Sound Identification and Speaker Recognition for Aircraft Cockpit Voice Recorder

Yang Lin [1,2]

[1]*Civil Aviation College, Nanjing University of Aeronautics and Astronautics, Nanjing,* 210016, *China*

[2]*Civil Aviation Safety Technical Center of CAAC, Beijing,* 100028, *China*

Email: yanglin@mail.castc.org.cn

Abstract: As air transportation systems have expanded around the world in recent decades, aviation safety and accident/incident prevention have assumed greater importance to governments and airlines. Aircraft accident investigation has a key role to play when an aircraft has an accident or unexpected incident during flight operations. Traditionally the Flight Data Recorder (FDR) has played the major role in establishing the causes of most accidents or incidents. However, information contained in the Cockpit Voice Recorder (CVR) is also very useful during such investigations by providing a better understanding of the real situation. The CVR can act effectively as a latent signal transducer for both the speech and non-speech audio information. Some typical techniques, such as sound identification, voice recognition, appear to offer significant clues in the analysis and classification of speech and non-speech CVR signals.

Keywords: cockpit voice recorder; sound identification; speaker recognition

1. Introduction

The CVR records audio information on four channels. Non-speech information from the Cockpit Area Microphone (CAM) is recorded on channel 1. CAM records thumps, clicks and other sounds occurring in the cockpit other than speech. Channel 2 and 3 of the CVR record speech audio information from the Captain and First Officer's audio selector panels. Channel 4 records the audio information from the jump seat/observer's radio panel.

2. Background Cockpit Sound Identification

It may be hard to believe that non-speech sounds are highly important to the investigation of aircraft damage because the background cockpit sounds can reveal problem areas of the aircraft during the time leading up to the accident. Non-speech data from the CAM can be analyzed with sound spectrum analysis to detect whirl flutter, as well as possibly distinguish the sound of a bomb explosion from the sound of cabin decompression. Spectrum analysis can also be used to confirm that the clicks and thumps recorded by the CAM are simply generated by cockpit controls, and the sound of the aircraft moving through the air.

Analysis background information recorded in aircraft CVRs has been proposed as a complement to the analysis of onboard FDRs in civil aircraft investigations. One reported case provides a good example of the analysis of CVR data playing a key part in an aircraft accident investigation. In 1992, a 19-seater commuter aircraft crashed during an evening training mission. At that time, the US Federal Aviation Agency (FAA) did not require the installation of FDR onboard all small commercial aircrafts, and the CVR onboard the crashed small jet was the only flight record available to provide clues to the causes of the accident. Fortunately, in this case, the CVR recording not only included the voice communication, but also structural acoustics as well as other sounds and noise sources. This allowed the accident investigation to focus on the non-speech sounds taken from the CVR tape. A close inspection of the time series from the CVR track revealed a periodic set of transient components occurring at a frequency of 0.86 Hz. Comparing this frequency with an independent dynamic analysis of the engine mount damage, the 0.86 Hz transient data were demonstrated by independent structural and flutter analyses to be quite close to the frequency experienced from a damaged engine mount. Moreover, there was a sudden loud sound at the end of the tape. This 25 millisecond long event was much louder than the sound in the cabin. Although this short length of the sound did not provide adequate audio listening time, there was enough signal time and amplitude to perform wavelet and voice recognition analysis. The conclusion drawn after the investigation was that the engine on the starboard wing separated during the flight. Subsequently, the fallen engine struck the tail of the aircraft, damaging most of the horizontal surfaces. The loss of the engine also led to the separation of the right wing panel outboard of the engine. As a result, the aircraft pitched down, rolled to the right and crashed [1,2,3].

The results of the accident investigation described above, and Pan Am Flight 103 disintegrated over Lockerbie, Scotland in 1989 due to a bomb explosion, motivated

Project supported by the National Natural Science Foundation of China (Grant No. 60776819).

978-1-935068-14-3 © 2010 SciRes.

us to explore the analysis of aircraft CVR sound sources for use in aircraft accident investigations.

