基于 NIST 评测的说话人分类及定位技术研究

杨 毅* 宋 辉 刘 加
(清华大学电子工程系 北京 100084)

摘 要：该文针对美国国家标准与技术研究院(NIST)的 NIST 评测，构建了一套多距离麦克风说话人分类及定位语音处理系统，针对 NIST 富标注评测中提出的说话人分类问题，提出改进的结合时延估计和聚类的说话人分类方法，在保证稳定性的前提下降低说话人分类的复杂度并提高准确率；提出一种新的相邻阵元间时延构造矩阵方程算法，可得到多个说话人的方向角。实验在标准会议环境下采集真实语音数据进行算法验证，说话人分类算法的正确率接近目前主要说话人分类系统的正确率，定位方向角误差在 3° 以内。实验结果说明，适当条件下多距离麦克风系统可作为合适的语音信号输入设备应用于多人多方会议环境。

关键词：说话人分类；说话人定位；多距离麦克风；时延聚类；时延矩阵
中图分类号：TN912.3 **文献标识码**：A **文章编号**：1009-5896(2011)05-1234-04
DOI: 10.3724/SP.J.1146.2010.00977

Speaker Diarization and Localization Technology Research Based on NIST Evaluation

Yang Yi Song Hui Liu Jia
(*Department of Electronic Engineering, Tsinghua University, Beijing* 100084, *China*)

Abstract: This paper builds one speaker diarization and localization speech processing system based on Multiple Distance Microphone (MDM) for NIST evaluation, and proposes a modified clustering algorithm based on time delay estimation, which can decrease the complexity of speaker diarization and improve the correct rate under the guarantee of stable performance. A new time delay matrix structure is proposed, which can acquire multiple speakers' direction angle. It is the real speech data collected under the standard session environment to validate the algorithms. The correct rate of proposed speaker diarization algorithm is similar with other speaker diarization system existed; Location algorithm direction angle error is less than 3°. The results show that under appropriate conditions, the MDM system can be a better input device applied to multiple dialogue scenes.

Key words: Speaker diarization; Speaker localization; Multiple Distance Microphone (MDM); Time delay clustering; Time delay matrix

1 引言

语音识别(speech recognition)也被称为自动语音识别(Automatic Speech Recognition, ASR)，其目标是将人类语音中的词汇内容转换为计算机可读的输入，例如二进制编码或者字符序列。国际上目前最主要的语音识别评测是由美国国家标准局(National Institute of Standards and Technology, NIST)举办的富标注[1](Rich Transcription, RT) 评测，富标注的目的是产生可被人理解且对进一步处理有用的记录来推动识别技术。说话人分类(Speech Diarization, SD)评测于 2005 年首次进入 RT 评测。说话人分类评测使用广播新闻、电话语音交谈、会议室语音等不同场景中的真实录音材料作为测试数据。SD 评测中，参评系统需要界定录音中说话人的时段，并在此基础上标注转写结果中的说话人，完成“who spoke when”的任务，SD 评测对以上任务均有相应的评价方法。除了 RT 评测外 NIST 还组织其它与语音相关的评测，从中衡量参评系统的性能。

2009 年 SD 评测[2]包括 SPKR(“who spoke when”), STT(Speech To Text), SASTT(Speaker Attributed Speech To Text)3 个部分。SPKR 评测是 SD 评测中的一个重要子任务，其目的是将声音

2010-09-07 收到，2010-12-16 改回
国家自然科学基金委员会与香港研究资助局联合科研基金(60931160443)，国家 863 计划项目(2008AA040201, 2008AA02Z414)和国家自然科学基金(90920302, 61005019)资助课题
*通信作者：杨毅 yangyy@mail.tsinghua.edu.cn

数据分成片段(segment)并按照不同说话人来分类。SPKR 的主要输入设备为包括单麦克风、麦克风阵列及多距离麦克风(Multiple Distant Microphone, MDM)[1]。系统首先对目标数据进行噪声消除，随后进行说话人定位和波束形成、以及语音检测和聚类。

