Mail address:
P. O. Box 84
NewYork, NY
10116 USA

World Organization to
Investigate the Persecution
of Falun Gong
追查迫害法轮功国际组织

hwww.upholdjustice.org
Tel：001-347-448-5790
Fax：001-347-402-1444

Speaker Verification Report on CCP Media Reports on "Tiananmen Self-Immolation"

1. Introduction

On March 12, 2003, the National Taiwan University Speech Processing Laboratory was asked to conduct tests on three episodes of a program entitled "Jiao Dian Fang Tan" broadcast by China Central TV (CCTV). The purpose of this test is to verify whether the two people who appeared repetitively in these three episodes, Liu Baorong and Wang Jindong, are actually the same people each time they appear.

The three videos used in this test are from the CCTV program "Jiao Dian Fang Tan." The videos consist of interviews of Liu and Wang regarding the Tiananmen self-immolation incident that occurred on January 23, 2001. Liu Baorong appears in Videos 1 and 2, and Wang Jindong appears in all three videos.

The recording environment for the interviews varies in the videos. When Liu Baorong is interviewed in Video 1, the recording environment is indoors and quiet. In Video 2, she is interviewed in her bedroom. The environment of Wang Jindong's interview in Video 1 is a hospital room. The first part of Video 2 is a hallway with echoes and the second part is a large, quiet room. Different recording conditions present challenges for the test results of speaker verification. Later in this section, we will discuss the method adopted in this report to resolve this issue.

For many years, the Taiwan University Speech Processing Laboratory has been dedicated to enhancing technology in Chinese language recognition and verification, and has accumulated many achievements. This test is conducted on the basis of speaker verification technology researched and developed by Weiren Chung for his Master's Thesis of June 2001. Speaker verification is a technology that verifies a speaker's identity based on the speaker's voice. Similar research done around the world can be traced back many years. Popular applications include financial transactions and crime investigation and prevention, among others.

According to Reference [1], popular models for speaker verification include the Gaussian Mixture Model (GMM), the Hidden Markov Model (HMM), and Eigenvoice. The Gaussian Mixture Model is a simplification of the Hidden Markov Model. Its principle is to separate one speaker's Training Corpus into groups according to the characteristics of the sound. A Gaussian distribution is used to describe each and every group of audio characteristics.

The Hidden Markov Model performs better in speaker verification than the Gaussian Mixture Model. But because its system is more complicated and requires more of a Training Corpus, it is not suitable for this test. The Eigenvoice Model was not adopted because its performance is not as good as the Gaussian Mixture Model.

As stated in the beginning of this section, different recording conditions present difficulties in speaker verification. Different recording conditions can produce verification results indicating that two speakers are different when, if fact, the two recordings came from the same person. This is due to differences in the environment (different microphones, noises and echoes, etc.). This situation is called False Rejection.

False Rejection occurs when the speaker is the same as the declared identity, but is rejected by the system. Subsequently, False Acceptance occurs when the speaker is different from the declared identity, but is accepted as being the same by the system. Usually, False Rejection and False Acceptance cannot be improved at the same time. There is a trade-off between the two. When one of them is lowered (by decreasing or increasing the threshold), the other one will increase.

To reach the requirement of high credibility, our test is designed to have minimal possibility of False Rejection and maximum possibility of False Acceptance. This way, because the possibility of False Rejection is very low, if the system still decides to reject, the possibility of accurate rejection is greatly improved.