2.1 Framework of Background Cockpit Sound Identification System

The framework of the system composes of three modules: feature extraction, audio search, and audio database[4]. When audio signals are fed into the system, it extracts audio features first; audio features are compared with the features in audio database. Audio candidates are generated according to the result of the match process.

Feature extraction module does some preliminary processing, such as down sampling, low band pass filtering. Then it computes the audio features using algorithm described in 2.2 below. Audio feature database stores the audio features computed in advance. Audio search module compares the features of possible identical audios and outputs the best candidates.

2.2 Audio Feature Extraction Approach

Because human hearing is most sensitive to the frequencies below 2,000 Hz, high frequency parts lose heavily when audios are encoded at very low bit rates. Accordingly, in this system audio signals are down sampled to 5,000 Hz first. Then signals are segmented into frames and weighted by hamming window. Fourier Transformation is performed and spectrum power is obtained. 33 overlapped frequency bands are used at equal logarithm interval. 32-bit audio feature is computed for each frame.

In order to make audio feature stable, frame length as long as 410 milliseconds is chosen. Frame shift is only 12.8 milliseconds. As a result, the frame boundaries of audio queries in the worst case are 6.4 milliseconds off from the boundaries used in the database that are pre-computed.

2.3 Audio Search Approach

2.3.1 Audio Feature Similarity Measurement

Each frame has one 32-bit audio feature. The similarity of two features is measured by Hanning distance, which is the number of different bits. The smaller the Hanning distance, the more similar the two features are, vice versa.

Bit Error Rate (BER) defines the similarity of two audio feature serials with same length. Let **X**, **Y** are two audio feature serials, **X**=$\left\{x_{1,}x_{2,}...,x_n\right\}$,

Y=$\left\{y_{1,}y_{2,}...,y_n\right\}$.

Where N is the frame number of the features. The BER between **X** and **Y** is

$$BER\ (X\ ,Y\ )=\frac{\sum_{i=1}^{N}H\ (x_i\ ,y_i\ )}{32\ N}$$

Where, ***H***(.) is Hanning distance between **X** and **Y.** Obviously, 0≤BER≤1, the lower BER is, the more similar the two feature serials are.

2.3.2 Beam Based Search Approach

When searching audio candidates in the audio database, it will be of very low efficiency if whole match comparison is processed at every possible starting frame. A beam-based search strategy is presented in this system to avoid low efficiency. The main idea of this approach is that it takes the current best score as the base and prunes away all branches whose scores are higher than the base plus the empirical threshold (beam width).

2.4 Experimental Results

First, we used Chinese National Project Speech Database as test data. All silent parts at the beginning and the end of the speech files are cut off. Speech files are merged into 5-minute long files. Totally 20 hours of speech is used as the audio database. Five hundred 3-second audio files are picked out from the database randomly. These 3-second audios are used as the audios to be identified. All these speech data are in PCM 16K sample rate format originally. They are encoded by various codecs at different bit rate and then decoded to PCM 8K sample rate wave files that are used in our experiments.

We performed two types of audio search tests in this study. First, the following three types of sounds generated in the cockpit were recorded as sound samples.

- Warning and alert signals such as GPWS, TCAS, engine fire, autopilot disengage, etc;
- Sounds generated by switches on central panel P2, glare shield P7, and forward overhead panel P5;
- Sounds generated by levers such as landing gear lever, thrust lever, speed-brake lever, and flap lever as well as stall warning signals generated by the levers.

Nearly twenty hours of audio data are saved in the audio database. Fifty audio segments that are less than three seconds long are used as queries to be identified.

The first type of audio search tests uses the cockpit sound samples we recorded earlier to test against all pre-recorded cockpit sound samples in the same database.

The second type of audio search tests uses real sound recorded on tape CVR and Solid State CVR. The sound samples to search with are the same as the first type of tests. The higher the score is, the more similar the two feature series are.

From the test results, we can see that, in first type of audio search tests, three types of sounds generated in the cockpit were recorded as sound samples have the higher score. While in second type of audio search tests, warning, alert signals and switch sounds have higher scores; the sounds generated by various levers have the lowest score.

