

基于模糊聚类决策树的分布式语者识别算法

黄继鹏, 陈 志, 芮 路, 王宇虹

(南京邮电大学 计算机学院, 江苏 南京 210023)

摘 要:为解决大规模语者识别问题中普遍存在的加性噪声、高计算复杂度等问题,提高大规模语者识别算法的抗噪性和鲁棒性,利用模糊聚类决策树,提出了一种分布式语者识别算法。该算法将训练数据等分成几个部分,对这几个部分分别使用基于模糊聚类的决策树算法进行训练;对于输入的测试样本,用建好的决策树进行分类,判断它属于哪棵树的哪个叶节点;在该选定的叶节点上使用梅尔频率倒谱系数和高斯混合模型识别方法识别该语者身份。对训练数据进行模糊聚类的过程主要包括四个步骤:根据相应的层提取语音特征;计算特征数据的均值和标准差得到信任间距集合;对集合使用 Lloyd 算法得到分隔向量;以分隔向量为基础进行聚类分组得到下一层的节点。实验结果表明,与传统的硬聚类算法相比,该算法能够提高语者识别的准确率和分类效率,对加性噪声具有良好的抗干扰能力。

关键词:语者识别;模糊聚类;决策树;分布式计算

中图分类号:TP391

文献标识码:A

文章编号:1673-629X(2017)08-0079-04

doi:10.3969/j.issn.1673-629X.2017.08.016

Distributed Speaker Identification Algorithm with Fuzzy Clustering Decision Tree

HUANG Ji-peng, CHEN Zhi, RUI Lu, WANG Yu-hong

(College of Computer, Nanjing University of Posts and Telecommunications, Nanjing 210023, China)

Abstract: In order to solve the problems of additive noise and high computational complexity in speaker identification and to improve the robustness and anti-noise ability of the large scale speaker identification algorithm, a distributed speaker identification algorithm with fuzzy clustering decision tree has been presented, which divides training data into several parts, and builds fuzzy clustering decision trees for these parts. For testing data, fuzzy decision trees has been employed, which are built in the previous step to decide which leaf node the people's speech belongs to. The speaker is identified by using the Mel-Frequency Cepstral Coefficients and the Gauss mixture model identification method on the selected leaf nodes. The process of fuzzy clustering on training data mainly includes four parts, i. e. extracting feature data from the corresponding layer, calculating the mean and standard deviation of the feature data, using Lloyd algorithm to get the separation vector, clustering to get the nodes of the next layer. The experimental result shows that compared with the traditional hard clustering algorithm, the proposed algorithm has improved the accuracy and classification efficiency of speaker identification, with the good anti-interference ability to the additive noise.

Key words: speaker identification; fuzzy clustering; decision tree; distributed computing

0 引 言

在语者识别中,给出一个输入语音,要求从系统提供的所有语者中选择一个来确定未知语者的身份^[1],这个过程通常用到梅尔频率倒谱系数、高斯混合模型

等方法^[2-4]。上述方法在低噪声条件下对小型语者表现非常好,但是在高噪声条件下会严重地降低识别性能,并且当语者数量明显增时,识别错误的可能性将大大增加^[5]。Kenny 等提出基于 I 向量的语音识别和语

收稿日期:2016-04-18

修回日期:2016-08-03

网络出版时间:2017-06-05

基金项目:国家自然科学基金资助项目(61501253);江苏省“六大人才高峰”第十一批高层次人才选拔培养资助项目(XXRJ-009);江苏省基础研究计划(自然科学基金)项目(BK20131382, BK20151506);江苏省重点研发计划(社会发展)项目(BE2016778);江苏省高等教育教学改革“重中之重”立项研究课题(2013JSJG005);国家级大学生创新创业训练计划项目(201410293011, 201510293014);江苏省高等学校大学生创新创业训练计划立项项目(201410293011Z、201510293014Z);南京邮电大学大学生创新训练计划立项项目(SZDG2014011, SZDG2015014, XYB2015036, XYB2015265)

作者简介:黄继鹏(1994-),男,研究方向为数据挖掘;陈 志,副教授,通信作者,CCF 会员(200014587M),研究方向为数据挖掘、传感器网络。
网络出版地址: <http://www.cnki.net/kcms/detail/61.1450.TP.20170605.1506.010.html>

