k(R)}}) \cdot p(X_{\mathrm{k(I)}}) \\ &= \frac{1}{\sigma_{\mathrm{x}}^2} \exp\{-\frac{2(\left|X_{\mathrm{k(R)}}\right| + \left|X_{\mathrm{k(I)}}\right|)}{\sigma_{\mathrm{x}}}\} \end{aligned} \quad (13)$$

## 4. The Likelihood Ratio Test Based on Hypothesis Test

Speech endpoint detection can be regarded as a binary hypothesis issue:

$$H_0 : \textit{speech donot appear} \quad \mathbf{X}(t) = \mathbf{N}(t)$$
$$H_1 : \textit{speech appear} \quad \mathbf{X}(t) = \mathbf{N}(t) + \mathbf{S}(t)$$

where, $H_0$ denote the situation of speech not appearing, $H_l$ denote the situation of speech appearing, $N(t)$ and $S(t)$ denote DFT coefficient vector of background noise and pure speech respectively. The conditional probability density of noise under the situation of $H$ and $H_l$ can be written as:

$$p(X_{\mathrm{k}} \left| \mathrm{H_n}=\mathrm{H_0}\right.) = \frac{1}{\lambda_{\mathrm{n,k}}} \exp\{-\frac{2(\left|X_{\mathrm{k(R)}}\right| + \left|X_{\mathrm{k(I)}}\right|)}{\sqrt{\lambda_{\mathrm{n,k}}}}\} \quad (14)$$

$$p(X_{\mathrm{k}} \left| \mathrm{H_n}=\mathrm{H_1}\right.) = \frac{1}{\lambda_{\mathrm{n,k}} + \lambda_{\mathrm{s,k}}} \exp\{-\frac{2(\left|X_{\mathrm{k(R)}}\right| + \left|X_{\mathrm{k(I)}}\right|)}{\sqrt{\lambda_{\mathrm{n,k}} + \lambda_{\mathrm{s,k}}}}\} \quad (15)$$

We can receive likelihood test of hypothesis test by Formulas (14) and (15). Likelihood ratio $\Lambda_{\mathrm{k}}$ of the kth frequency band can be denoted as:

 

$$\Lambda_{\mathrm{k}} \equiv \frac{p(X_{\mathrm{k}} \left| \mathrm{H_n}=\mathrm{H_1} \right.)}{p(X_{\mathrm{k}} \left| \mathrm{H_n}=\mathrm{H_0} \right.)} \tag{16}$$

Because the signal samples $\mathbf{X}_{\mathrm{k}}(k=1,2\cdots M)$ are uncorrelated and have the same distribution, the likelihood ratio of M dimension observed vector of two hypothesis is:

$$\begin{aligned}\Lambda &= \frac{p(\mathbf{X}\left|H_n=H_1\right.)}{P(\mathbf{X}\left|H_n=H_0\right.)} = \prod_{k=0}^{M-1}\frac{p(X_k\left|H_n=H_1\right.)}{p(X_k\left|H_n=H_0\right.)} \\ &= \prod_{k=1}^{M}\frac{1}{1+\xi_k}e^{\left\{2\left(\left|X_{k(R)}\right|+\left|X_{k(I)}\right|\right)(\left|X_k\right|-\sqrt{\lambda_{n,k}})/\sqrt{\left|X_k\right|\lambda_{n,k}}\right\}}\end{aligned} \tag{17}$$

where, $\xi_k$ is the forward signal-to-noise, define as $\xi_k = \frac{\lambda_{s,k}}{\lambda_{n,k}}$, we assume that all the frequency vectors are uncorrelated.