3. Automatic Speaker Recognition

The speech information recorded by the CVR can be analyzed with spectrum analysis in order to match the recorded voices to the appropriate person.

Automatic speaker recognition, automatically extract information transmitted in speech signal, which can be classified into identification and verification, is the task of identifying a speaker based on his or her voice in CVR recording. Speaker identification is the process of determining which registered speaker provides a given utterance. Speaker verification is the process of accepting or rejecting the identity claim of a speaker. Speaker recognition methods can also be divided into text-dependent and text-independent methods. The former require the speaker to say key words or sentences having the same text for both training and recognition trials, whereas the latter do not rely on a specific text being spoken.

3.1 Framework of Automatic Speaker Recognition System

The framework of the system composes of three modules: feature extraction, speaker modeling, and speaker recognition. When audio signals are fed into the system, the speaker features are drawn from the input speech segments. Furthermore, the influence of channel and environment is restrained by robust techniques. During speaker modeling process, input front-end features characterize the speaker. GMM or SVM modeling approach is used to train the target speaker models, which compose the speaker model database.

3.2 Speaker Feature Extraction Approach

There are two main aspects of speaker features. First, physiologic structure is different by individual. Second, the uttered habits are different. It can be described as prosody features. In the field of speech signal processing, the former is embodied on the structure of frequency. The classical features include cepstral and pitch. And the latter is embodied on the variability of the speech based on the spectral structure. The classical features include the delta cepstral and delta pitch.

In speaker recognition, the cepstral is used mostly and could achieve a good performance. Besides, it can be extracted more easily than other features. At present, the Mel Frequency Cepstral Coefficients (MFCC) is used successfully in speaker recognition, which is proved in applications. In feature extractors of speaker systems, all of the feature vectors are processed by CMS and feature warping method.

Using the delta cepstral information based on time domain is proved that the performance of speaker recognition is enhanced mostly. In our system, speech data are parameterized every 25ms with 15ms overlap between contiguous frames. For each frame a feature vector with 52 dimensions is calculated: 13 Mel Frequency Perceptual Linear Predictive (MFPLP) coefficients, 13-delta cepstral, 13 double delta cepstral and 13 triple cepstral.

3.3 Two of the Speaker Models

The GMM system uses a 100-3800 Hz bandwidth front end consisting of 24 MEL filters to compute 13 cepstral coefficients (C1-C13) with cepstral mean subtraction, and their delta, double delta, and triple-delta coefficients, producing a 52 dimensional feature vector. The feature vectors are modeled by a 2048-component GMM.

The cepstral SVM system is based on the cepstral sequence kernel[5]. All of them use basic features, which are similar to the cepstral GMM system. The only difference is that MFCC features are appended with only delta and double delta features.

3.4 Speaker Recognition Approaches

The speaker identification is that given the test speech segment, the system needs to choose the true speaker from the speaker models database. The key function is calculating the log likelihood of the input test speech features and one target speaker model. Its calculated method is denoted as the follows:

$$S(X) = \log p\left(X \mid \lambda_{hyp}\right) - \log p\left(X \mid \lambda_{UBM}\right)$$

Where $S(X)$ is the final output score, $p\left(X \mid \lambda_{hyp}\right)$ is the probability of the speech segment based on the hypothesis model, $p\left(X \mid \lambda_{UBM}\right)$ is the probability of the speech segment based on UBM. The final output score $S(X)$ is according as the final answer "YES" or "NO" by comparing with the system threshold.

3.5 Evaluation and Experimental Results

3.5.1 NIST Evaluation

We used NIST 06 SRE tasks and data as training and test data. The task of speaker detection includes single speaker verification and conversational speaker verification based on telephone database.