传统的单麦克风具有体积小、价格低廉等优点，但不具备对环境噪声处理以及声源定位的能力；麦克风阵列由多个按照特定几何位置摆放的全向麦克风组成，对空间信号进行时空域联合处理，其能力包括：混响条件下的声源定位、增强语音信号、辨识与分离声源等[3,4]，但麦克风阵列系统算法对各个设备之间采样的误差敏感，因此对音频数据的同步性要求十分严格；多距离麦克风[2]是由多个单麦克风组成的信号输入系统，各个麦克风由不同设备控制(如 PDA 或其他便携设备等)，对麦克风的排列和间距没有任何限制，因此麦克风采集的信号不同步，因其结构简单、使用方便和节约成本的优势，在多人多方对话环境中得到广泛采用。

2　多距离麦克风系统说话人分类算法

说话人分类的目的是解决“将声音数据分成片段按说话人来分类(segment an audio recording into speaker homogeneous regions)”的问题。NIST 说话人分类评测条件为：话者个数未知、麦克风位置未知、房间声学环境未知[2]。即在时间和空间先验信息均缺失的场景下判断多个说话人的身份并对声音数据按说话人身份进行分类。普遍多人多方会议场景中说话人和麦克风的相对空间位置固定不变，并且说话人时空域位置不重叠，因此可采用对不同说话人建立聚类模型[5]的算法，将声音数据片段按不同说话人分类，此外还需排除来自静音段和过低信噪比段的估计值来降低误差。

本文时延估计算法采用 PHAT 加权算法。PHAT 加权在混响下可得到比传统广义互相关算法更尖锐的峰值[6,7]，其加权函数为

$$W(\omega)=\frac{1}{\left|X_1(\omega)X_2^*(\omega)\right|} \tag{1}$$

时延估计为

$$\tau=\arg\max R_{x_1x_2}(n) \tag{2}$$

其中 $R_{x_1x_2}(n)$ 为信号 $X_1(\omega)$ 和 $X_2(\omega)$ 之间的互相关函数。

针对每个麦克风采集到的语音信号的每一帧都得到相应的时延估计，需要对时延估计进行优化处理，来避免多说话人场景下时延跳变引起的信号不连续，降低错位率。假设所有说话人和麦克风的相对空间位置固定不变，说话人时空域位置不重叠。定义有 J 个说话人和 N 个麦克风，存在 $J\cdot[N\cdot(N-1)]/2$ 个时延，如式(3)所示。

$$\left.\begin{aligned}\boldsymbol{\tau}_{s_1}&=[\tau_{s_1,m_0,m_1}\quad \tau_{s_1,m_1,m_2}\quad \cdots\quad \tau_{s_1,m_{N-1},m_N}]^{\mathrm{T}}\\&\vdots\\\boldsymbol{\tau}_{s_J}&=[\tau_{s_J,m_0,m_1}\quad \tau_{s_J,m_1,m_2}\quad \cdots\quad \tau_{s_J,m_{N-1},m_N}]^{\mathrm{T}}\end{aligned}\right\} \tag{3}$$

$\boldsymbol{\tau}_{s_j}$ 代表第 j 个说话人的时延向量，τ_{s_j,m_i,m_j} 代表第 j 个说话人到麦克风 m_i 和 m_j 的时延。当声源个数大于 1 时，需要对声音数据片段进行说话人分类，对 J 个说话人采用聚类算法。聚类算法是将数据集划分为若干个子集的过程，并使得同一集合内的数据对象有比较相近的特性，而不同集合中的数据对象特性相差较大。K-means 算法为常用聚类算法，理论简单，计算速度快[8]。K-means 算法采用欧式距离作为相似性评价，其公式为

$$D(X,Y)=\left\{\sum_i|x_i-y_i|^2\right\}^{1/2} \tag{4}$$

K-means 算法随机选取 k 个点作为初始聚类中心，然后根据每个样本到中心的距离对样本归类；重新计算每个类的中心再次聚类，直至平方误差准则函数稳定在最小值。其平方误差准则函数为

$$J_c=\sum_{j=1}^{k}\sum_{i=1}^{n_j}\left\|x_i-m_j\right\|^2 \tag{5}$$

其中 k 为类的个数，n_j 为第 j 类中样本个数，m_j 为第 j 类数据的中心。

本文采用改进的 K-means 算法，主要目的是克服传统算法性能受初值选取及孤立点影响的问题。首先计算各个集合中每个样本的区域密度，选择密度最大的点为初始点，将与之距离最大的点作为下一个初始点，直至选择的初始点个数达到要求；针对孤立点的影响，在每次迭代计算中心时，计算全部样本与中心的距离，根据一定的误差准则(如最小平方误差准则)筛选样本，对满足误差准则的数据求平均值作为新的迭代计算中心，如式(6)所示。

$$\mathrm{Func}=\sum_{j=1}^{J}\sum_{n=1}^{M}\left\|\hat{\boldsymbol{\tau}}[n]-\boldsymbol{\tau}_j\right\|^2 \tag{6}$$

