基于参考说话人模型和双层结构的说话人辨认快速算法

王　刚，邬晓钧，郑　方，王琳琳，张陈昊

（清华信息科学技术国家实验室 技术创新与开发部 语音和语言技术中心，清华大学计算机科学与技术系，北京，100084）

摘　要：为了提高基于高斯混合模型通用背景模型(GMM-UBM)的说话人辨认系统的运算效率，提出一种基于参考说话人模型的双层结构用于目标说话人剪枝，采用矢量量化方法从目标说话人模型集合中训练参考说话人模型，利用语音与参考说人模型的偏差来描述说话人的发音特性，将辨认语音偏差向量和目标说话人偏差向量的相似性作为距离度量进行目标说话人剪枝。实验结果表明：在基于GMM-UBM 的说话人辨认系统中，对包含5,200个目标说话人和1,000个集外说话人的测试集进行开集辨认的条件下，在提高辨认的运算效率12.5倍同时而识别率仅下降0.3%。

关键词：双层结构；说话人快速辨认；参考说话人模型

中图分类号：　TP391

An algorithm for efficient speaker identification using reference speaker model based two-layer structure

Gang Wang, Xiaojun Wu, Thomas Fang Zheng, Linlin Wang and Chenhao Zhang

(Center for Speech and Language Technologies, Division of Technical Innovation and Development, Tsinghua National Laboratory for Information Science and Technology , Department of Computer Science and Technology, Tsinghua University, Beijing, 100084, China)

Abstract: To improve the GMM-UBM based speaker identification system's computation efficiency, a fast algorithm using reference speaker model based two-layer structure is proposed. Vector quantization is used to train the reference speaker models using target speaker models. The deviations between one speaker and reference speaker models are used to model the speaker's acoustic characteristics. The correlation between the deviations' vectors is used to evaluate the similarity degree between the identified speech and target speakers and to prune those target speakers with lower similarity degree. Experimental results proved that on the database containing 5,200 target speakers and 1,000 out-of-set imposters, the proposed algorithm improved the identification efficiency by 12.5 times with performance degradation of only 0.3%.

Key words: Two-Layer Structure; Speaker Identification; Reference Speaker Model

说话人辨认是说话人识别的一种，把待辨认的语音判定为是否属于N个目标说话人当中的某一位，是一个多选一的问题[1]。说话人辨认在近十几年来一直都是研究热点，也在许多领域进行了实际应用，如司法和金融领域。目前说话人辨认最流行的方法是高斯混合模型通用背景模型(Gaussian Mixture Model-Universal Background Model, GMM-UBM) [2]，支持向量机（Gaussian Mixture Model-Support Vector Model, GMM-SVM）[3]，或者以GMM-UBM为基础进行的一定地改进，如联合因子分析（Joint Factor Analysis, JFA）[4]等。当前的说话人辨认系统在一定条件下已经能达到很高的准确率[1-2]，但是随着目标说话人数量的增多[5]（几千甚至上万或更大时），目前的说话人辨认系统的时间性能往往较难满足要求，尤其是对于那些实时性要求较高的系统。例如在安全监听当中，需要快速辨认监听语音是否属于目标说话人集合中的某一个，不仅要求系统有好的辨认准确率还要有很高的辨认速度。

基于GMM-SVM的说话人辨认系统[3]需要将待辨认语音训练成高斯混合模型作为SVM的输入，训练高斯混合模型一般采用最大后验概率（Maximum *a posterior*, MAP）[6]算法、最大似然线性回归（Maximum likelihood linear regression, MLLR）[7]算法等，相当耗时且很难改进。因此大多数的快速辨认算法都是对基于GMM-UBM的系统进行改进。在基于GMM-UBM的说话人辨认系统中，运算量主要集中在两部分[8-9]。一是待辨认语音的特征数量。每一帧特征都需要从UBM所有混合中挑选核心分布，计算每一帧特征矢量对UBM 中所有单高斯分布的似然分, 从中挑选出分数最高的前N（4或5）个高斯分布作为核心分布[2]。一般来说UBM为了全面的覆盖整个声学空间其混合数都较大（1,024或2,048）[2]，而且似然分的运算是耗时较大的运算。二是目标说话人的数量。每一帧特征在挑选完UBM核心分布之后还要对所有目标说话人模型上对应