We can know from Formula (17) that $\lambda_{\mathrm{n,k}}$ and $\xi_k$ have great influence on the veracity of likelihood ratio test. The estimate of $\lambda_{\mathrm{n,k}}$ of traditional speech endpoint detection updates in speech intermission time. The power spectrum changes when speech appears in cockpit voice background, where the impulse noise does not appear in other time. So, the estimation of noise power spectrum should be updated really both when speech appear and when speech do not appear. We adopt the method of long time power spectrum smooth to compute $\lambda_{\mathrm{n,k}}$ [16]. From [16] we know that the estimation of the kth fourier transform coefficient variance is:

$$\hat{\lambda}_{n,k}(t+1) = \varsigma_{\lambda_n}\hat{\lambda}_{n,k}(t) + (1-\varsigma_{\lambda_n})E[\left|N_k(t)\right|^2\left|X_k(t)\right.] \tag{18}$$

where, $\hat{\lambda}_{n,k}(t)$ is the estimation of $\lambda_{n,k}(t)$ and $\varsigma_{\lambda_n}$ is the smooth coefficient. Considering the two situation of speech appearing and not appearing, the estimation of the noise power spectrum of current frame is:

$$\begin{aligned}&E[\left|N_k(t)\right|^2\left|X_k(t)\right.] \\ &= E[\left|N_k(t)\right|^2\left|X_k(t), H_n=H_0\right.]P(H_n=H_0\left|X_k(t)\right.) \\ &\quad + E[\left|N_k(t)\right|^2\left|X_k(t), H_n=H_1\right.]P(H_n=H_1\left|X_k(t)\right.)\end{aligned} \tag{19}$$

where: $E[\left|N_k(t)\right|^2\left|X_k(t), H_n=H_0\right.] = \left|X_k(t)\right|^2$

$$\begin{aligned}&E[\left|N_k(t)\right|^2\left|X_k(t), H_n=H_1\right.] \\ &= (\frac{\hat{\xi}_k(t)}{1+\hat{\xi}_k(t)})\hat{\lambda}_{n,k}(t) + (\frac{1}{1+\hat{\xi}_k(t)})^2\left|X_k(t)\right|^2\end{aligned}$$

The prior signal-to-noise $\xi_k$ can be estimated, following literature [17], as:

$$\hat{\xi}_k(t) = (1-\varsigma_{SNR})\frac{\left|\hat{S}_k\right|^2(t-1)}{\hat{\lambda}_{n,k}(t-1)} + \varsigma_{SNR}P[\hat{\gamma}_k(t)-1] \tag{20}$$

where, $P[x]=\begin{cases}x, & x\ge 0\\ 0, & others\end{cases}$, $\gamma_k = \frac{\left|X_k\right|^2}{\lambda_n}$ is posterior signal-to-noise, $\hat{\gamma}_k(t)$ is it's estimation, $\varsigma_{SNR}$ is the weight of direct judge estimate, $\left|\hat{S}_k\right|^2(t-1)$ is the speech amplitude breadth of pre-frame which has estimated by using MMSE.

We can gain likelihood estimate by substituting (18) and (20) in (7).We can judge whether speech appear or not based on traditional MAP criterion [18].

## 5. The Judge Criterion Based on Conditional MAP

The decision-making of speech endpoint detection based on traditional MAP criterion is:

$$\frac{p(H_n=H_1\left|\mathbf{X}\right.)}{P(H_n=H_0\left|\mathbf{X}\right.)} \underset{H_0}{\overset{H_1}{\gtrless}} 1 \tag{21}$$

where, $H_n$ denote the nth frame right hypothesis. According to Bayesian formula, the criterion of likelihood ratio is:

$$\frac{p(\mathbf{X}\left|H_n=H_1\right.)}{P(\mathbf{X}\left|H_n=H_0\right.)} \underset{H_0}{\overset{H_1}{\gtrless}} \frac{p(H_n=H_0)}{P(H_n=H_1)} \tag{22}$$

However, the speech appear model $H_1$ include speech do not appear model $H_0$. It causes the computing of likelihood ratio partial to $H_1$ [10]. In order to make up the difference, the Formula (22) is adjusted as:

$$\frac{p(\mathbf{X}\left|H_n=H_1\right.)}{P(\mathbf{X}\left|H_n=H_0\right.)} \underset{H_0}{\overset{H_1}{\gtrless}} \alpha\frac{p(H_n=H_0)}{P(H_n=H_1)}, \quad \alpha>1 \tag{23}$$

The speech endpoint detection of interframe has strong relativity. The probability of that speech frame's next frame turns into speech frame is very large. The relativity was validated by FSM [11].

