语音验证结果报告

台大语音实验室
2003 年 4 月 19 日

第一章 绪论

2003 年 3 月 12 日台大语音实验室受委託对 3 集中国中央电视台的焦点访谈节目做检验，检验的内容为验证 2 位在 3 集中连续出现的人物－刘葆荣及王进东－是 否真如节目所宣称的为同一人。

此 3 集录影是中国中央电视台播出的焦点访谈节目，内容为 2001 年 1 月 23 日“天安门广场自焚事件”的调查及访谈。其中刘葆荣和王进东为自焚未遂者。刘葆 荣在第一集和第二集出现接受採访，王进东则在全部三集均有出现接受採访。第一集访问刘葆荣时的录音环境为安静的室内，第二集则是刘葆荣家的卧房。

第一集访问王进东的环境为医院的病房，第二集前半为有迴音的走廊，后半为安静的大房间，第三集亦为安静的大房间。不同的录音条件对语者验证结果的可信度构成了极大的考验。本章后半部分将讨论本报告所採用的解决方法。

台大语音实验室多年来致力于提升中文语音辨识技术，已累积丰硕的成果。本测试实验係以 2001 年 6 月研究生锺伟仁在其毕业学位论文中所研发的语者验证(Speaker Verification)技术为基础进行【1】。

语者验证(Speaker Verification)是一种根据说话者的语音与其所宣称的身份，验证说话者是否真如其人的技术。相关的研究在国际上可追溯到许多年前。常见的用途包括金融交易，犯罪侦防等。

根据【1】，常用于语者验证的模型有高斯溷合模型(Gaussian Mixture Model, GMM)、隐藏式马可夫模型(Hidden Markov Model, HMM)、及特性语音(Eigenvoice)。其中高斯溷合模型是隐藏式马可夫模型的简化，原理为把同一语者的训练语料(Training Corpus)依声学特性分群，然后把每一群声学特性用一个高斯分佈来描述。高斯溷合模型也是本报告所使用的方法。隐藏式马可夫模型在语者验证的表现比高 斯溷合模型好【1】，但因系统较为複杂且需要更多的训练语料，因此并不适合用于本测试。特性语音因其表现不如高斯溷合模型【1】，因此亦不採用。

如本章开头所述，不同的录音条件对语者验证构成了极大的困难。因为不同的录音条件可以造成就算是两段同一个人的讲话录音，也可能因环境的差异（如不同的麦克风、不同的噪音、及不同的迴音等）而被验证为不同一人。此种情形称为 False

Rejection。 False Rejection 是说话者的确为其所宣称的身份，但却被系统拒绝（Reject）的情况。相反地，False Acceptance 则是说话者并非其所宣称的身份，但却被系统接受（Accept）的情况。通常 False Rejection 及 False Acceptance 二者无法兼顾，亦即其间是互相取捨（Trade-off）的关係，亦即将其中一个降低的时候（如提高或降低门槛值时），另一个一定会 上升。

为了达到可信度的要求，本测试实验设计为让 False Rejection 的可能性儘量地低，而 False Acceptance 的可能性儘量高。这样一来，因为 False Rejection 的可能性很低，所以如果系统仍然判定为 Rejection，则其为正确（的 Rejection）的可能性就大大地提高。

因为本测试实验採用门槛值（Threshold）做为判定接受（Accept）或拒绝（Reject）的标准，高于门槛值即接受，低于门槛值即拒绝，因此 选 择一个合理但较低的门槛值即可达到降低 False Rejection 但提高 False Acceptance 的目的。

观察 3 集节目，可发现採访受访者的女记者多次出现在节目中，且其录音的条件包含最多种不同的情况（如外场、医院、卧室，监狱，走廊等）。因此如果可以适当地设计门槛值使这些不同录音条件的语音片段都被验证为同一个人，即门槛值低到让这些女记者的语音片段都被系统接受，则可以达到最大的可信度。（注：3 集中进行採访的女记者并不一定都是同一人，但因为是考虑最差情况也要验证通过，因此并不影响）

第二章 理论背景

2.1 语者验证器本报告所使用的语者验证器为一种对数相似值比侦测器（Log-Likelihood Ratio Detector），如下图：

当测试语音经过前端处理抽取出特徵向量（Feature Vector）后，分别将特徵向量对语者特定模型及语者背景模型作对数相似度（Log-Likelihood）的计算，然后再相减得到最后的分数。这麼做 的目的是使最后的分数降低对语者自己语音的变异性（Inner-Speaker Variation），但留下语者间的变异性（Inter-Speaker Variation）【1】。

2.2 语者背景模型（Background Speaker Model）

语者背景模型用来在语者验证中帮助分数的正规化动作，使分数降低对语者自己语音的变异性（Inner-Speaker Variation），但留下语者间的变异性（Inter-Speaker Variation）【1】。