作者简介：王刚（1976-），男（汉），河北省卢龙县，博士研究生。

通讯作者：郑方，研究员，E-mail：fzheng@tsinghua.edu.cn

的UBM核心分布上计算似然分，当目标说话人集合大时运算量也会大。

目前已有一些快速辨认算法：快速挑选UBM上核心分布的方法，哈希高斯混合模型算法(Hash GMM, HGMM）[10]，结构化高斯混合模型方法（Structural Gaussian Mixture Model-SGMM）[11]，基于树形UBM的核心挑选（Tree Based Kernel Selection, TBKS）算法[12]等，这些算法通过将UBM组织成某种排序结构并利用剪枝来降低UBM中单高斯分布的计算数量，从而达到加速的目的，目前这类算法相对较成熟在提升速度的同时性能下降很小甚至可以忽略，但在大规模目标说话人条件下，挑选核心分布的计算量仅占说话人辨认总计算量的很小的一部分，因而其对整个辨认来说其加速贡献很小；压缩语音特征数量的方法, 特征矢量重排序的剪枝算法(Observation Reordering by Pruning, ORBP) [13], 预量化（Pre-Quantization， PQ）方法[14]等下采样方法（Down-Sampling或Sub-Sampling），这类方法利用了相邻帧语音的相关性较大似然分差异较小且每帧语音计算的先后顺序与最终的似然分无关[15]，首先对语音进行下采样（典型采样间隔为4帧）[13]，开始只使用一组采样语音计算似然分然后利用该得分剪枝掉那些得分很低的目标说话人模型，再增加一组采样语音计算并更新似然分然后剪枝，直到所有语音全部被使用。该类算法能够提升辨认速度，但由于其采样语音的得分与全部语音得分相比较存在一定的波动，算法的稳定性存在一定欠缺；目标说话人剪枝方法，说话人模型聚类算法（Speaker Models Clustering, SMC）[5]和分层结构的说话人辨认（Hierarchical Speaker Identification, HSI)[16]等算法，首先利用某种聚类算法将目标说话人模型聚类得到一组聚类中心，辨认阶段待辨认语音首先在聚类中心上计算似然分，然后对属于得分最高的聚类中心那一类的目标说话人模型计算似然分进行辨认。该类方法容易受目标说话人数量的影响，因为大规模目标说话人聚类后各类之间的区分性往往会变差，会有较多的目标说话人处于类间的边界位置，当与待辨认语音对应的真实目标说话人处于这样的位置时，就很可能会由于类的选择而无法被正确检出。当然可以增大辨认的类数来减弱这种影响，但这会导致运算速度的提高幅度下降。

SMC和HSI算法性能受目标说话人规模的影响较大，其根本原因在于减枝后保留的目标说话人与聚类中心相似程度较大，而未必与待辨认语音相似程度高，当待辨认语音与处于易混淆的两类间的边界时，SMC和HSI仅保留了某一类中的目标说话人，而理想的结果是保留这两类边界处的目标说话人。本文提出了一种基于参考说话人模型的目标说话人剪枝算法（Reference Speaker Model based Speaker Pruning, RSMSP），快速挑选与待语音的相似程度较高的目标说话人进行辨认。与SMC类似，RSMSP也使用了矢量量化中的聚类方法[17]对目标说话人模型进行聚类得到多个聚类中心（即参考说话人模型），由于参考说话人模型模拟了同一类中的目标说话人的共性的声学特性，语音与某个参考说话人模型的偏差在一定程度上描述了语音同该类目标说话人发音特性的相似程度，那么语音与多个参考说话人模型的偏差就能够充分地描述语音的发音特性，由这多个偏差组成的向量称为偏差向量。两段语音偏差向量之间的相似性就可以度量两段语音的说话人之间的相似程度。目标说话人的偏差可在训练阶段计算好，在辨认阶段计算待辨认语音的偏差并比较待辨认语音和所有目标说话人之间的相似程度，并根据相似程度来挑选与待辨认语音相似程度最高的那部分目标说话人进行辨认。此外，设计了双层结构来进一步提高剪枝效率。

本文安排如下：首先介绍基于 RSMSP 的算法，然后介绍双层结构的快速辨认算法，接下来介绍本文的实验设置、实验结果以及结果的分析，最后是对本文研究工作的总结与展望。

