第 26 卷 第 4 期
2000 年 7 月

自 动 化 学 报
ACTA AUTOMATICA SINICA

Vol. 26, No. 4
Jul., 2000

基于 Mellin 变换的语音新特征与频率归正说话人自适应技术

陈景东　　徐　波　　黄泰翼

（中国科学院自动化研究所模式识别国家实验室　邮编　100080）

（E-mail：jchen@spl.me.gu.edu.au，xubo@nlpr.ia.ac.cn）

摘　要　为了减小由于说话人之间声道形状的差异而引起的非特定人语音识别系统性能的下降，研究了两种方法，一种是基于最大似然估计的频率归正说话人自适应方法，另一种是基于 Mellin 变换的语音新特征. 在非特定人孤立词语音识别系统上的初步实验表明，这两种方法都可以提高系统对不同说话人的鲁棒性，相比之下，基于 Mellin 变换的语音新特征具有更好的性能，它不仅提高了系统对不同话者的识别性能，而且也使系统对不同话者的误识率的离散程度大大减小.

关键词　Mellin 变换，频率归正，自适应.

SPEAKER NORMALIZATION AND NOVEL ROBUST SPEECH FEATURE BASED ON MELLIN TRANSFORM

CHEN Jingdong　　XU Bo　　HUANG Taiyi

(*National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, P. O. Box* 2728, *Beijing* 100080)

Abstract　One major source of interspeaker variability in speaker-independent (SI) speech recognition is the variation of the vocal tract shape, especially the vocal tract length (VTL) among individual speakers. If the model of the vocal tract is assumed to be a uniform tube with length L, then the formant frequencies of utterances of a given sound are inversely proportional to L. Since the VTL can vary from approximately 13cm for females to over 18cm for males, formant center frequencies can vary by as much as 25% among speakers. This source of variability results in state-of-the-art SI speech recognizers working poorly for outlier speakers whose vocal tract shapes differ significantly from those of speakers in the training set. In an effort to reduce the degradation in speech recognition performance caused by variation of the VTL among speakers, two methods are investigated in this paper. One is to remove the variability with a technique of speaker normalization. Another is to extract new feature based on the Mellin transform (MT). Because of

收稿日期　1998-08-17　　收修改稿日期　2000-01-06

the scale invariance property of the MT, the new feature is insensitive to variation of VTL among different speakers. Experiments show that both methods can improve the performance of an SI recognizer, while the latter approach is more effective than the former one.

Key words　Mellin transform, frequency normalization, adaptation.

1　简介

在基于 HMM 的非特定人语音识别之中，说话人之间的差异主要是声道形状，尤其是声道长度(VTL)的差别[1]. 声道长度随着说话人的年龄和性别等因素的变化而变化. 对于成年女性来说，其声道长度约为 13cm，而对于成年男性，其声道长度则接近于 18cm[2]，儿童的声道长度比成年人的短，例如 8 岁小孩的平均声道长度大约在 10cm 左右. 如果假设声道是长度为 L，横截面积为 1 的均匀管道，可以证明，语音频谱共振峰频率与 L 成反比[3]，因此对成年人来说，发相同音其语音频谱共振峰频率的变化可高达 25%. 大量的研究及实验结果表明，由声道长度差异而引起的共振峰频率的变化是导致一个非特定人语音识别系统对集外人的识别率远远低于其对训练集内说话人的识别率的主要原因之一[4,5].

为了克服由于说话人声道长度的差异而引起的识别系统性能的下降，人们提出了基于声道长度归正技术的说话人自适应方法[2,6]，其基本思想是利用少量的自适应语料，估计出说话人的声道长度(或频率归正系数)，然后通过频率归正算法，将该说话人的语音频谱进行归正处理，归正后的语音频谱很像是一个具有标准声道长度的说话人的语音频谱. 大量的实验表明，将说话人的语音频谱归正后再进行训练和识别，可以提高系统的识别性能[7,8].

声道长度归正算法根据其频率归正系数估计的不同可以分为两大类，即最大似然法和参数法. 最大似然法通过最大似然准则[7]来直接估算频率归正系数，而参数法则通过估计共振峰频率来间接估计频率归正系数[8]. 这两种方法各有优缺点，最大似然法的性能稳定，但其运算量非常大；参数法虽然运算量小，但由于共振峰频率与文本和语境有关，加之噪声的影响，很难准确地估计出频率归正系数，因此利用参数法有时反而会使系统性能变坏.

对于如何提高非特定人语音识别系统的鲁棒性，本文则从一种新的途径出发，提出了一种基于 Mellin 变换的语音信号新特征，由于 Mellin 变换的尺度不变性，这种新的特征对话者的声道长度是不敏感的，因此可以大大减小由于声道长度不同而引起的说话人之间的差异. 在非特定人孤立词识别系统上的一系列实验表明，同 Mel 倒谱相比，这种新特征可以大大提高系统对不同说话人的识别率[9]，即使同基于最大似然的说话人自适应技术相比，这种新的特征仍可以提高系统的性能.