其中 $\left\|\hat{\boldsymbol{\tau}}[n]-\boldsymbol{\tau}_j\right\|^2$ 是每一帧时延估计向量 $\hat{\boldsymbol{\tau}}[n]$ 到簇质心向量 $\boldsymbol{\tau}_j$ 的距离，$\boldsymbol{\tau}_j[n]$ 是簇中 J 个数据的中心向量。此外还需排除来自静音段和过低信噪比段的时延估计值降低误差，如式(7)所示。

$$\tau[n]=\begin{cases}\hat{\boldsymbol{\tau}}[n-1], & \hat{\boldsymbol{\tau}}[n]<\mathrm{Thr}\\ \hat{\boldsymbol{\tau}}[n], & \hat{\boldsymbol{\tau}}[n]\geq\mathrm{Thr}\end{cases} \tag{7}$$

其中 $\tau[n]$ 代表第 n 帧时延估计，Thr 为时延估计阈值。

3 多距离麦克风系统说话人定位算法

说话人定位的目的是解决“语音信号在何地出现”的问题。基于多距离麦克风的说话人定位需要在时间和空间先验信息缺失场景下判断多个说话人的方位。基于麦克风阵列的说话人定位算法在文献[9]中有详细介绍。本文采用矩阵法解决基于多距离麦克风的说话人定位问题。矩阵法通过特征分解得到多距离麦克风结构矩阵的估计，通过对相邻阵元间时延构造矩阵方程并求解，可得到说话人方向角。该方法假设阵元满足空域采样定理的基本要求，具有广泛的适用性和实际应用价值。

假设有 k 个声源，则 N 元多距离麦克风阵在 t 时刻的输出为

$$\boldsymbol{y}(t)=\begin{bmatrix}y_1(t)\\y_2(t)\\\vdots\\y_N(t)\end{bmatrix}=\begin{bmatrix}a_{1,1}&a_{1,2}&\cdots&a_{1,k}\\a_{2,1}&a_{2,2}&\cdots&a_{2,k}\\\vdots&\vdots&\ddots&\vdots\\a_{N,1}&a_{N,2}&\cdots&a_{N,k}\end{bmatrix}\begin{bmatrix}s_1(t)\\s_2(t)\\\vdots\\s_N(t)\end{bmatrix}+\begin{bmatrix}n_1(t)\\n_2(t)\\\vdots\\n_N(t)\end{bmatrix}$$
$$=[a_1\ \ a_2\ \ \cdots\ \ a_k]\boldsymbol{s}(t)+\boldsymbol{n}(t)=\boldsymbol{A}\boldsymbol{s}(t)+\boldsymbol{n}(t)\qquad(8)$$

其中 $\boldsymbol{A}=\begin{bmatrix}a_{1,1}&a_{1,2}&\cdots&a_{1,k}\\a_{2,1}&a_{2,2}&\cdots&a_{2,k}\\\vdots&\vdots&\ddots&\vdots\\a_{N,1}&a_{N,2}&\cdots&a_{N,k}\end{bmatrix}$ 由多距离麦克风结构决定，$\boldsymbol{s}(t)$ 为语音信号向量，$\boldsymbol{n}(t)$ 为噪声信号向量，第 $i(i=1,\cdots,k)$ 个声源信号对第 m 个麦克风有：$\hat{a}_{mi}=k\exp\left[\dfrac{-j2\pi(d_{x,m}\cos\alpha_i\cos\beta_i+d_{y,m}\sin\alpha_i\sin\beta_i)}{\lambda}\right]$。