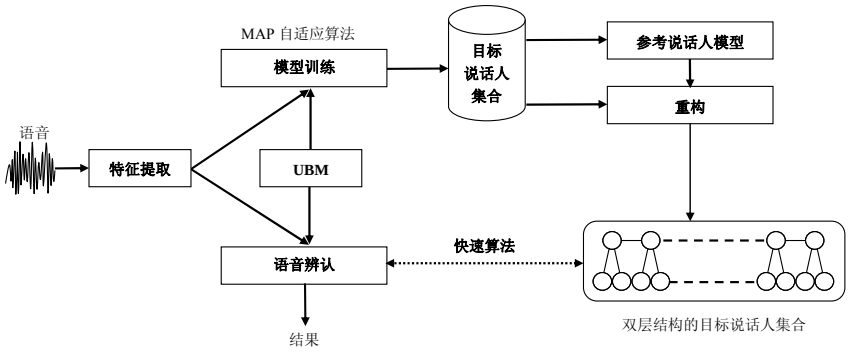

图 1. 算法流程示意图

1 基于参考说话人模型的剪枝算法

图 2 是挑选与待辨认语音相似的目标说话人的示意图，三角形符号代表参考说话人模型（矢量量化后的聚类中心），空心圆形代表目标说话人模型，实线代表目标说话人的偏差，实心圆形代表待辨认语音，虚线代表待辨认语音的偏差，所有目标说话人的偏差均在训练阶段计算，利用偏差确定待辨认语音与哪些目标说话人最相似。算法如下：

1．目标说话人模型是从 UBM 上利用 MAP 算法[6]自适应均值得到。

2．以所有目标说话人模型作为聚类样本，利用最小最大方法从目标说话人模型集合中挑选 K 个模型作为聚类的初始中心。

3．利用 K-means[18-19]算法将目标说话人集合聚成 K 类，得到的 K 个聚类中心作为参考说话人模型。

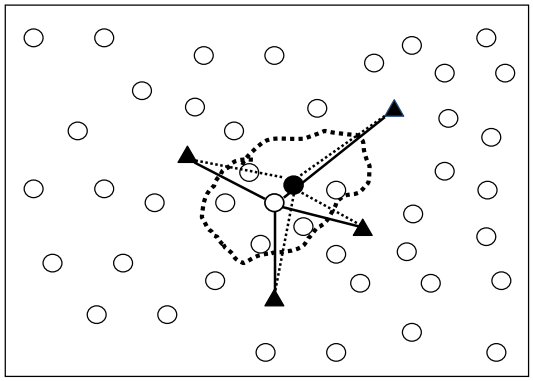

图 2. 挑选相似目标说话人示意图

4．计算目标说话人语音对 K 个参考说话人模型的似然分作为偏差，利用这 K 个偏差形成一个偏差矢量 V_U。

5．计算待辨认语音与 K 个参考说话人的偏差得到偏差矢量 V_F。

6．计算 V_F 与所有 V_U 的相关性并按照大小进行排序，挑选出前 L 个目标说话人。

7．计算待辨认语音对 6 中得到 L 个目标说话人模型的似然分，根据得分辨认说话人。

聚类时高斯混合模型之间的距离度量采用 KL 距离[11]。对于两个协方差是对角矩阵的单高斯分布 $g_m \sim N(\mu_m, \Sigma_m)$和 $g_n \sim N(\mu_n, \Sigma_n)$，其 KL 距离使用公式（1）计算：

$$d(g_m, g_n) = \sum_{i=1}^{D} \left(\frac{\left(\sigma_{mi}^2 - \sigma_{ni}^2\right) + \left(\mu_m^i - \mu_n^i\right)^2}{\sigma_{ni}^2} + \frac{\left(\sigma_{ni}^2 - \sigma_{mi}^2\right) + \left(\mu_m^i - \mu_n^i\right)^2}{\sigma_{mi}^2} \right) \quad (1)$$