The paper combined the relativity of interframe with MAP criterion. It is different from traditional forward probability $P(H_n\left|\mathbf{X}\right.)$. The present observed value and the decision-making of pre-frame were used for computing forward probability. It was denoted as $P(H_n\left|\mathbf{X},H_{n-1}\right.)$, and the decision-making verification of speech endpoint detection decision-making was adjusted:

$$\frac{p(H_n=H_1\left|\mathbf{X},H_{n-1}=H_i\right.)}{P(H_n=H_0\left|\mathbf{X},H_{n-1}=H_i\right.)} \underset{H_0}{\overset{H_1}{\gtrless}} \alpha \quad i=0,1 \tag{24}$$

where, $\alpha$ is threshold. The estimation of likelihood ratio becomes:

 

$$\frac{p(\mathbf{X}|H_n = H_1, H_{n-1} = H_i)}{P(\mathbf{X}|H_n = H_0, H_{n-1} = H_i)} \underset{H_0}{\overset{H_1}{\gtrless}} \alpha \frac{p(H_n = H_0|H_{n-1} = H_i)}{P(H_n = H_1|H_{n-1} = H_i)}, \quad i = 0,1 \tag{25}$$

In the actual cockpit voice, because of the lack of prior information, distributed parameters, $p(\mathbf{X}|H_n = H_1, H_{n-1} = H_i)$ and $P(\mathbf{X}|H_n = H_0, H_{n-1} = H_i)$, have not been estimated, and the distributed parameters of current frame were decided by the current observed value. So it was predigested as:

$$p(\mathbf{X}|H_n = H_j, H_{n-1} = H_i) = P(\mathbf{X}|H_n = H_j), \quad i = 0,1, \quad j = 0,1. \tag{26}$$

Formula (25) is changed to:

$$\frac{p(\mathbf{X}|H_n = H_1)}{P(\mathbf{X}|H_n = H_0)} \underset{H_0}{\overset{H_1}{\gtrless}} \alpha \frac{p(H_n = H_0|H_{n-1} = H_i)}{P(H_n = H_1|H_{n-1} = H_i)}, \quad i = 0,1 \tag{27}$$

Its form of log is:

$$\log \frac{p(\mathbf{X}|H_n = H_1)}{P(\mathbf{X}|H_n = H_0)} \underset{H_0}{\overset{H_1}{\gtrless}} \log\left[\alpha \frac{p(H_n = H_0|H_{n-1} = H_i)}{P(H_n = H_1|H_{n-1} = H_i)}\right] \triangleq \eta_i, \quad i = 0,1 \tag{28}$$

The Formula (27) or (28) is the judge criterion of speech endpoint detection. $\eta_i$ is the threshold. When preframe is speech frame, $\eta_1$ will be regarded as the threshold of the current frame. When preframe is nonspeech frame, $\eta_0$ will be regarded as the threshold of the current frame. Multiple thresholds can provide more freedom and can enhance the robusticity of speech endpoint detection. Considering the relativity of interframe, parameter distribution has the trait as follows:

$$\frac{p(H_n = H_0|H_{n-1} = H_0)}{P(H_n = H_1|H_{n-1} = H_0)} > \frac{p(H_n = H_0|H_{n-1} = H_1)}{P(H_n = H_1|H_{n-1} = H_1)} \tag{29}$$

It indicates that the probability of nonspeech frame's next frame become nonspeech frame is large. When the preframe is nonspeech frame, $\eta_0$ is larger than $\eta_1$. It is all the same for speech frame.