在规模比较大的语者验证系统中，为了简化系统设计的複杂度，通常就拿语者不特定模型（Speaker Independent Model）来当成每一位语者的背景模型【1】。
语者不特定模型由全部语者的语料训练得到。

2.3 语者特定模型（Speaker Dependent Model）

语者特定模型的目标是模拟每一位语者的声学特徵。每一位语者的模型都代表该位语者的语音声学特性。语者特定模型由语者不特定模型经由贝氏调适法调适而来，调适的语料即为该位语者的语料。

第三章 实验方法及结果

3.1 录音将 3 集节目（zf1.rm、zf2.rm、zf3.rm）经由 RealPlayer 播放的同时启动音效卡直接将声音讯号录製下来（在音效卡内部播放同 时录製，并不经由任何外部的导线）。取样参数为：

取样频率	8kHz
取样大小(Sample Size)	16bit
声道(Channels)	2

3.2 切音在前述录制下来的声音中切出需要的片段如下：

名称	语者	来源	长度(分:秒)	时间分佈
Zf1_liubaorong	刘葆荣	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	刘葆荣	Zf2.rm	0:32	6:40-7:30*
Zf1_wangjindong	王进东	Zf1.rm	0:06	4:30-4:34 13:10-13:21* 13:30-13:31*
Zf2_wangjindong	王进东	Zf2.rm	0:30	9:06-9:24* 9:58-10:20*
Zf2_wangjindong2	王进东	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	王进东	Zf3.rm	0:55	9:07-9:22

				9:30-10:13
Zf1_reporter	访问刘思影的女记者	Zf1.rm	0:05	9:11-9:18*
Zf1_reporter2	访问刘云芳的女记者	Zf1.rm	0:09	12:36-12:44*
Zf1_reporter3	访问王进东的女记者	Zf1.rm	0:07	13:07-13:18*
Zf1_reporter4	访问何海华、王娟的女记者	Zf1.rm	0:15	13:44-13:48* 13:52-14:01*
Zf1_reporter5	访问刘葆荣的女记者	Zf1.rm	0:05	15:22-15:28*
Zf2_reporter	访问陈果的女记者	Zf2.rm	0:15	3:05-3:06 4:02-4:53*
Zf2_reporter2	访问郝惠君的女记者	Zf2.rm	0:11	3:48-3:50 5:45-6:03*
Zf2_reporter3	访问崔丽的女记者	Zf2.rm	0:05	5:35-5:42*
Zf2_reporter4	访问刘葆荣的女记者	Zf2.rm	0:03	6:51-6:53
Zf2_reporter5	访问刘云芳的女记者	Zf2.rm	0:03	8:09-8:11
Zf2_reporter6	访问王进东的女记者	Zf2.rm	0:03	9:01-9:05*
Zf2_reporter7	第二次访问王进东的女记者	Zf2.rm	0:31	10:59-12:00* 16:21-16:27
Zf3_reporter	访问冯海军的女记者	Zf3.rm	0:13	2:04-2:13* 3:05-3:25*
Zf3_reporter2	访问马乐的女记者	Zf3.rm	0:16	4:22-4:25 6:21-6:42* 8:22-8:27

* 中间旁白或其他人的声音，已去除其中 Zf1_liubaorong 的长度为 2 分 36 秒，Zf2_wangjindong2 的长度为 4 分 08 秒，因其长度最 长，分别做为训练刘葆荣及王进东的语者特定模型的语料。因为女记者的语料长度都太短，无法训练模型，因此再将女记者的语料依节目组合如下：

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

其中 Reporter-1_2、Reporter-2_3、及 Reporter-1_3 分别用来训练三个不同的语者特定模型，分别与 Zf3_reporter_all、Zf1_reporter_all、及 Zf2_reporter_all 做验证，以训练门槛值。最后还有一个训练语者 不特定模型的语料：

ZFAll_vocal	三集节目中所有的语音

3.3 抽取特徵向量（Feature Vector）本报告所使用的语音特徵向量（Feature Vector）为 39 维的 MFCC（Mel-Frequency Cepstral Coefficient）係数：

预强调滤波器(Pre-emphasis Filter)	1-0.97z-1
音框长度(Frame Size)	32ms
音框平移(Frame Shift)	10ms
滤波器组 (Filter Bank)	梅尔刻度三角滤波器组(Mel-Scale Triangular Filter Banks)
滤波器数 (Number of Filter Bank)	26
低频截止频率 (Low Cut-off Frequency)	300Hz
高频截止频率 (High Cut-off Frequency)	3400Hz
特徵向量 (Feature Vector)	12 维尔频率倒频谱係数加时间轴上正规化之短时能量(MFCC_E)加一阶及二阶回归係数(MFCC_E_D_A)共 39 维