其中，D 为特征向量的维数；对于两个混合数为 M 高斯混合模型 λ_1 和 λ_2，其 KL 距离按公式（2）计算。

$$KL(\lambda_1, \lambda_2) = \sum_{i=1}^{M} w_i d\left(g_{1i}, g_{2i}\right) \quad (2)$$

g_{1i} 和 g_{2i} 分别是 λ_1 和 λ_2 中的第 i 个单高斯分布，w_i 为第 i 个分布的权重。偏差向量相关性的计算根据公式（5）：

$$V_U = [P(X_U|C_i)]^T,\ i = 1, 2, \ldots, K \quad (3)$$

$$V_F(X_F) = [P(X_F|C_i)]^T, i = 1, 2, \ldots, K \quad (4)$$

$$\rho_{12} = C_{12}/\delta_1\delta_2 \quad (5)$$

C_i 是参考说话人模型，X_U 是目标说话人的语音，V_U 是目标说话人的偏差矢量，X_F 是待辨认语音，$V_F(X_F)$是 X_F 的偏差矢量，δ_1 和 δ_2 分别是 V_U 和 $V_F(X_F)$的标准差，C_{12} 是 V_U 和 $V_F(X_F)$的协方差。两段语音的相关性越大则二者属于同一说话人的可能性越大，反之则越小。

2 双层结构算法

一般来说待辨认语音仅与一小部分参考说话人模型的距离较小而与大部分参考说话人模型的距离较大，这些较大的距离对于待辨认语音和目标说话人的区分能力较弱，计算相似程度还会造成一定的精度影响同时增加计算时间消耗。同时，待辨认语音与目标说话人集合中的大多数差异较大。基于此设计了一种双层辨认结构，上层用来对参考说话人模型进行快速挑选并对目标说话人进行粗剪枝，下层用来对目标说话人进行精确减枝和辨认。算法如下：

1．首先使用上一小节中的方法训练 K_D 个下层参考说话人模型，记做 DRSM；使用这 K_D 个参考说话人模型作为输入训练出 K_U 个上层参考说话人模型，记做 URSM。

2．计算目标说话人与 DRSM 的偏差向量 V_T^D；计算目标说话人与 URSM 的偏差向量 V_T^U；计算 DRSM 与 URSM 的偏差向量 V_D^U。

3．计算待辨认语音与 K_U 个 URSM 的似然分，并拼接成 K_U 维的上层偏差向量 V_I^U。

4．计算 V_I^U 与所有 V_D^U 的相关性并按照大小进行降序排列，挑选出前 J 个 DRSM 并记录索引，并根据这 J 个 DRSM 的索引从 V_T^D 抽取对应的子

偏差向量V_T^{DJ}；计算V_I^U与所有V_T^U的相关性并按照大小进行降序排列，挑选出前R个目标说话人；

5．计算待辨认语音与步骤 4 中挑选的 J 个 DRSM 的偏差得到 J 维的下层偏差矢量V_I^{DJ}；

6．从计算V_I^{DJ}与步骤 4 中挑选的 R 个目标说话人的V_T^{DJ}的相关性，并按照大小进行降序排列，挑选出相关性最大的前 L 个目标说话人；

7．计算待辨认语音在步骤 6 中得到的 L 个目标说话人模型上的似然分，根据得分高低辨认说话人。

3 说话人辨认实验

3.1 实验设置与数据库

实验数据来自 CCC-VPR2C2005-6000[20]，语音是在电话信道下录制，采样频率 8KHz，采样精度 8 位，单声道录音，选择 5,200 个说话人作为目标说话人，1,000 个说话人作为集外说话人组成实验数据集。目标说话人的训练语音为 30 秒，每个说话人有 2 条待辨认语音。

实验中的系统基于 GMM-UBM，基线系统的 UBM 和说话人模型的混合数都选择 1,024，UBM 采用文献[12]中的树形结构，每帧挑选核心分布仅需计算 100 个单高斯分布的似然分。

语音特征采用 MFCC[21]，帧宽 20*ms*，帧移 10*ms*。对每一帧语音数据，其预加重系数为 0.97，经过 Hamming 窗后，使用基于能量的方法去除了部分静音数据；MFCC 的提取使用的 Mel 三角滤波器组的个数为 30，滤波器组等带宽，中心频率等间隔分布，提取 16 维特征系数及其一阶差分共 32 维特征系数，最后对特征进行 CMS-CVN[22]后得到最后的语音特征序列。