## 6. Experimentation

In order to test the validity of the paper's algorithm, the cockpit voice background sound of airplane normal station and wrecked station have been picked up respectively, and two teams experimentation of speech endpoint detection based on GMM and the paper's algorithm have been done.

### 6.1. The Establishment of Experimentation

In environment of lab, we record 200 sentences of 6 persons (3 men and 3 women) to form storage of pure speech and training GMM. The test group makes up of other 40 sentences. Because cockpit voice background sound is complex, excessive, so its bandwidth is broad (150Hz-6800Hz), and its signal is not calm and is transient. Different kind airplanes have different cockpit voice background sounds. Its characteristics are different from F16 noise provided by group NOISEX-2. So that, the cockpit voice background sound used in simulation test was recorded in the real environment. Its sample frequency is 16KHz and quantitative change bite is 16 and single channel is format wave. We can get airplane normal station and wrecked station speech with noise group by adjusting breadth of pure speech and adding it to cockpit voice background sound. The extracting of character is showed in Table 1.

When training GMM, the GMM parameter with 25 characteristic vector (12 rank mail cepstral coefficient and its differential coefficient, short time power differential coefficient) was gained by using the expectation-maximization (EM) algorithm. The smooth coefficient $\varsigma_{\lambda_n}$, the weight $\varsigma_{SNR}$ for judging forward signal-to-noise estimation and the known threshold $\eta_i$ based on preframe should be chosen carefully to ensure the robusticity.

### 6.2. The Result of Experiment

We define that $P_d$ is the ratio that the speech frame is detected as the speech frame correctly and $P_f$ is the ratio that the nonspeech frame is detected as the speech frame. The performance of the two algorithms is depicted by the ROC curve which denote the relation of $P_d$ and $P_f$. Figure 3 shows a real example of speech endpoint detection. Its last time is 1s. Figure 4 and Figure 5 show the ROC curve, which is the cockpit voice background speech endpoint detection of airplane normal station and wrecked station, of the two algorithms.

In Figure 3, the broken line of pure speech graph is the manual mark place of speech begin point. When the air-

**Table 1. the condition of character extracting.**

| | |
|---|---|
| Sample frequency | 16kHz |
| Quantitative change bite | 16bite |
| Advance add quantity | $1\text{-}0.97z^{-1}$ |
| Length of frame | 20ms |
| Moving of frame | 10ms |
| Function of window | Hamming |

 