Mail address:
P. O. Box 84
NewYork, NY
10116 USA

World Organization to
Investigate the Persecution
of Falun Gong
追查迫害法轮功国际组织

hwww.upholdjustice.org
Tel: 001-347-448-5790
Fax: 001-347-402-1444

Title	Speaker	Source	Length (m:s)	Time Distribution	
Zf1_liubaorong	Liu Baorong	Zf1.rm	2:36	1:34-1:43* 2:06-2:17* 2:22-2:34* 2:39-2:59* 3:09-3:47* 4:55-5:30 9:54-10:11 15:17-15:40*	
Zf2_liubaorong	Liu Baorong	Zf2.rm	0:32	6:40-7:30*	
Zf1_wangjindong	Wang Jindong	Zf1.rm	0:06	4:30-4:34 13:10-13:21* 13:30-13:31*	
Zf2_wangjindong	Wang Jindong	Zf2.rm	0:30	9:06-9:24* 9:58-10:20*	
Zf2_wangjindong2	Wang Jindong	Zf2.rm	4:08	10:28-10:40* 11:08-11:55* 12:01-12:19* 12:44-12:55* 13:12-14:46 14:58-15:42 15:57-16:41*	
Zf3_wangjindong	Wang Jindong	Zf3.rm	0:55	9:07-9:22 9:30-10:13	
Zf1_reporter	Reporter interviewing Liu Siying	Zf1.rm	0:05	9:11-9:18*	
Zf1_reporter2	Reporter interviewing Liu Xuefang	Zf1.rm	0:09	12:36-12:44*	
Zf1_reporter3	Reporter interviewing Wang Jindong	Zf1.rm	0:07	13:07-13:18*	
Zf1_reporter4	Reporter interviewing He Haihua, Wang Juan	Zf1.rm	0:15	13:44-13:48* 13:52-14:01*	
Zf1_reporter5	Reporter interviewing Liu Baorong	Zf1.rm	0:05	15:22-15:28*	
Zf2_reporter	Reporter interviewing Chen Guo	Zf2.rm	0:15	3:05-3:06 4:02-4:53*	
Zf2_reporter2	Reporter interviewing Hao Haijun	Zf2.rm	0:11	3:48-3:50 5:45-6:03*	
Zf2_reporter3	Reporter interviewing Cui li	Zf2.rm	0:05	5:35-5:42*	Other people's voices are eliminated
Zf2_reporter4	Reporter interviewing Liu Baorong	Zf2.rm	0:03	6:51-6:53	
Zf2_reporter5	Reporter interviewing Liu Xuefang	Zf2.rm	0:03	8:09-8:11	
Zf2_reporter6	Reporter interviewing Wang Jindong	Zf2.rm	0:03	9:01-9:05*	

3

World Organization to
Investigate the Persecution
of Falun Gong
追查迫害法轮功国际组织

Mail address:
P. O. Box 84
NewYork, NY
10116 USA

hwww.upholdjustice.org
Tel：001-347-448-5790
Fax：001-347-402-1444

The duration of Zf1_liubaorong is 2 minutes 36 seconds. Zf2_wangjindong2 is 4 minutes and 08 seconds. Because durations of these two are the longest, they are used separately as the Training Corpus of Liu Baorong's and Wang Jindong's Speaker Dependent model.

Because the female reporter's corpus durations are all too short to train a model, we assembled the female reporter's corpus according to the videos as follows:

Zf1_reporter_all	Zf1_reporter + zf1_reporter2 + zf1_reporter3 + zf1_reporter4 + zf1_reporter5
Zf2_reporter_all	Zf2_reporter + zf2_reporter2 + zf2_reporter3 + zf2_reporter4 + zf2_reporter5 + zf2_reporter6 + zf2_reporter7
Zf3_reporter_all	Zf3_reporter + zf3_reporter2
Reporter- 1_2	Zf1_reporter_all + Zf2_reporter_all
Reporter- 2_3	Zf2_reporter_all + Zf3_reporter_all
Reporter- 1_3	Zf1_reporter_all + Zf3_reporter_all

Reporter-1_2, Reporter-2_3, and Reporter-1_3 were used to train three different Speaker Dependent models. To train the value of the threshold, these models were tested and verified separately with Zf3_reporter_all, Zf1_reporter_all, and Zf2_reporter_all.

Finally, there is a corpus to train the Speaker Independent model:

ZFALL_vocal	All voices in all three videos

3.3 Obtaining the Feature Vector

The Feature Vector used in this report is 39 MFCC£¨Mel-Frequency Cepstral Coefficient

Pre-emphasis Filter	$1-0.97z^{-1}$
Frame Size	32 ms
Frame Shift	10 ms
Filter Bank	Mel-Scale Triangular Filter Banks
Number of Filter Banks	26
Low Cut-off Frequency	300 Hz
High Cut-off Frequency	3400 Hz
Feature Vector	12 Mel_Frequency Cepstral Coefficients and one short-time energy and its delta and delta-delta (Total: 39)

The program used to obtain the Feature Vector is from HCopy of HTK 3.0 in Reference [2].

3.4 Training the Speaker Independent Model

The Training Corpus is ZFAll_vocal. The training method is to obtain the initial model through Vector Quantization. When the number of clusters is less than 8, Modified K-means is used. When the number of clusters is greater than 8, Binary Split is used. After obtaining the initial model, Expectation Maximization was performed to obtain the final model [1].

According to [1], in speaker verification using the Gaussian Mixture Model, the error rate is the lowest when Number of Mixtures is 512 or 1024. To cut down on the amount of computation, this report adopted the following:

Number of Mixtures	512

3.5 Speaker Dependent Model

The Speaker Dependent model comes from the Speaker Independent model described in the previous section, after adaptation using the Bayesian Adaptation method. Only the average vector is adapted. Mixture weight and variance are substituted using the coefficients in the Speaker Independent model.