3.2 实验结果与分析

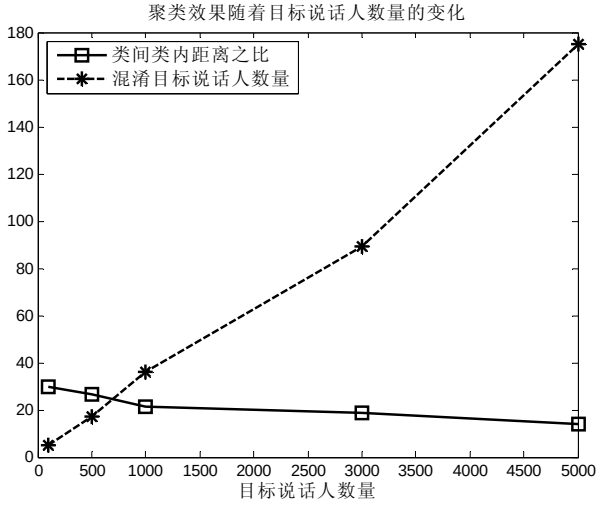

图 4. 聚类效果与目标说话人数量曲线

实验中聚类用的说话人是从实验数据库 6,200 个说话人中随机抽取而得到的子集，聚类数是根据目标说话人数量选择多个聚类数然后从聚类结果中选取最优的类数，评价聚类效果的参数是类间平均距离与类内平均距离之比，比值越大则分类效果好反之则越差。图中数据是多次实验的平均值。混淆目标说话人数量是聚类之后类间容易混淆的目标说话人，计算目标说话人与 K 个类中心的距离并排序，如果最小的 3 个距离值之间的差异较小则该目标说话人定义为混淆目标说话人。

从图 4 中可以看到，随着目标说话人数量的增多，类间区分性在降低，混淆目标说话人的数量及比例在增加。这就说明随着目标说话人数量的增多类间的混淆程度在增加，聚类效果在变差，对 SMC 和 HSI 方法的影响也会随之加大。

在表 1 中，K 是参考说话人数量，L 是挑选到相似目标说话人数量，*Top*-3 是前三选正确率，在后续实验中 K 选择 256，L 择 300；*Factor* 是算法的加速因子，是基线系统的运行时间与实验系统的运行时间之比。

表 1. 参考说话人数量实验

K	*L*	*Factor*	*Top*-3(%)
64	100	31.7	90.3
	300	14.3	93.5
	500	9.2	93.8
128	100	22.5	91.2
	300	12.1	93.9
	500	8.3	94.1
256	100	14.2	91.9
	300	**9.1**	**94.7**
	500	6.5	94.8
512	100	8.5	92.2
	300	6.4	94.8
	500	5.1	94.9

表 2. 剪枝性能比较

算法	*L*	*Factor*	*AE*	*Top*-3(%)
SMC	50	17.0	0.40	88.4
	100	14.6	0.72	89.3
	300	**9.4**	**0.89**	**92.9**
	500	6.9	0.92	93.2
RSMSP	50	16.7	0.26	90.1
	100	14.2	0.39	91.9
	300	**9.1**	**0.56**	**94.7**
	500	6.5	0.79	94.8

在表 2 中，*AE* 表示平均偏差，描述保留的目标说话人与真实的目标说话人的平均相似程度。使用保留目标说话人模型与真实目标说话人模型之间的 KL 距离来描述二者之间的差异程度，*AE* 为所有保留目标说话人与真实目标说话人 KL 距离的平均值，*AE* 越小说明保留目标说话人当中与真实目标

说话人相似程度高的越多，那么辨认性能自然也就会好。SMC 方法的聚类数为 256。RSMSP 与 SMC 相比，在保留相同数量的目标说话人时，更多与真实目标说话人接近的目标说话人保留。

表 3. 算法性能比较

算法	*Factor*	*Top*-3(%)
Baseline	1.0	94.9
SMC	9.4	93.2
RSMSP	9.1	94.7
RSMSP +TL	**12.5**	**94.6**

RSMSP 算法与 SMC 算法相比较，由于增加了偏差向量相关性的运算其运算速度会有所下降，但辨认性能会更接近于基线系统。双层结构实验中 K_D 取 256，K_U 取 32，J 取 64，R 取 1,000，L_1 取 1,000，L_2 取 300，双层结构的运算速度较 SMC 有明显提高，辨认性能较基线系统下降了 0.3%。

4 结论与展望

本文提出一种应用于说话人辨认的基于双层结构和参考说话人模型的目标说话人剪枝算法，能够快速挑选与待辨认语音相似的目标说话人，可在一定程度上降低大规模目标说话人对于目标说话人聚类剪枝算法 HSI 和 SMC 的性能的影响，对包含 5,200 个目标说话人和 1,000 个集外说话人的测试集进行的开集辨认的实验结果表明：推荐算法在提高运算效率 12.5 倍的同时而识别率变化仅下降 0.3%。未来需继续研究快速挑选相似模型的方法，将特征剪枝重排序算法应用于双层结构的粗糙模型层来进一步提升运算速度。

参 考 文 献

[1] J. Campbell, Speaker recognition: a tutorial, Proc. IEEE, 1997, Vol. 85 (9), pp: 1437-1462.