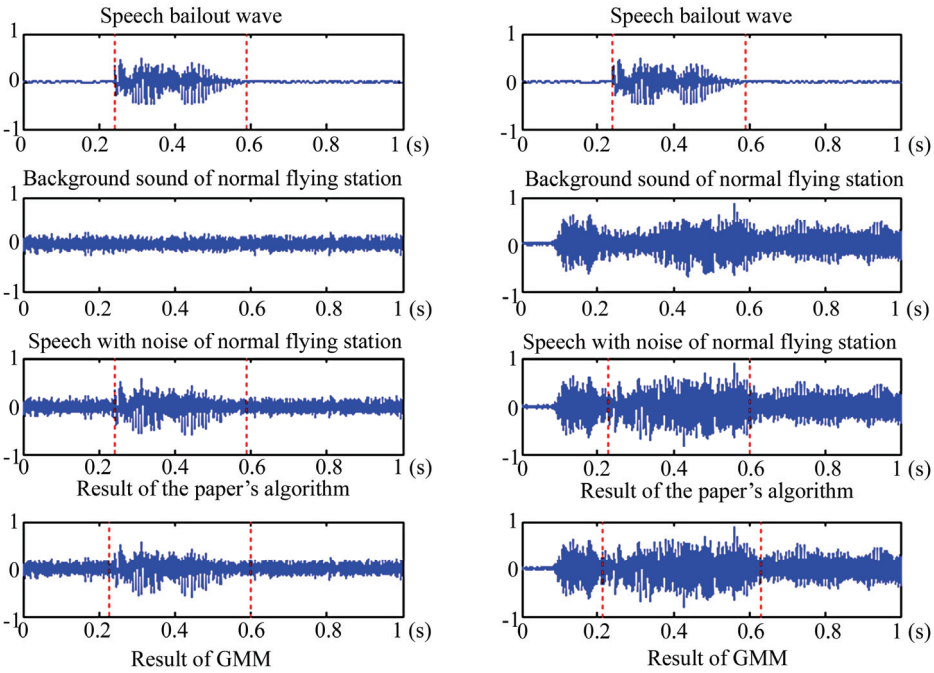


**Figure 3. A real example of speech endpoint detection.**

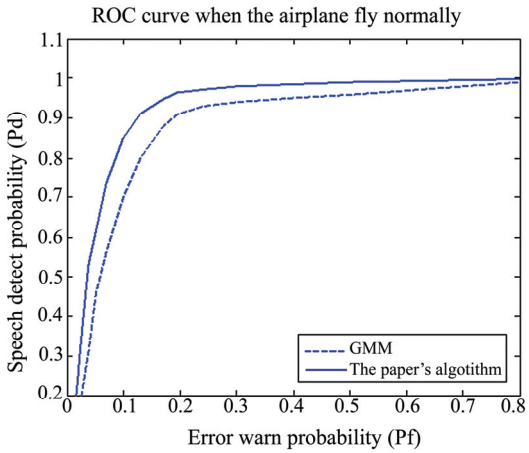


**Figure 4. The ROC curve of airplane normal station**

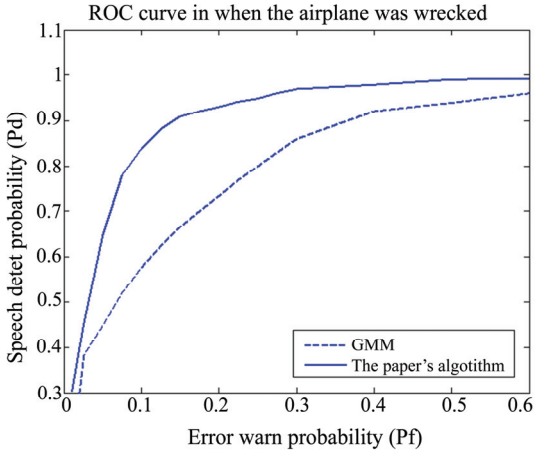


**Figure 5. The ROC curve of airplane wrecked station.**

plane fly normally, the noise in cockpit primary is smooth engine noise and the quiver noise arosed by aerosphere mussy flu. So the veracity of the speech endpoint detection result of the two algorithms almost the same. When the airplane was wrecked, the noise in cockpit is very intensive. The prior half part is the airplane speech alarm sound, the posterior half part is the alarm ring. There is strike sound of pilot pull switch in it. In the complex and nonsmooth background sound, speech was almost silenced. From Figure 3 we can see that the paper's algorithm, modelling directly for speech with noise, robuster than GMM algorithm, modelling noise and speech respectively, and gets better effect of speech endpoint detection.

From Figure 4 we can see, when the airplane fly normally, the ROC curve's best work points of GMM and the paper's algorithm are [0.180,0.885] and [0.135,0.920] respectively. Compared with GMM, The error warn probability and the detect probability of the paper's algorithm reduce 25% and increase 4% respectively. The cause of the phenomena is that the draw up precision of complex Laplacian transformation higher than that of GMM. Adding the application of the relativity of the interframe, his total precision is better than GMM. From Figure 4 we can see, when the airplane was wrecked, the speech endpoint detection algorithm of the paper is better than GMM obviously. The best work points of the two algorithms are [0.141,0.910] and [0.275,0.820] respec-

 

tively. Compared with GMM, The error warn probability and the detect probability of the paper's algorithm reduce 49% and increase 10% respectively. The cause of the phenomena is that GMM modeling noise and speech respectively is not applicable for the environment of wrecked station. When the airplane was wrecked, there are many kinds of noise and they are transient, which is difficult to establish a universal model. Then, the paper's algorithm models the total speech with noise directly and exhibits preferable robusticity.