设 $\boldsymbol{R}\boldsymbol{A}=\boldsymbol{A}\boldsymbol{\Phi}$，其中 $\boldsymbol{R}$ 为声源方向矩阵，$\boldsymbol{\Phi}$ 为对角阵。通过对矩阵 $\boldsymbol{R}$ 进行特征值分解，得到特征值及对应特征向量，可得到对 $\boldsymbol{\Phi}$ 和 $\boldsymbol{A}$ 的估计，由此可以得到声源方向。根据麦克风之间的相位差，则可对第 $i(i=1,\cdots,k)$ 个声源信号构造如下方程：

$$\begin{bmatrix}\dfrac{2\pi}{\lambda}(d_{x,2}-d_{x,1})&\dfrac{2\pi}{\lambda}(d_{y,2}-d_{y,1})\\\vdots&\vdots\\\dfrac{2\pi}{\lambda}(d_{x,N}-d_{x,N-1})&\dfrac{2\pi}{\lambda}(d_{y,N}-d_{y,N-1})\end{bmatrix}\cdot\begin{bmatrix}\cos\alpha_i&\cos\beta_i\\\sin\alpha_i&\sin\beta_i\end{bmatrix}=\begin{bmatrix}\arg\left[\dfrac{\hat{a}_{1,i}}{\hat{a}_{2,i}}\right]\\\vdots\\\arg\left[\dfrac{\hat{a}_{N-1,i}}{\hat{a}_{N,i}}\right]\end{bmatrix}\qquad(9)$$

其中 $d_{x,m}$ 和 $d_{y,m}$ 分别代表阵元在直角坐标系中的横坐标和纵坐标，α_i 和 β_i 分别代表声源信号与 x 轴的夹角和与 z 轴的夹角，$\begin{bmatrix}\cos\alpha_i&\cos\beta_i\\\sin\alpha_i&\sin\beta_i\end{bmatrix}$ 为待求矩阵。设

$$\boldsymbol{z}_i=[\cos\alpha_i\quad\cos\beta_i\quad\sin\alpha_i\quad\sin\beta_i]^{\mathrm{T}}\qquad(10)$$

将式(9)改写为 $\boldsymbol{D}\boldsymbol{z}_i=\boldsymbol{P}_i$，该方程在最小均方准则下 $\boldsymbol{z}_i$ 的最优解为

$$\hat{\boldsymbol{z}}_i=[\boldsymbol{D}^{\mathrm{H}}\boldsymbol{D}]^{-1}\boldsymbol{D}^{\mathrm{H}}\boldsymbol{P}_i\qquad(11)$$

将该最优解代入式(9)解得

$$\left.\begin{aligned}\hat{\alpha}_i&=\arctan[\hat{z}_{2,i}/\hat{z}_{1,i}]\\\hat{\beta}_i&=\arcsin\sqrt{\hat{z}_{1,i}^2+\hat{z}_{2,i}^2}\end{aligned}\right\}\qquad(12)$$

式(12)即为入射信号的 2 维声源方向估计。

4 实验结果及结论

实验中采用多个自带麦克风的笔记本电脑作为信号采集和处理设备，在中型会议室进行安静环境下的数据采集，如图 1 所示为数据采集实验设备及声源空间 2 维位置图，会议室尺寸为 3 m×5 m×3 m，会议混响 RT60 ≈ 1.4 s (根据公式 RT60 = $k(v/\mathrm{Sa})$ 计算，其中 k 为常数，v 为房间体积，Sa 为房间内各个吸收表面的吸收系数总和)。其中笔记本电脑的分布可随机摆放，但实验开始后各自位置固定不变。声源为 2 个不同空间位置的说话人在会议室进行录制，录制过程中声源和麦克风相对位置不变。2 个说话人轮流发言，录制时长 30 min 的多路语音信号，采样率 16 kHz。说话人分类实验的判别率如表 1 所示，该判别率按测试结果属于某类说话人的语音帧数占测试总语音帧数的比例来计算。采用贝叶斯信息标准(BIC)的聚类算法对双说话人进行分类的帧层面分类纯度为 83.64%[10]，本文算法得到的帧层面分类纯度为 86.95%。

