讲话人识别系统的鼻腔参数研究

李 倩[1]，庄 琳[2]，达 钊[1]，郭霞生[1]，章 东[1]
(1. 近代声学教育部重点实验室，南京大学声学研究所，南京 210093；2. 南京森林警察学院，南京 210046)

摘要：基于鼻腔发声的极零数学模型，结合线性预测、线性预测失真尺度等知识，利用同态预测法从语音中提取出个人的鼻腔参数。用此鼻腔参数尝试区分不同讲话者，并与共振峰，美尔频率倒谱系数的区分效果作比较。实验结果表明：鼻音参数相对于这些简单的时频参数(共振峰，美尔频率倒谱系数等)对于区分不同讲话者更有效，在声纹检测中有潜在的应用价值。

关键词：声纹；鼻音；特征提取；线性预测失真度
中图分类号：TB556 **文献标识码**：A **文章编号**：1000-3630(2012)-03-0291-05
DOI 编码：10.3969/j.issn1000-3630.2012.03.011

Research on the nasal parameters of speaker recognition system

LI Qian[1], ZHUANG Lin[2], DA Zhao[1], GUO Xia-sheng[1], ZHANG Dong[1]
(1. *Key Laboratory of Modern Acoustics*, *Institute of Acoustics*, *Nanjing University*, *Nanjing* 210093, *China*;
2. *Nanjing Forest Police College, Nanjing* 210046, *China*)

Abstract: In this paper, based on the pole-zero model of the nasal sound, relevant parameters have been extracted from the individual's nasal voice by using linear prediction, linear prediction distortion and homomorphism theories. Experimental results show that the nasal parameters are more effective than some simple frequency parameters (such as formants) in distinguishing different speakers, which means that the nasal parameters has potential value in speaker identification.

Key words: voiceprint; nasal sound; feature extraction; linear prediction distortion

0 引 言

近年来，声纹识别技术越来越多地应用于信息、公安司法及银行证券领域，以区分不同人的声音特征。用于声纹识别特征参数有很多，如美尔频率倒谱系数(MFCC)、线性感知系数(PLP)，共振峰、时域能量等[1]。但这些特征对环境变化和说话人差异很敏感[2]，识别率还有待进一步提高。此外，同一个人声音的易变性和声学环境的影响，也极易导致声纹系统的错误识别[3]，近年来，研究者们一直关注着能在声学环境变化中保持稳定的特征参数，如韵律、音素等[2, 4-6]。

鼻腔是声道中唯一不随语音内容而改变形状的部分，鼻音相对其它声音相对稳定[7]。另外感知说话人辨识 PSI 实验证明，不同讲话人之间的鼻音的发音明显不同，即使实验样本改变，利用鼻音识别讲话人的有效性仍比其它的辅音要高[8]，由于英文语音系统与普通话系统的辅音/m/、/n/是相同的[9]，因此可以认为文献[8]的 PSI 实验结果可适用于普通话/m/、/n/鼻音系统。由于鼻音具有稳定、高效地区分不同讲话人的特性，可以用于提取稳定的声纹参数。用带鼻音音节的鼻声母段提取参数来识别讲话者，比之一般识别方法中使用整个音节的语音特性，其运算量要小得多，识别有效性也更高。有研究人员利用极零模型(ARMA)获得所有汉语鼻声母音节的极点和零点系数的谱参数，系统的识别率可高达 87.92%[10,11]。

由于鼻音所具有的上述优点，鼻腔参数的提取方法值得深入地进行探讨和研究。本文提出了一种新的提取鼻腔参数的方法，其与传统的 ARMA 方法的不同之处在于：在鼻腔数字模型的基础上，用同态处理的方法将极零模型转化为线性预测谱模型，再与线性失真尺度的检查相结合。利用该方法，本文探讨了针对不同鼻音(m、n)对不同说话人进行区分的有效性，并与共振峰方法及 MFCC 的识别性能进行了比较与分析。

收稿日期：2011-06-13；**修回日期**：2011-07-26
基金项目：国家自然科学基金(11161120324、10974093)，中央高校基本科研业务费专项资金资助(1103020402、1112020401)
作者简介：李倩(1986－)，女，江苏人，硕士，研究方向为声学。
通讯作者：章东，E-mail: dzhang@nju.edu.cn