The Speaker Dependent model and its corpus for adaptation used in this report are as follows:

4

Mail address:

P. O. Box 84
NewYork, NY
10116 USA

World Organization to Investigate the Persecution of Falun Gong
追查迫害法轮功国际组织

hwww.upholdjustice.org
Tel: 001-347-448-5790
Fax: 001-347-402-1444

Speaker Dependent Model	Corpus for Adaptation
Zf1_liubaorong.sd.modal	Zf1_liubaorong
Zf2_wangjindong2.sd.modal	Zf2_wangjindong2
Reporter-1_2.sd.modal	Reporter-1_2
Reporter-2_3.sd.modal	Reporter-2_3
Reporter-1_3.sd.modal	Reporter-1_3

3.6 Speaker Verification

Derived from the diagram in Section 2.1, the formula for calculating the verification score of each segment of test speech is: T is the test speech's number of frames, is time t's Feature Vector. S and $S_{\bar{i}}$ are the Speaker Dependent model and Speaker Independent model, respectively.

Verification Scores:

SD Model	Test Corpus	Verification Score
Zf1_liubaorong	Zf2_liubaorong	-0.042003
	Zf1_wangjindong	-0.201615
Zf2_wangjindong2	Zf2_wangjindong	0.128923
	Zf3_wangjindong	0.325247
Reporter-1_2	Zf3_reporter_all	0.146295
Reporter-2_3	Zf2_reporter_all	0.022340
Reporter-1_3	Zf1_reporter_all	0.012399

The first verification score is the verification of Liu Baorong's voice in the second video using the first video's interview of Liu Baorong as the training model.

The second, third, and fourth verification scores are based on using Video 2's second interview of Wang Jindong as the training model to verify the interview of Wang Jindong in Video 1, the first interview in Video 2, and Video 3.

The fifth verification score is based on using the reporter's voice in Video 1 and Video 2 as the Training Corpus to verify the reporter's voice in Video 3. The sixth and seventh verification scores are different combinations of the similar tests.

As described in Section 1, in order to achieve credibility, the value of the test threshold should be set so that all three test corpuses of the reporter will be accepted by the test. Accordingly, the smallest value of test scores 5, 6, and 7 is selected, which is 0.012399.

Threshold	0.012399

Verification Results:

Reference Speaker	Test Speaker	Score	Threshold	Results
Liu in video 1 (Zf1_liubaorong)	Liu in video 2 (Zf2_liubaorong)	**-0.042003**	**o.o12399**	**Rejection**
Second interview in video 2 Wang Jindong (Zf2_wangjindong2)	Wang Jindong in video 1 (Zf1_wangjindong)	**-0.201615**		**Rejection**
	First interview in video 2 Wang Jindong in (Zf2_wangjindong)	**0.128923**		**Acceptance**
	Wang Jindong in video 3 (Zf3_wangjindong)	**0.325247**		**Acceptance**

Female reporter in video 1 and 2 (reporter-1_2)	Female reporter in video 3 (Zf3_reporter_all)	0.146295		**Acceptance**
Female reporter in video 2 and 3 (reporter-2_3)	Female reporter in video 1 (Zf1_reporter_all)	0.022340		**Acceptance**
Female reporter in video 1 and 3 (reporter-1_3)	Female reporter in video 2 (Zf2_reporter_all)	0.012399		**Acceptance**

Acceptance in the Test Result means the test speaker and the reference speaker (speaker in the training model) are determined to be the same person. Rejection means they are not the same person.

From the table above, under the condition of "Minimizing the possibility of False Rejection" (that is "Try best not to reject" or "To accept if two voices have certain similarity"), based on the test result of the voices available to this experiment, the conclusion can be made that Liu Baorong in the first video and Liu Baorong in the second video are not the same person.

The two Wang Jindongs in the second video and the Wang Jindong in the third video are determined to be the same person. Wang Jindong in the first video and Wang Jindong in the other two videos can be determined not to be the same person.

4. Conclusion

By using Gaussian Mixture Model speaker verification technology, this report has reached the conclusion that Liu Baorong and Wang Jindong in the first video are not the same as Liu Baorong and Wang Jindong in the second video.

In Section 3.3, this report used the model with a mixer of 512. Actually, this report also conducted an experiment using a mixer of 256 and 128. Except for the fact that the score was slightly different, the conclusion reached (Acceptance or Rejection) was completely the same.

References

[1] Weiren Chung, "An Initial Study on Speaker Recognition and Verification," 2001, National Taiwan University, Master's Degree Thesis
[2] Steve Young, Dan Kershaw, Julian Odell, et. Al, "The HTK Book (for HTK version 3.0"), July 2000

To investigate the criminal conduct of all institutions, organizations, and individuals involved in the persecution of Falun Gong; to bring such investigations, no matter how long it takes, no matter how far and deep we have to search, to full closure; to exercise fundamental principles of humanity; and to restore and uphold justice in society.