说话人方向角的坐标轴设定按照参考文献[11]制定如下规则：选取距离声源最远的参考设备标记为设备 0，其配置的麦克风坐标设为(0,0)；选择距离声源最近的参考设备标记为设备 1，其配置的麦克风坐标设为(x,0)，据此确定坐标原点和 X 轴；与 X 轴垂直的方向则设为 Y 轴；全部方向角的测量在

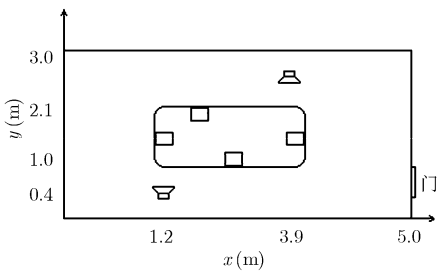

图 1 实验设备及声源摆放位置图

表 1 说话人分类实验判别率

	测试值	
真实值	第 1 类	第 2 类
第 1 类	87.2%	13.3%
第 2 类	12.8%	86.7%

此 2 维坐标系内进行。说话人方向角的实验结果如表 2 所示，声源方向角为两个说话人在坐标轴上的方向角，通过矩阵法估计声源方向角，给出与真实值的误差结果。

表 2 多距离麦克风定位算法性能

声源方向角	MDM 算法	算法误差值
(0°，90°)	(3°，91°)	(3°，1°)
(30°，90°)	(28°，91°)	(2°，1°)
(45°，90°)	(43°，92°)	(2°，2°)
(30°，60°)	(29°，58°)	(1°，2°)

实验表明，本文采用的多距离麦克风说话人分类算法可以快速有效地进行说话人聚类；说话人定位算法的定位误差在 3° 以内。相对于传统的单麦克风及麦克风阵列设备的应用[12]而言，多距离麦克风系统更加适用于多人多方会议场景。进一步研究将集中于混响较大的多方会议环境下如何提取对干扰鲁棒的说话人分类特征及定位算法。

参 考 文 献

[1] Martin A F and Przybocki M A. NIST 2003 language recognition evaluation. National Institute of Standards and Technology, URL: http://www.nist.gov/speech/publications/papersrc/, 2003.

[2] Huijbregts M, Leeuwen D V, and Thomas H. The AMI RT09s Speaker Diarization System. URL: http://www.nist.gov/, 2009.

[3] 许海国. 抗噪声语音识别系统的前端处理[D]. [硕士论文], 北京: 清华大学, 2002.

[4] Ward D B and Williamson R C. Particle filter beamforming for acoustic source location in a reverberant environment [C]. IEEE International Conference on Acoustic Speech Signal Processing. Orlando, USA, 2002: 1777–1780.

[5] Kühne M, Togneri R, and Nordholm S. Robust source localization in reverberant environments based on weighted fuzzy clustering [J]. *IEEE Signal Processing Letters*, 2009, 16(2): 85–88.

[6] Knapp C H and Carter G C. The generalized correlation method for estimation of time delay [J]. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 1976, 24(4): 320–327.

[7] Georgiou P G, Kyriakakis C, and Tsakalides P. Robust time delay estimation for sound source localization in noisy environments[C]. 1997 IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, NY, Oct. 19–22, 1997: 4.

[8] 杨芳, 湛燕, 田学东, 郭宝兰. 使用遗传算法实现 K-means 聚类算法的 K 值选择[J]. 微机发展，2003, 13(1): 25–29.

[9] Huang Yiteng. Real-time acoustic source localization with passive microphone arrays [D]. Georgia Institute of Technology, 2001.

[10] 张薇. 电话语音的多说话人分割聚类研究[J]. 清华大学学报(自然科学版), 2008, 48(4): 574–577.

[11] Raykar Vikas C. Kozintsev Igor, and Lienhart Rainer. Three-dimensional position calibration of audio sensors and actuators on a distributed computing platform [P]. US, Patent 7035757.

[12] 赵贤宇, 王作英. 用于语音识别的鲁棒自适应麦克风阵列算法[J]. 清华大学学报(自然科学版), 2004, 44(10): 1433–1436.

杨　毅：　女，1978 年生，助理研究员，研究方向为多通道语音信号处理.

宋　辉：　男，1982 年生，博士生，研究方向为麦克风阵列.

刘　加：　男，1954 年生，教授，博士生导师，研究方向为语音信号处理.