1 鼻腔特征参数提取与匹配

1.1 鼻腔数字模型的建立

声道的终端由鼻腔和口腔构成，为了提取鼻腔参数，需要了解鼻腔在发声过程中的作用。发声过程中，鼻腔与口腔的连接情况如图 1 所示。将喉部声带处至口腔看成管道，鼻腔可近似看成管道的旁支，起到滤波作用。对于鼻音节而言，如 ma，先发 m，此时口腔关闭，鼻腔打开；在 m 到 a 过渡段时，口腔与鼻腔同时打开；最后发出 a 音，此时鼻腔关闭。

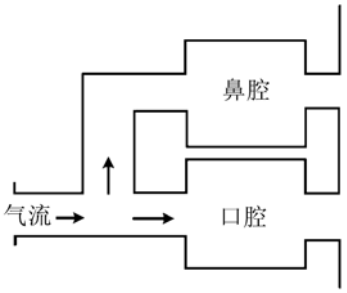


图 1 鼻腔与口腔的连接示意图
Fig.1 Illustration of nasal cavity and oral cavity.

根据上述发音过程，使用一个极零模型来表征鼻腔与口腔的并联。考虑到声门及口唇的辐射，本文采用如下系统函数来表示语音产生过程：

$$H(z)=S(z)/U(z)=G[1+\sum_{i=1}^{d}b_i z^{-1}]/[1+\sum_{i=1}^{m}a_i z^{-1}] \quad (1)$$

式中：$U(z)$为声门信号的 z 变换；$S(z)$为语音信号的 z 变换；$H(z)$为系统函数。如果模型的阶数足够高，理论上可以用全极点模型来近似表示极零模型。但若系统中的零点过多，所需用于近似的极点个数将成倍增长，导致运算量大大增加。此外，与全极点模型相比，极零模型的线性预测谱系数(LPC)分析需要求解非线性方程以找到最优参数。

1.2 鼻腔数字模型参数的提取及个体性识别

基于鼻腔的极零模型，可以用同态预测法及线性预测谱提取模型的特征参数。首先利用同态处理，倒谱的方法，将一个极零模型转变为全极点模型，再利用线性预测谱，对转化后全极模型的变换语音序列提取线性预测谱参数。每个语音样本可提取一组这样的参数，用于建立个人码本。语音谱需要每千赫两个极点(可以是一对共轭极点)来表征声道响应，因此全极模型的分子多项式(零点)和分母多项式(极点)的最佳阶数 Q、P 的选择可根据采样率来确定[11]。对不同人进行识别时，可以比较两个语音(码本)之间的差异；语音差异用线性失真尺度来度量，两个语音差异越大，则失真越大。

在进行失真测试时，仅由线性预测系数的差值不能完全表征两个语音信息的差别，可以直接用由上述系数所描述的信号的功率谱来进行比较，为此采用 I-S（Itakura-Saito)距离：

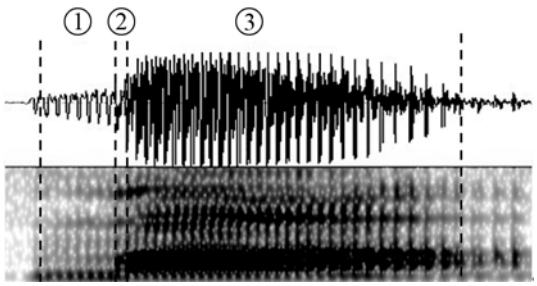
$$d_{IS}(X,Y)=\frac{\boldsymbol{a}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{a}}{\sigma^2}-\ln\sigma_X^2+\ln\sigma^2-1 \quad (2)$$

其中，$\boldsymbol{a}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{a}=r_a(0)r_x(0)+2\sum_{i=1}^{p}r_a(i)r_x(i)$，$\boldsymbol{R}$ 为信号 $x(n)$ 的 p 阶自相关矩阵，$\boldsymbol{a}=[a_1,a_2,\cdots,a_p]$ 为 Y 的线性预测系数，σ_x^2 为 $x(\omega)$的增益，$r_a(i)$为线性预测系数的自相关函数，$r_x(i)$为信号 $x(n)$的自相关函数。

2 结果及讨论

本文研究所用的语音数据取自中国科学院声学研究所语音库 SCSC(汉语普通话单音节语音语料库)，在专业的录音室中录制，用 16 kHz 的声卡采集存储。随机选择 10 个人的鼻音(ma、na)4 遍单音字作分析用。PSI 实验表明，C-V(辅音加元音)结构的单音字中，元音为/o/、/a/的单音字比其它元音的单音字更易于区分，故本文采用 m-a、n-a 的 C-V 结构[8]。