抽取特徵向量的程式是藉由 HTK 3.0 所附的 HCopy【2】。 3.4 训练语者不特定模型（Speaker Independent Model）训练语料为 ZFAll_vocal。训练的方式是先由向量量化（Vector Quantization）求得初始模型，当群数小于 8 时，使用 K 平均值修正法（Modified K-means）；当群数大于 8 后，改用二值分裂法（Binary Split）。得到初始模型后再做期望值最大化（Expectation Maximization）得到最终的模型【1】。根据【1】，在高斯溷合模型的语者验证实验中，溷合数（Number of Mixtures）为 512 或 1024 时验证错误率为最低。为简少计算量，因此本报告取溷合数为：

混合數（Number of Mixtures）	512

3.5 调适语者特定模型（Speaker Dependent Model）语者特定模型由上一节所述的语者不特定模型（ZFAll_vocal）经由贝氏调适法调适而来。且只对平均值向量作调适， 而溷合加权值及变异 量则用语者不特定模型的参数代替【1】。本报告所用到的语者特定模型及其调适语料如下：

语者特定模型	调适语料
Zf1_liubaorong.sd.model	Zf1_liubaorong
Zf2_wangjindong2.sd.model	Zf2_wangjindong2
Reporter-1_2.sd.model	Reporter-1_2
Reporter-2_3.sd.model	Reporter-2_3
Reporter-1_3.sd.model	Reporter-1_3

3.6 语者验证由 2.1 图，最后计算每段测试语音的验证分数的公式为：

$$Score = \frac{1}{T}\sum_{t=1}^{T}\log f(\overrightarrow{x_t} \mid S) - \frac{1}{T}\sum_{t=1}^{T}\log f(\overrightarrow{x_t} \mid S')$$

上表第一列为以第一集的刘葆荣访问内容训练模型，验证第二集的刘葆荣的声音的分数。第二、三、四列分别为以第二集第二次访问王进东的内容训练模型，验证第一 集、第二集第一次、及第三集访问王进东的内容。第五列为以第一集、第二集的女记者声音为训练语料，验证第三集的女记者声音的分数。第六、七列类推。如第一 章所述，为了取得可信度，因此设计门槛值为使这三段女记者的测试语料都通过验证，所以取第五、六、七列分数的最小值为门槛值，即 0.012399。

门槛值（Threshold）	0.012399

验证结果为：

参考语者	测试语者	分数 (Score)	门槛值	验证结果
第一集的刘葆荣 (Zf1_liubaorong)	第二集的刘葆荣 (Zf2_liubaorong)	-0.042003	0.012399	拒绝
第二集的第二次访问王进东 (Zf2_wangjindong2)	第一集的王进东 (Zf1_wangjindong)	-0.201615		拒绝
	第二集的第一次访问王进东 (Zf2_wangjindong)	0.128923		接受
	第三集的王进东(Zf3_wangji	0.325247		接受

	ndong)			
第一、二集的女记者 (reporter-1_2)	第三集的女记者 (Zf3_reporter_all)	0.146295		接受
第二、三集的女记者 (reporter-2_3)	第一集的女记者 (Zf1_reporter_all)	0.022340		接受
第一、三集的女记者 (reporter-1_3)	第二集的女记者 (Zf2_reporter_all)	0.012399		接受

验证结果为接受意指该测试语者与参考语者（即训练模型的语者）经本测试判定为同一人，拒绝意指判定为不同一人。因此由上表，在本实验所设定的“儘可能把 False Rejection 降至最低”，亦即“儘可能不去拒绝”，或“只要两段声音有一定相似度，就会被接受”的条件下，在本实验所拥有的语料下所测试的结果为第 一集的刘葆荣与第二集的刘葆荣应可判断不是同一人，第二集的第一、二个王进东与第三集的王进东应可判断是同一人，第一集的王进东与其余 2 集的王进东应可判 断不是同一个人。第四章结论本篇报告经由高斯溷合模型的语者验证技术，判别出了焦点访谈录像中第一集的刘葆荣与王进东分别与第二集的刘葆荣与王进东应可判 断不是同一个人的结论。 在 3.3 节中，本篇报告採用溷合数为 512 的模型，其实本篇报告亦有做溷合数为 256 或 128 的实验，除了数字少许不同外，结论（验证接受或拒绝）却是完 全相同的。

参考文献

1. 锺伟仁，“语者辨认与验证之初步研究(An Initial Study on Speaker Recognition and Verification)”，2001 年，国立台湾大学电信工程学研究所 硕士学位论文
2. Steve Young, Dan Kershaw, Julian Odell, et. al, “The HTK Book (for HTK version 3.0)”, July 2000

报 告全文(Word File)