者身份确认方法,该方法通常需要大量表现良好的数据,但当把 I 向量应用于大规模语者识别时,计算复杂度可能会很高^[6]。Graves 等提出使用递归神经网络进行语者识别的方法,在噪声较小的情况下,该方法准确率较高,但对于高噪声的数据效果不尽如人意,并且使用深度网络在数据量较大时计算复杂度很高^[7]。Chorowski 等研究了使用机器翻译模型进行语音识别错误率较高的原因,提出了一种基于注意力模型的方法该方法可以防止单帧过于集中,提高了语音识别的鲁棒性和抗噪声性,但计算复杂度较高^[8]。

可见,普遍存在的加性噪声和大规模语者识别的应用需求使得语者识别算法设计需要解决以下问题:

(1) 语音数据的噪声可能会导致训练和测试之间的错误匹配,降低语者识别的正确性。

(2) 当识别的语者数量显著增加时,识别的错误率也将可能提高。在传统的梅尔频率倒谱系数和高斯混合模型的方法中,语者规模增长时,识别准确性持续减小,当规模增长到 100 时,准确性发生最大下降;与 10 个语者的案例相比,630 个语者大约有 30% 的准确性损失^[4]。

(3) 当语者规模变大时,识别效率随之下降,而计算复杂性相应提高^[9]。

良好的语者识别算法需要有较强的鲁棒性和抗噪性,能够克服大规模语者识别错误率高、计算复杂度高等问题。为此,结合决策树^[10]和模糊聚类^[11],设计了一种分布式语者识别算法,以解决上述问题。

1 分布式语者识别框架设计

基于模糊聚类决策树的分布式语者识别包括四个过程:将训练数据等分成三个部分;对等分过的数据分别使用基于模糊聚类的决策树分类;决定测试语者属于哪棵树的哪个叶节点;对该选定的叶节点使用梅尔频率倒谱系数和高斯混合模型识别技术识别该语者身份。在决策树的每一层的建树过程中采用模糊聚类,即在每一层上一个语者可能属于多个节点。图 1 给出了基于模糊聚类决策树的分布式语者识别框架。

- 1: 输入样本 S
- 2: 将样本分为 3 份, $s_1, s_2, s_3 \in S$
- 3: for all $s_i \in S$
- 4: 使用模糊聚类算法建成决策树 t_i
- 5: 输入测试语音
- 6: 决定该语者属于哪个叶节点
- 7: 使用 MFCC+GMM 算法进行识别
- 8: 输出语者身份 9: end for

图 1 基于模糊聚类决策树的分布式语者识别框架

在该框架中,首先输入样本 S , 将其分为 s_1, s_2, s_3 用于分布式处理;对每个样本,使用模糊聚类算法建成一棵决策树 t_i , 将大量的语者分类到不同的叶子节点,缩小语者规模;对于输入的测试语音,先判断该语者属于哪个叶节点,再在所属叶节点的人群中使用梅尔频率倒谱系数和高斯混合模型的算法进行识别,最后输出语者身份。上述分布式策略用于降低计算复杂度,而用模糊聚类决策树分类将目标语者缩小,制造出传统语者识别方法适合的语者规模,以提高对加性噪声的抗性和识别精确性。

此外,在分布式语者识别框架中,决策树建树过程会从语者语音信号提取语音的音调、语音信号正脉冲的均值、语音信号正脉冲的偏度、语音信号负脉冲的均值、语音信号负脉冲的偏度、语音信号正脉冲的宽度等六种特征,决策树每一层将提取一个特征。给定一个连续的语音输入,使用 YIN 算法^[12]将语音分解成等长的 N_F 帧,一帧的长度是 25 ms,帧移位长度是 10 ms。在语者语音信号提取中,音调的提取方法为:获得第 i 帧的音调值 P_i 和有声概率 P_{ri} ($i = 0, 1, \dots, N_F$);去掉 50 ~ 550 Hz 范围之外的音调值,同时去掉从有声概率低于 0.8 的帧中提取的音调值;得到音调值的集合。其他五个语音特征的提取方法:计算每一帧的能量 E_i 和过零率 Z_i , 并判断该帧是否有声,若不是,不进行操作,若是,则用 Levinson - Durbin 算法^[13-15]计算线性预测系数;通过使用线性预测系数得到线性预测剩余信号;从 LP 剩余信号中提取五个声源特征。

提取的特征表示为 $F_{i,j}$, i 是当前节点上的语者索引, j 是特征索引, $j = 0, 1, \dots, N_i$, N_i 表示语者 i 的特征值总数。