由于鼻音的浊音特性，可在平稳段采用长时分析，得到平均的参数。对 m、n 鼻辅音进行处理，截取声母段原则为，截取鼻音的中间部分，省略掉鼻音的开始部分和过渡到元音的过渡段部分，截取时长 30 ms 的平稳段(鼻音的平均音长为 77 ms)如图 2 所示，辅音、过渡段及元音部分也可从语谱图上清晰地看出。采用 Praat 软件(University of Amsterdam)进行手动截取及分析[12]。对截取的声音段用汉明窗分帧，帧长为 256 样点，相邻两帧之间用二比一交叠。20 ms 内语音可视为平稳，这里一帧为 16 ms，在这时间段内，语音信号的频谱特性和某些物理特征参量可近似地看作不变。根据采样率，全极模型的 P 取 16，Q 取 10，得到 4 遍语音的 LPC 系数。



(1) 声母段 m (2) 辅音至元音的过渡段 (3) 元音段 a
图 2 语音信号 ma 的时域波形截取示意图
Fig.2 Waveform in time domain for the truncated signal ma.

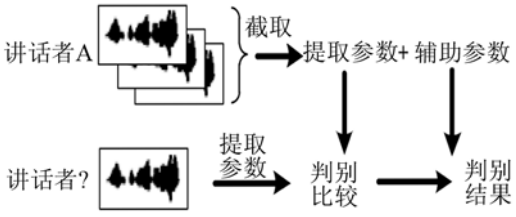


图 3　识别基本原理图

Fig.3　Illustration of speaker recognition.

用前三遍声音得到语音同态分析后的 LPC 系数，根据鼻音的浊音的周期特性进行规整并作为码本。再用第四遍的语音同态分析后的 LPC 系数与码本作失真测量，得到区分结果。整个识别过程如图 3 所示(判别比较过程中可加入辅助识别参数，以提高识别率) 。

2.1　实验结果

这里对 10 位讲话者进行 m 和 n 码本的分析与识别测试，图 4(a)中给出了对于 m 鼻音进行分析比较的结果，白色柱形图表示同一人(intraspeaker)的平均失真差异 d_{xy}，黑色柱形图则代表了不同人(interspeaker)之间的平均差异，其中同一人平均差异数据为该讲话人的三遍语音失真平均值。图 4(b)中给出了对于 n 鼻音的类似分析结果。可见，对于 m 和 n 两类码本，在所选取的样本中，每一个讲话人的同一人的自身差异要比不同人的差异值小。

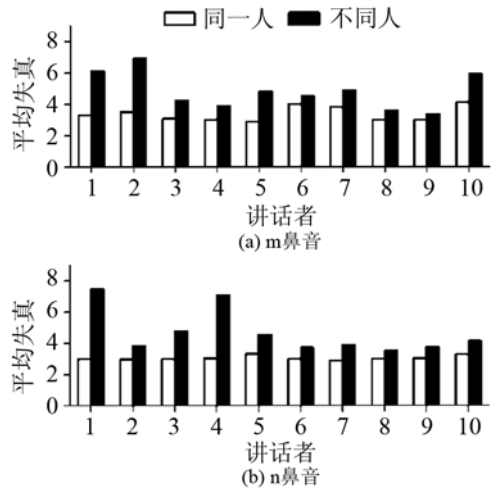


图 4　同一人与不同人的鼻音差异

Fig.4　Comparison of nasal sounds between intraspeaker and interspeaker

表 1　同一人与不同人结果差异的定量比较

Table 1　Quantitative comparison of nasal sounds between intraspeaker and interspeaker

讲话者	1	2	3	4	5
鼻音 m	84.8%	97.2%	37.7%	29.4%	65.5%
鼻音 n	149.3%	28.5%	59.4%	134.2%	37.6%
讲话者	6	7	8	9	10
鼻音 m	12.7%	28.7%	20.7%	12.6%	43.7%
鼻音 n	24.4%	34.0%	16.7%	23.9%	25.6%

为了便于分析比较，需要进一步将识别结果定量化，将图 4 中同一人与不同人的结果之比较((不同人差异–同一人差异)/同一人差异×100%)列入表 1 中。表中数据均为正值表明不同人的差异始终是大于同一人的，由此即可区分是否是同一人。所选取的 10 个讲话人之间 m、n 鼻音最大的差异分别为 97.2%(讲话人 2)和 149.3%(讲话人 1)；而相差较小的分别为 12.6%和 16.7%。可见此方法得到的鼻腔参数用于区分不同讲话人具有较好的敏感度；同时可看出，鼻音 m，n 在区别上有互补的趋势。在鼻音 m 差异较小的情况下，鼻音 n 的差异相对较高；在鼻音 n 差异较小的情况下，鼻音 m 的差异相对较高。因此，结合两者可以达到更好的区分效果。