## 7. Conclusions

The speech endpoint detection of airplane cockpit voice background was put forward by the paper. The two teams' experiment denotes that the algorithm can preserve preferable veracity and robusticity in the airplane normal station and wrecked station.

## 6. References

[1] Y. M. Guo, Q. Fu, and Y. H. Yan, "Speech endpoint detection in complex noise environment [J]," Journal of Acoustics, Vol. 31, No. 6, pp. 549–554, 2006.

[2] D. L. Cheng, C. J. Yi, H. Y. Yao, *et al.*, "The primary research of voice information identify methods of airplane cockpit voice recorder [J]," Control of Noise and Quiver, Vol. 3, pp. 81–84, 2006.

[3] J. L. Shen, J. W. Hung, and L. S. Lee, "Robust entropy-based endpoint detection for speech recognition in noisy environments [C]," In Proceedings of ICSLP, pp. 232–235, 1998.

[4] J. L. Shen and C. H. Yang, "A novel approach to robust speech endpoint detection in car environment [C]," In Proceedings of ICASSP, Vol. 3, pp. 1751–1754, 2000.

[5] C. Jia and B. Xu, "An improved entropy-based endpoint detection algorithm [C]," In Proceedings of ISCSLP, 2002.

[6] J. A. Haigh and J. S. Mason, "Robust voice activity detection using cepstral feature [C]," In Proceedings of IEEE TELCON'93, pp. 321–324, 1993.

[7] X. D. Wei, G. R. Hu, and X. L. Ren," Speech endpoint detection with noise using cepstral feature [J]," Journal of Shanghai Jiao Tong University, Vol. 34, No. 2, pp. 185–188, 2001.

[8] E. Nemer, R. Goubran, and S. Mahmoud, "Robust voice activity detection using higher-order statistics in the LPC residual domain [J]," IEEE Transactions on Speech and Audio Processing, Vol. 9, No. 3, pp. 217–231, 2001.

[9] R. Q. Yan and Y. S. Zhu, "Speech endpoint detection based on the analysis of signal recursion [J]," Journal of Communication, Vol. 1, pp. 35–39, 2007.

[10] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection [J]," IEEE Signal Processing Letters, Vol. 6, No. 1, pp. 1–3, 1999.

[11] A. Davis, S. Nordholm, and R. Togneri, "Statistical voice activity detection using low-variance spectrum estimation and an adaptive threshold [J]," IEEE Transactions on Audio, Speech, Language Process, Vol. 14, No. 2, pp. 412–424, 2006.

[12] M. Fujimoto, K. Ishizuka, and H. Kato, "Noise robust voice activity detection based on statistical model and parallel non-linear Kalman filter [C]," ICASSP'07, pp. 797–800, 2007.

[13] J. H. Chang, J. W. Shin, and N. S. Kim, "Likehood ratio test with complex Laplacian model for voice activity detection [C]," In Proceedings of Euro Speech, pp. 1065–1068, 2003.

[14] M. J. F. Gales, "Models based techniques for noise robust speech recognition [D]," Cambridge University, 1995.

[15] H. Hirsch and C. Ehrlicher, "Noise estimation techniques for robust speech recognition [A]," ICASSP'95 Proceedings, pp. 153–156, 1995.

[16] N. S. Kim and J. H. Chang, "Space enhancement based on global soft decision [J]," IEEE Signal Processing Letters, Vol. 7, No. 5, pp. 108–110, 2000.

[17] W. H. Shin, B. S. Lee, Y. H. Lee, *et al.*, "Speech/non-speech classification using multiple features for robust endpoint detection [C]," In Proceeding of ICAASSP, Vol. 3, pp. 1399–1402, 2000.

[18] J. J. Lei, "The research of some issues in noise robust speech identification [D]," Doctor Thesis of Beijing University of Posts and Telecommunications, 2007.

 