2 基于模糊聚类决策树的分布式语者识别算法设计

根据分布式语者识别框架,从节点模糊聚类分类、识别语者身份两个方面,设计基于模糊聚类决策树的分布式语者识别算法。

2.1 节点模糊聚类分类

在图 1 的框架中,语音样本数据等分成三个部分,这些等分后的语音样本数据分别作为一棵决策树的根节点 C_1 进行建树;每一个语音样本都来自不同的语者, $C_{n_1, n_2, \dots, n_{L+1}}$ 表示 L 层的第 n_{L+1} 个节点。此外,使用基于模糊聚类的决策树对等分过的数据进行分类,对已建好的一棵树,先对根节点进行分组,得到的子节点执行相同的步骤继续进行分组直到建成决策树。图 2 给出了一个节点的模糊聚类分类过程。

```

1: if 节点上的语者数 > 预设值
2:   提取特征值
3:   计算  $\mu_i, \delta_i$ 
4:   构建信任间距集合  $\{\mu_i - \lambda\delta_i, \mu_i + \lambda\delta_i\}$ 
5:   使用 Lloyd 的算法得到分隔向量  $[P_0, P_1, \dots, P_M]$  和下面组群数  $M$ 
6:   创建  $M$  个子节点
7:   for all 语者  $i \in C_1$ 
8:     索引  $m_i = 1$ 
9:   while  $m_i \neq M + 1$ 
10:    if  $[\mu_i - \lambda\delta_i, \mu_i + \lambda\delta_i] \cap [P_{m-1}, P_m] > 0$ 
11:     语者  $i \in C_{1,m}$ 
12:   end if
13:    $m = m + 1$ 
14: end while
15: end for
16: end if

```

图2 一个节点的模糊聚类分类过程

根据图2,在节点的模糊聚类分类中,首先判断当前节点中的样本数量是否大于预设值,若不大于则该节点为叶节点,不需要再进行分组;然后对节点上的样本进行特征提取,每一层只提取一种特征,依次提取音调、语音信号正脉冲的均值、语音信号正脉冲的偏度、语音信号负脉冲的均值、语音信号负脉冲的偏度和语音信号正脉冲的宽度。

在完成节点样本特征提取后,根据式(1)、式(2)计算每个语者特征数据的平均值和标准差。

$$\mu_i = \frac{\sum_{j=1}^{N_i} F_{i,j}}{N_i} \quad (1)$$

$$\delta_i = \sqrt{\frac{\sum_{j=1}^{N_i} (F_{i,j} - \mu_i)^2}{N_i - 1}} \quad (2)$$

其中, μ_i 为语者 i 的特征数据的平均值; δ_i 为语者 i 的特征数据的标准差; $F_{i,j}$ 为提取出的特征, i 是当前节点上的语者索引, $j(j=0,1,\dots,N_i)$ 是特征索引, N_i 为语者 i 的特征值总数。

构建一个可信任的间距 $[\mu_i - \lambda\delta_i, \mu_i + \lambda\delta_i]$, λ 是一个预定的系数;得到所有语者两个统计数据 $\mu_i \pm \lambda\delta_i$ 的集合 $D = \{\mu_i - \lambda\delta_i, \mu_i + \lambda\delta_i\}$ 。

在获得每个语者特征数据的平均值和标准差后,对集合 D 使用 Lloyd 算法^[16]得到分隔向量 $[P_0, P_1, \dots, P_M]$, M 为 Lloyd 算法采用的语者组的总数。以此分隔向量为基础,创建全部 M 个子节点。对每个语者 $i(i \in C_1)$ 进行分组,判断其属于 $C_{1,m}(m=1,2,\dots,M)$ 中的哪一个, m 初始为 1,若 $[\mu_i - \lambda\delta_i, \mu_i + \lambda\delta_i] \cap [P_{m-1}, P_m]$ 不为空,则语者 i 属于 $C_{1,m}$,如此直到 m 等于

M ,所有语者都分组完毕,所得的 $C_{1,m}$ 即是决策树的下一层。

2.2 识别语者身份

根据基于模糊聚类决策树的分布式语者识别框架,对已建好的三棵决策树,分别同时从决策树的根节点开始,对测试语者进行分类,直到其中一棵树分类完成,即识别出该语者身份。识别过程如图3所示。

```

1: while 该节点不为叶节点
2:   特征值提取
3:   异常值去除
4:   计算特征平均值  $\mu$ 
5:    $m = 1$ 
6:   while ! ( $P_{m-1} \leq \mu \leq P_m$ )
7:      $m = m + 1$ 
8:   end while
9: end while
10: 使用 MFCC+GMM 识别语音身份