2.2　讨 论

从上述结果可知，同一人的多遍鼻音数据相比较，差异相对较小，不同人之间的数据差异则相对较大，亦即证明此参数对于区别不同的讲话人是有意义的。为了反映此参数的优劣性，本文还与共振峰参数，MFCC 参数进行了对比。

共振峰反映个人差异体现在频率差异上，共振峰频率本身不能提供幅度信息[1]，但其带宽则能反应出共振峰幅度所对应的能量[13]。由于此前所采用的鼻腔参数 d_{xy} 中有增益项，为了更好地比较，表 2 不仅给出了 10 个讲话人鼻音 m 的前四个共振峰的平均频率(F)，还给出了平均带宽(B)。若用欧氏距离表示共振峰的差异，则可得到同一人的 m 鼻音差异与不同人 m 鼻音差异如图 5 所示。

图 5 表明，共振峰在用于区分不同讲话人时，效果不是很明显。无论是从频率还是带宽考虑，某些样本的同一人差异甚至比不同人的差异要高得多。即使将共振峰的频率与带宽结合起来考虑，也存在着同一人差异比不同人差异要高的情况，如讲话者 10。造成这一结果的原因是：(1) 由于高频区的频率差较大，放大了欧氏距离；(2) 鼻音的语音能量主要集中在低频区，带宽在反映能量的个人差异不是很明显。为了提高共振峰参量的识别效果，需要考虑加入其它参数，如融合谐波能量曲线特征，反映出被共振峰调制的谐波能量变化趋势[1]。再如，可以通过大量的实验得到各个共振峰之间最佳的权重因子，这时要考虑到鼻音的个性特征表现在鼻腔与咽腔的细微结构上，在频谱上个性特征在高频处的体现如何权重的问题[11]。相比于本文提到的方法，这些手段的工作量相对较大，且分析过程繁琐。

表 2 10 人鼻音 m 前四个共振峰平均频率及带宽值

Table 2 Averaged frequency and bandwidth of the resonant peaks (from 1st to 4th) for nasal m from 10 persons

共振峰及带宽	1	2	3	4	5
F1	309.4	362.4	264.2	264.2	279.6
B1	58.7	148.6	34.7	123.9	145.9
F2	2249.6	2391.8	1445.4	1666.6	1580.4
B2	1661.9	29.9	928	413.4	1017
F3	2631.2	3671.4	2389.2	2784	2495.9
B3	66.2	687.5	152.6	67.1	60.2
F4	3473.6	4668.2	3837.9	4179.9	3920.9
B4	585.6	648.3	625.3	862.8	621.8
共振峰及带宽	6	7	8	9	10
F1	291.9	335.8	295.6	289.5	329.3
B1	229.2	75.8	59.3	57.4	117.8
F2	1533.1	2402.3	1056.7	2196.8	2242.1
B2	1481	424.4	664	747.2	57.9
F3	2653.5	3299.3	2755.3	3496.3	3454.9
B3	91.2	450.2	133.4	939.4	342.1
F4	3794.2	3884.7	3393.9	4610.2	4411.8
B4	2678.9	245.5	946	1155.7	1047.9

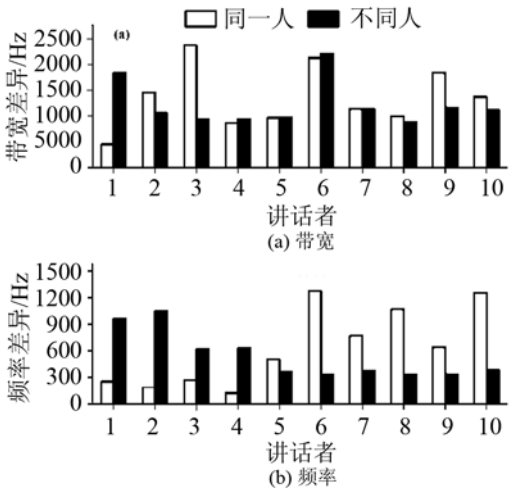


图 5 同一人与不同人的 m 鼻音差异

Fig.5 Comparison of nasal m between intraspeaker and interspeaker:

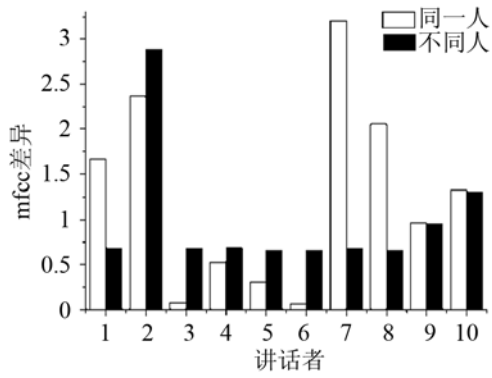


图 6 同一人与不同人的 m 鼻音 MFCC 差异

Fig.6 Comparison of MFCC for nasal m between intraspeaker and interspeaker

若用欧氏距离表示 MFCC 的差异，则可得到同一人的 m 鼻音差异与不同人 m 鼻音 MFCC 差异如图 6 所示。

图 6 表明，MFCC 在用于区分不同讲话人的 m 鼻音时，效果不明显。10 人中有 5 人的同一人差异比不同人差异高，如讲话者 7，同一人差异比不同人的差异要高得多。

为了更好地对比本文提出的方法，采用了直接计算法求解极零模型[10]，对同一人与不同人的极零模型预测系数加权组合距离进行了计算，结果如图 7 所示。

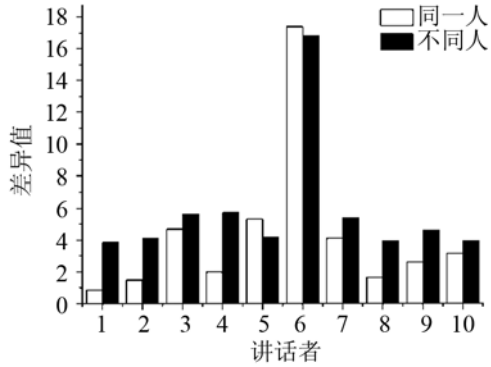


图 7 极零模型预测系数加权组合距离

Fig.7 Predicted distance with weighted coefficient array based on the pole-zero model.

图 7 结果表明，极零模型预测系数加权组合距离对识别不同人效果较好，虽然讲话者 5、6 的同一人距离比不同人距离大，但是两者的差异值并不会像 MFCC 及共振峰参数那样，相距较大。大多数讲话者的同一人距离比不同人距离小，在数据量比较小的情况下，识别效果较好。

表 3 定量比较了本文方法与文献[10]方法识别同一人与不同人的结果差异。同表 1 的分析方法，表中数据为((不同人差异–同一人差异)/同一人差异×100%)的值，负号表示同一人的差异比不同人的差异大。

表 3 鼻音 m 同一人与不同人结果差异的定量比较

Table 3 Quantitative comparison of nasal m between interspeaker and intraspeaker.

讲话者	1	2	3	4	5
本文方法	84.8%	97.2%	37.7%	29.4%	65.5%
文献[10]法	351．0%	176.7%	20.1%	191.0%	-26.4%
讲话者	6	7	8	9	10
本文方法	12.7%	28.7%	20.7%	12.6%	43.7%
文献[10]法	-3.2%	32.0%	140.5%	78.3%	24.9%

表 3 的结果表明，对于讲话者 1、2、4、8、9，文献[10]法的定量差异要比本文的方法好，如讲话者 1，文献[10]法的差异为 351%，比本文方法

84.8%要高出很多。对于讲话者 3、5、6、10，本文方法的差异比文献[10]的方法要好，且不存在同一人差异比不同人差异大的情况。两种方法各有优势，可以做到互补，以保证判定结果的正确性。

据上述分析可知，d_{xy} 的鼻腔参数要比共振峰参数及 MFCC 参数操作简单且识别精度更高。由于 d_{xy} 包含了语音倒谱、LPC 系数及幅度的信息，其所保留的语音信号的独有信息更加完善，且计算过程较为简单。与简单地运用共振峰参数及 MFCC 相比，效果更加好。从另一个方面考虑，声纹参数并不是独立的，利用声纹参数的互补性质，可以提高系统的识别率。