There is a single basic cost model for measuring speaker detection performance, to be used for all speaker detection tests. For each test, a detection cost function will be computed over the sequence of trials provided. Each trial must be independently judged as "true" (the model speaker speaks in the test segment) or "false" (the model speaker does not speak in the test segment), and the correctness of these decisions will be tallied. This detection cost function is defined as a weighted sum of miss and false alarm error probabilities:

$$\begin{aligned} C_{\text{Det}} &= C_{\text{Miss}} \times P_{\text{Miss} \mid \text{Target}} \times P_{\text{Target}} \\ &+ C_{\text{FalseAlarm}} \times P_{\text{FalseAlarm} \mid \text{NonTarget}} \times (1 - P_{\text{Target}}) \end{aligned}$$

The parameters of this cost function are the relative costs of detection errors, C_{Miss} and $C_{\text{FalseAlarm}}$, and the *a priori* probability of the specified target speaker, P_{Target}. The parameter values is used as the primary evaluation of speaker recognition performance for all speaker detection

978-1-935068-14-3 © 2010 SciRes.

tests, C_{Miss} =10, $C_{\text{FalseAlarm}}$ =1, P_{Target} =0.01. Besides, the Equal Error Rate (EER) is also described the performance. EER is denoted as the point value when the miss rate is equal to the false alarm rate,

$$EER = P(\text{Miss} \mid \text{Target}) = P(\text{False} \mid \text{Non Target}) .$$

3.5.2 Experiments and Results

The experiments are assigned based on 2006 NIST speaker recognition evaluation database.

- The performance of the speaker identification system The correct detection rate is more than 92 percents, in case that the speaker number of closed database is no more than 50 and the number of candidate is 5.
- The performance of the speaker verification system. The EER is no more than 10 percents. It is similar to the best level compared with other systems.

3.5.3 Real CVR Speaker Recognition Performance Test

First we created Captain (CAP), First Officer (FO), and Observer (OBS) speech segments manually from four audio files from a 30 min tape CVR. The segments were then saved as small wav files, which were used to train speaker models for CAP, FO, and OBS. Due to limited data availability, CAP and OBS speaker modeling used about 30 seconds of data, and FO speaker modeling used about 60 seconds of data.

Next for each speaker, we randomly chose four segments to test; for each testing segment, a score was calculated for each of the three speakers, and the speaker with the highest score was identified. In most cases, speaker recognition was correct. Because we have limited pilots' data available, we also see some mistakes.

Correct rate is usually used to evaluate the effectiveness. The correct rate C is defined as follow:

C=number of correctly recognized test utterances / total number of test utterances

In this test, correct rate is 10 out of 12, which is 83.3%.

4. Acknowledgement

I hereby express gratitude to my dear partner Prof. Pan Jielin, without his effort, this paper can not be accomplished. In the process of compilation, he made great contribution on data collecting and analyze. This thesis is rather a common achievement than a private possession.

References

[1] Vogt, C., Coughlin, S., Lauber, J. K., Hart, C. A., and Hammer Schmidt, J., "Loss of Control Business Express- Beechcraft 1900C N811BE Near Block Island, Rhode Island," Aircraft Accident/Incident Summary Report, National Transportation Safety Board, Washington, D.C., April 1993.

[2] Stearman, R. O., Schulze, G. H., and Rohre, S. M., "Aircraft Damage Detection from Acoustic and Noise Impressed Signals Found by a Cockpit Voice Recorder," Conference Institute of Noise Control Engineering, 1997.

[3] Stearman, R. O., Schulze, G. H., Rohre, S. M., and Buschow, M. C., "Aircraft Damage Detection from Acoustic Signals Found by a Cockpit Voice Recorder", 133rd Meeting Lay Language, Acoustical Society of America, 1997.

[4] A. Kimura, K. Kashino, T. Kurozumi, and H. Murase, “Very quick audio searching: introducing global pruning to the time-series active search,” in Proc. of Int. Conf. on Computational Intelligence and Multimedia Applications, Salt Lake City, Utah, May 2001.

[5] D. Reynolds, T. Quatieri, and R. Dunn, “Speaker Verification Using Adapted Mixture Models,” Digital Signal Processing, vol. 10, pp.181-202 (2000).