```

图3 语者身份识别过程

在图3中,从树的根节点开始,对测试语者进行分类直到结束或找到测试语者属于的叶节点,每个节点上执行的分类步骤相同。首先判断该节点是否为叶节点,若是则对该节点使用梅尔频率倒谱系数和高斯混合模型进行身份识别;对测试语者的语音进行特征提取和异常值去除,且相应的层只提取相应特征,得到特征集合 $\{F_k\}$, $k=1,2,\dots,K$, K 为特征值总数。使用式(3)计算特征值的平均值:

$$\mu = \frac{\sum_{k=1}^K F_k}{K} \quad (3)$$

此后,通过比较平均值和从模糊聚类中的 Lloyd 算法得到的分隔向量 $[P_0, P_1, \dots, P_M]$ 来做分类决定,令 $m=1$,判断是否有 $P_{m-1} \leq \mu \leq P_m$,若不是,则令 $m=m+1$,重复此步骤继续判断直到 m 等于 M ;若是,则该测试语者被分类到子节点且此层分类结束。在比较的基础上,当有且仅有一个节点在 L 层是可用的且 L 层的基于决策树的分类结束时,分类将会从 L 层的该可用节点上以相同的方式继续进行,直到一个叶节点最终可用。最终,对选中的叶节点使用梅尔频率倒谱系数和高斯混合模型识别技术识别测试语者身份。

3 实验结果及分析

实验采用从 www.audible.com 等在线有声读物网站收集的数据,将所有的 mp3 样例以 11.025 kHz 的抽样率转换成 wav 格式,假设这些语者都不相同,共得到 1 300 个语者的语音,每个语音的时长为 20 s,在高斯白噪声下的信噪比为 25 dB。

在实验中,1 300 段语音用来构建一棵六层决策

树,音调特征参数包括构建信任间距的 λ 值和 Lloyd 算法采用的聚类数量;对于另外五个声源特征,除了以上列出的两个参数,还有一个额外的关于异常值去除的比例参数。表 1 给出了决策树每一层所构建信任间距的 λ 值和 Lloyd 算法采用的聚类数量以及每一层语者缩减率和准确率。

表 1 决策树每一层参数和性能指标

层数	特征	聚类数目	λ	语者缩减率/%	准确率/%
1	音调	16	0.8	49.78	98.77
2	正脉冲的均值	32	1.1	72.81	98.4
3	正脉冲的偏度	16	0.55	85.33	97.71
4	负脉冲的均值	16	0.8	92.03	97.48
5	负脉冲的偏度	8	0.85	93.23	97.26
6	正脉冲的宽度	16	0.7	94.75	97.01

实验将 1 300 段语音通过训练的决策树做分类测试,在某一层的分类精确性是以在该层上被分到正确节点的语者所占百分比计算的。为了计算某一层的语者缩减率,给该层的节点加权并且所加权重是由该层上语音被分到正确节点的百分比决定的。例如,1 000 个语者全部被正确的分到树的某一层,其中有 100 个语者被正确地分到该层上语者规模(容量)为 200 的节点上,那么当计算该层所有节点的加权平均语者时,该节点分配的权重是 $100/1\,000=10\%$ 。根据表 1,从树的高层到低层,分类精确性持续下降,语者缩减率持续增加。在 25 dB 的情况下,实验中六层决策树在底层能够获得 97.01% 的分类精确性和 94.75% 的语者缩减率,性能表现很好。

下面比较模糊聚类决策树和传统聚类决策树的性能表现。实验使用与表 1 相同的六个特征构建两棵树,图 4 和图 5 给出了两种不同聚类方法的实验结果。

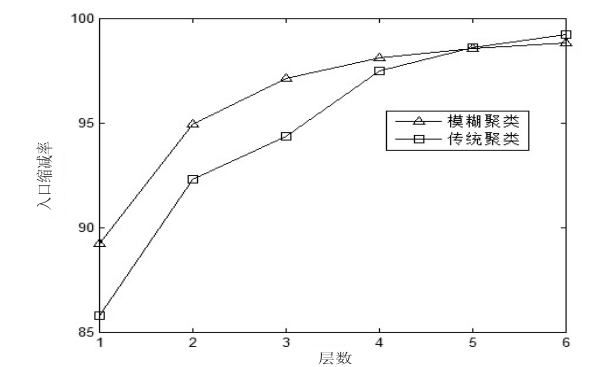


图 4 模糊聚类决策树和传统聚类决策树语者缩减率对比

从图 4 和图 5 可以看出,模糊聚类决策树的分类精确性比传统聚类决策树的精确性高得多,而两棵树的语者缩减率基本相同。可见,模糊聚类对构造所采用的决策树好了传统的硬聚类方法。