3　结 语

每个人发音时，鼻腔的共振腔各异，鼻辅音有着个人独特的共振特性，且这些共振腔不能随意的改变。讲话者识别中必须要求有较大的不同人差异和较小的同一人差异，且此语音不能被讲话者自己控制，鼻音恰能满足这些条件[8]。因此用鼻音来提取声纹参数有较好的应用前景。本文从鼻音的 ARMA 模型出发，提出了一种鼻腔参数提取方法。此方法结合了极零点模型、同态分析、倒谱、线性预测谱及线性失真尺度等，从鼻音声母 m、n 中提取个人参数。实验结果表明，不同人之间鼻音差异是很明显的；鼻音参数对于区分不同讲话者是有效的。在讨论中可知，鼻音参数相比于一些简单的时频参数(如共振峰、MFCC 等)，在区分讲话者时更加有效，且处理过程简单。鉴于鼻音在识别中的优点，在进行说话人识别时，如果是文本识别，在编辑文本应该考虑多加入鼻音节的字；如果是非文本识别，在处理鼻音语音段中，可采用本文提到的方法提取鼻音参数。后续工作将着重于加大数据量，从而获得更加可靠的数据集和更加完善的结果。另一方面，鼻音的性质可能会受到风寒及喉病的影响，这些因素也应在今后的工作中考虑到[12]。

参 考 文 献

[1] 张建平, 李明. 长时语音特征在说话人识别技术上的应用[J]. 声学学报, 2010, **35**(2): 267-269.
ZHANG Jianping, LI Ming. Long span prosodic features for speaker recognition[J]. Acta Acustica, 2010, **35**(2): 267-269.

[2] 张晴晴, 潘接林, 颜永红. 基于发音特征的汉语普通话语音声学建模[J]. 声学学报, 2010, **35**(2): 254-260.
ZHANG Qingqing, PAN Jielin, YAN Yonghong. Tonal articulatory feature-based acoustic modeling for Chinese Putonghua speech recognition[J]. Acta Acustica, 2010, **35**(2): 254-260.

[3] Zheng N H, Ching P C. Speaker recognition using complementary information from vocal source and vocal tract[M]. The Chinese University of Hong Kong. 2006: 1-8.

[4] 姜晓庆, 田岚崔, 崔国辉. 多语种情感语音的韵律特征分析和情感识别研究[J]. 声学学报, 2006, **31**(3): 217-221.
JIANG Xiaoqing, TIAN Lan, CUI Guohui. Statistical analysis of prosodic parameters and emotion recognition of multilingual speech[J]. Acta Acustica, 2006, **31**(3): 217-221.

[5] Bocklet T, Shriberg E. Speaker recognition using syllable-based constraints for cepstral frame selection[C]// Proc. ICASSP 2009, Taipei, Taiwan, April 2009, 4525-4528.

[6] Baker B, Vogt R, Sridharan S Gaussian mixture modeling of broad phonetic and syllable events for text-independent speaker verification[C]// Proc. Interspeech2005, Lisbon, Portugal, 2005, 2429-2432.

[7] Kitamura T, Honda K Unchanged parts in the vocal tract during the utterance of the vowels[C]// Proc. Gen. Meet. Phonet.Soc. Jpn. 2003, 105-110.

[8] Kanae A, Takayuki A Speaker-dependent characteristics of the nasals[J]. Forensic Science International, 2009, **185**(4): 21-28.

[9] 刘赵杰, 邵健, 张鹏远, 赵庆卫, 颜永红, 冯稷. 汉语自然口语中声调识别的研究[J]. 物理学报, 2007, **56**(12): 7064-7067.
LIU Zhaojie, SHAO Jian, ZHANG Pengyuan, ZHAO Qingwei, YAN Yonghong, FENG Ji. Research on tone recognition in Chinese spontaneous speech[J]. Acta Physics, 2007, **56**(12): 7064-7067.

[10] 刘莹. 一种用鼻音声母识别讲话者的方法[J]. 声学学报, 1995, **20**(3): 232-234.
LIU yin. A method to recognize speaker using nasal[J]. Acta Acustica, 1995, **20**(3): 232-234.

[11] 林宝成, 陈永彬. 基于 ARMA 模型的汉语讲话者识别[J]. 声学学报, 1998, **23**(3): 229-234.
LIN Baocheng, CHEN Yongbin. Recognition of chinese speaker based on ARMA model[J]. Acta Acustica, 1998, **23**(3): 229-234.

[12] Kanae A, Tsutomu S, Takayuki A. Idiosyncrasy of nasal sounds in human speaker identification and their acoustic properties[J]. Acoust. Sci. & Tech, 2006, **27**(4): 233-235.

[13] Jessica E H, Elaine T S, Gina M C. Formants of children, women, and men: The effects of vocal intensity variation[J]. J. Acoust. Soc. Am, 1999, **106**(3): 1532-1542.