[2] Reynolds D A, Quatieri T, Dunn R. Speaker Verification using Adapted Gaussian Mixture Models [J]. Digital Signal Processing, 2000, 10: 19-41

[3] W. M. Campbell, D. E. Sturim, D. A. Reynolds and A. Solomonoff. SVM based speaker verification using a GMM supervector kernel and NAP variability compensation. ICASSP, 2006, pp: 97-100

[4] P. Kenny, P. Ouellet, N. Dehak et al. A Study of Inter-Speaker Variability in Speaker Verification [A]. IEEE Trans. on Audio, Speech, and Language Processing, 2008, Vol. 16, No. 5, pp 980-988

[5] V. R. Apsingekar and P. L. De Leon, Speaker Model Clustering for Efficient Speaker Identification in Large Population Applications, IEEE Trans. on Audio, Speech, and Language Processing, 2009, Vol. 17, NO. 4, pp: 848-853

[6] J. L. Gauvain, and C.-H. Lee. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains, IEEE Trans. Speech Audio Process, 2, 1994, 291–298

[7] C. J. Leggetter and P. C. Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. Computer Speech and Language. 1995, 9:171-185

[8] Z.-Y. Xiong, T. F. Zheng, Z.-J. Song, F. Soong, W.-H. Wu, A tree-based kernel selection approach to efficient Gaussian mixture model–universal background model based speaker identification, Speech Communication, 2006, Vol. 48, pp: 1273-1282

[9] G. Wang, X.-J. Wu, T. F. Zheng, L.-L. Wang and C.H. Zhang, Regression-class Tree based Method for Efficient Speaker Identification, APSIPA ASC, 2010, pp: 462-465

[10] R. Auckenthaler and J. Mason, Gaussian selection applied to text-independent speaker verification. In Proc. A Speaker Odyssey—Speaker Recognition Workshop, 2001

[11] B. Xiang and T. Berger, Efficient text-independent speaker verification with structural Gaussian mixture models and neural network. IEEE Trans. Speech Audio Process. 2003. Vol. 11 (5), 447–456

[12] Z.-Y. Xiong, F. Zheng, Z.-J. Song and W.-H. Wu, Tree-structure universal background model based efficient speaker identification, Journal of Tsinghua University (Sci. & Tech.), 2006, Vol. 46, No. 7, pp: 1305-1308. (In Chinese)

[13] Z.-Y. Xiong, T. F. Zheng, Z. j. Song, W. h. Wu. Combining Selection Tree with Observation Reordering Pruning for Efficient Speaker Identification Using GMM-UBM. ICASSP, 2005, pp: 625-628

[14] T. Kinnunen, E. Karpov, and P. Franti, “Real-time speaker identification and verification,” IEEE Trans. Audio, Speech, Lang. Process., 2006, Vol. 14, No. 1, pp. 277–288

[15] Pellom and J.H.L. Hansen. An efficient scoring algorithm for gaussian mixture model based speaker identification. IEEE Signal Processing Letters, 1998, 5(11):281–284

[16] B. Sun, W. Liu and Q. Zhong, “Hierarchical speaker identification using speaker clustering,” in Int. Conf. Natural Lang. Process. Knowledge Eng., 2003, pp. 299–304

[17] Z.-Q. Bian and X.-G. Zhang, Pattern Recognition. Tsinghua University Press, 2000. (In Chinese)

[18] A. V. Hall. Methods for demonstrating resemblance in taxonomy and ecology, Nature, 1967, Vol. 214, pp: 830-831

[19] X.-D. Huang, A. Acero, H. Hon, 2001. Spoken Language Processing: A Guide to Theory, Algorithm, and System Development. Prentice-Hall.

[20] Chinese Corpus Consortium, Online available, http://www.CCCForum.org/

[21] Bingxi W, Dan Q, Xuan P. Practial fundamentals of speech recognition, National Defense Industry Press. (In Chinese)

[22] Jing D. Studies on multi-Speaker recognition over telephone, Phd thesis, Beijing, Department of Computer Science and Technology, Tsinghua University, 2007. (In Chinese)