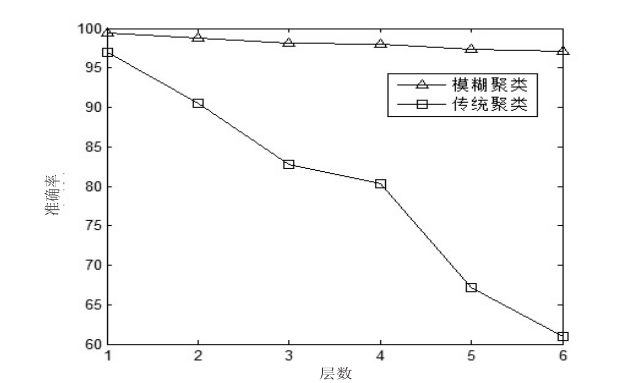


图 5 模糊聚类决策树和传统聚类决策树准确率对比

4 结束语

为解决存在加性噪声的大规模语者识别问题,提出了一种基于模糊聚类决策树的分布式语者识别算法。该算法划分训练数据,使用基于模糊聚类的决策树分别进行分类,通过决定测试语者属于哪棵树的哪个叶节点,缩小识别语者的规模,结合梅尔频率倒谱系数和高斯混合模型来识别未知语者的身份。实验结果表明,利用基于模糊聚类的决策树能够显著提高分类准确率,而分布式建树极大地提高了分类效率,同时对加性噪声有良好的抗干扰力。

在后续研究中,可以考虑对分类算法、语音特征提取算法等进行优化。此外,设计脱离梅尔频率倒谱系数和高斯混合模型的新识别算法,寻找合适的深度学习架构来进行语者身份识别等也是值得探索的方向。

参考文献:

[1] Togneri R, Pullella D. An overview of speaker identification: accuracy and robustness issues[J]. IEEE Circuits and Systems Magazine, 2011, 11(2): 23-61.

[2] Reynolds D A, Rose R C. Robust text-independent speaker identification using Gaussian mixture speaker models[J]. IEEE Transactions on Speech and Audio Processing, 1995, 3(1): 72-83.

[3] Reynolds D A, Quatieri T F, Dunn R B. Speaker verification using adapted Gaussian mixture models[J]. Digital Signal Processing, 2000, 10(1): 19-41.

[4] Reynolds D A. Speaker identification and verification using Gaussian mixture speaker models[J]. Speech Communication, 1995, 17(1): 91-108.

[5] Hasan M R, Jamil M, Rahman M G, et al. Speaker identification using mel frequency cepstral coefficients[C]//Proceedings of the 3rd international conference on electrical & computer engineering. [s. l.]: IEEE, 2004: 565-568.

[6] Kenny P, Boulianne G, Ouellet P, et al. Joint factor analysis versus eigenchannels in speaker recognition[J]. IEEE Trans-

得到分类器,最终实现对人体动作的分类。同时使用文中方法与文献[4]中的算法在 Buffy 数据库上对特定动作进行分类对比。实验结果表明,文中方法具有很高的分类准确率,能够实现对人体运动动作的有效分类。

参考文献:

[1] Wang J Z, Gemen D, Luo J, et al. Real-world image annotation and retrieval; an introduction to the special section[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2008, 30(11):1873-1876.

[2] Leung M K, Yang Y H. First sight; a human body outline labeling system[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1995, 17(4):359-377.

[3] Eichner M, Ferrari V. Better appearance models for pictorial structures[C]//British machine vision conference. [s. l.]: [s. n.], 2009.

[4] Eichner M, Marin-Jimenez M, Zisserman A, et al. 2d articulated human pose estimation and retrieval in (almost) unconstrained still images[J]. International Journal of Computer Vision, 2012, 99(2):190-214.

[5] Kellokumpu V, Pietikäinen M, Heikkilä J. Human activity recognition using sequences of postures[C]//LAPR conference on machine vision applications. Japan: [s. n.], 2005:570-573.

[6] Liu Hong, Zhang Qiaoduo. Human action classification based on sequential bag-of-words model[C]//International conference on robotics and biomimetics. [s. l.]: IEEE, 2014:2280-2285.

[7] 胡 琼,秦 磊,黄庆明. 基于视觉的人体动作识别综述

(上接第 82 页)

actions on Audio, Speech, and Language Processing, 2007, 15(4):1435-1447.

[7] Graves A, Mohamed A, Hinton G. Speech recognition with deep recurrent neural networks[C]//Proceedings of the 2013 IEEE international conference on acoustics, speech and signal processing. [s. l.]: IEEE, 2013:6645-6649.

[8] Chorowski J K, Bahdanau D, Serdyuk D, et al. Attention-based models for speech recognition[C]//Advances in neural information processing systems. [s. l.]: Neural Information Processing Systems Foundation, 2015:577-585.

[9] Hu Y, Wu D, Nucci A. Fuzzy-clustering-based decision tree approach for large population speaker identification[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2013, 21(4):762-774.

[10] Safavian S R, Landgrebe D. A survey of decision tree classifier methodology[J]. IEEE Transactions on Systems, Man, and Cybernetics, 1991, 21(3):660-674.

[11] 孙吉勇,刘力数据,赵连宇. 聚类算法研究[J]. 软件学报,

[J]. 计算机学报, 2013, 36(12):2512-2524.

[8] Sun Qianru, Liu Hong. Learning spatio-temporal co-occurrence correlograms for efficient human action classification[C]//2013 IEEE international conference on image processing. Melbourne: IEEE, 2013:3220-3224.

[9] Xiong Ziyu, Radhakrishnan R. Audio events detection based highlights extraction from baseball, golf and soccer games in a unified framework[C]//International conference on acoustics, speech, and signal processing. Hong Kong: Institute of Electrical and Electronics Engineers Inc, 2003:632-636.

[10] 张学工. 关于统计学习理论与支持向量机[J]. 自动化学报, 2000, 26(1):32-42.

[11] Vapnik V N. Estimation of dependences based on empirical data[M]. New York: Springer-Verlag, 1982.

[12] Cortes C, Vapnik V. Support-vector networks[J]. Machine Learning, 1995, 20(3):273-297.

[13] Hsu C W, Lin C J. A comparison of methods for multi-class support vector machines[J]. IEEE Transactions on Neural Networks, 2002, 13(2):415-425.

[14] 相 洁,陈俊杰. 基于 SVM 的 FMRI 数据分类:一种解码思维的方法[J]. 计算机研究与发展, 2010, 47(2):286-291.

[15] Rother C, Minka T, Grabcut; interactive foreground extraction using iterated graph cuts[J]. ACM Transactions on Graphics, 2004, 23(3):307-312.

[16] Hernándezvela A, Reyes M, Ponce V, et al. GrabCut-based human segmentation in video sequences[J]. Sensors, 2012, 12(11):15376-15393.

[17] Schölkopf B, Platt J, Hofmann T. Learning to parse images of articulated bodies[C]//Conference on advances in neural information processing systems. [s. l.]: [s. n.], 2007:1129-1136.

2008, 19(1):48-61.

[12] de Cheveigné A, Kawahara H. YIN, a fundamental frequency estimator for speech and music[J]. Journal of the Acoustical Society of America, 2002, 111(4):1917-1930.

[13] Musicus B R. Levinson and fast Choleski algorithms for Toeplitz and almost Toeplitz matrices[D]. [s. l.]: Massachusetts Institute of Technology, 1988.

[14] Sim K S, Lim M S, Yeap Z X. Performance of signal-to-noise ratio estimation for scanning electron microscope using autocorrelation Levinson - Durbin recursion model[J]. Journal of Microscopy, 2016, 263(1):64-77.

[15] Selvaperumal S K, Nataraj C, Thiruchelvam V, et al. Speech to text synthesis from video automated subtitling using levinson durbin method of linear predictive coding[J]. International Journal of Applied Engineering Research, 2016, 11(4):2388-2395.

[16] Lloyd S P. Least squares quantization in PCM[J]. IEEE Transactions on Information Theory, 1982, 28(2):129-137.