2　基于最大似然估计的声道归正技术

基于最大似然估计的声道归正技术就是通过最大似然估计器，估计出测试者声道长

度与参考说话人声道长度的比值，然后将测试者的语音归正成为参考说话者的语音，从而减小测试语音与归正的 HMM 模型之间的不匹配，以提高非特定人语音识别系统对说话人的鲁棒性. 下面我们简要介绍一下归正系数的估计，以及如何利用归正后的语音进行训练和识别.

2.1 归正系数的估计

设 $s_{i,j}(t)$ 为话者 i 的对应于脚本 $W_{i,j}$ 的语音信号，$s_{i,j,t}[n]$，$n=1,2,\cdots,T$（T 为总帧数）为 $s_{i,j}(t)$ 经过采样和分帧后的第 t 帧离散语音信号，$S_{i,j,t}(\omega)$ 为 $s_{i,j,t}[n]$ 的频谱，$\vec{C}_{i,j,t}$ 为对应于 $s_{i,j,t}[n]$ 的 Mel 倒谱特征矢量，因此，$s_{i,j}(t)$ 的整个倒谱矢量可以记为 $C_{i,j}=\{\vec{C}_{i,j,1},\vec{C}_{i,j,2},\cdots,\vec{C}_{i,j,T}\}$.

利用系数 α 进行频率归正的定义为

$$S^{\alpha}_{i,j,t}(\omega)=S_{i,j,t}(\alpha\omega). \tag{1}$$

通过归正后的频谱计算得到的 Mel 倒谱系数记为 $C^{\alpha}_{i,j}=\{\vec{C}^{\alpha}_{i,j,1},\vec{C}^{\alpha}_{i,j,2},\cdots,\vec{C}^{\alpha}_{i,j,T}\}$. 最后记 $C^{\alpha}_{i}=\{C^{\alpha}_{i,1},C^{\alpha}_{i,2},\cdots,C^{\alpha}_{i,N_i}\}$ 为话者 i 所有可利用的语音经过归正后计算得到的倒谱矢量集，$W_i=\{W_{i,1},W_{i,2},\cdots,W_{i,N_i}\}$ 为话者 i 的所有可利用语音的脚本，$\hat{\alpha}_i$ 代表话者 i 的最佳频率归正系数，λ 为已训练好的非特定人 HMM 模型，则最佳频率归正系数可通过以下公式来估算

$$\hat{\alpha}_i=\arg\max_{\alpha}P(C^{\alpha}_{i}|\lambda,W_i). \tag{2}$$

由于频率归正是一种非线性变换，所以要求出(2)式的解析解是非常困难的，我们一般是利用数值方法进行估计，其估计方法如下：

1) 根据共振峰的最大可能变化范围，将 $\hat{\alpha}_i$ 固定在区间[0.88, 1.12]上；

2) 将区间[0.88,1.12]等分成 M 等份，相应的 $\hat{\alpha}_i$ 记作 $[\hat{\alpha}^1_i,\hat{\alpha}^2_i,\cdots,\hat{\alpha}^M_i]$；

3) 对应于每一个 $\hat{\alpha}^k_i$，对语音信号进行归正和解码，求出其相应得分

$$P^{\hat{\alpha}^k_i}_i=P(X^{\hat{\alpha}^k_i}_i|\lambda,W_i)；$$

4) 最佳的频率归正系数为 $\hat{\alpha}_i=\arg\max_{\hat{\alpha}_i}P^{\hat{\alpha}^k_i}_i$.

2.2 训练过程

L. Li 等在文[2]中介绍了一种训练方法，其步骤为：

1) 将训练集中的所有话者的语音均分为两部分，分别记作 T 和 A；

2) 利用 T 中的语音训练 HMM 模型，所得的模型参数记为 λ_{T}；

3) 利用 λ_{T} 对 A 中的语音进行解码，由于假设声道长度是话者的独有特性，与话者发什么音没有关系，所以可以利用 A 中每位说话者的全部语音来估算出其最佳的频率归正系数 $\hat{\alpha}$，然后对 A 中的语音进行归正处理，并利用归正后的语音训练 HMM 模型，相应的模型参数记作为 λ_{A}；

4) 利用 λ_{A} 对 T 中的语音进行解码，并估算出每位话者的最佳频率归正系数 $\hat{\alpha}$，然后对 T 中的语音进行归正处理，并利用归正后的语音训练 HMM 模型，相应的模型参数记作为 λ_{T}；

5) 重复步骤 3)和步骤 4)，至到估计出来的每位话者的频率归正系数变化不大为止；

6) 利用估计出来的频率归正系数，对训练集中所有的语音进行归正处理，并利用它

们来训练 HMM 模型，训练所得的模型参数即为我们需要归正模型，记作为 λ_N.

也许有人会问，为什么要将训练集的语料均分为两部分，其实，如果训练语料足够多时，没有必要将其均分为两部分，但将其分为两部分却是必要的. 因为如果训练集不分开，我们会发现估计频率归正系数→语音频谱归正处理→训练模型参数这个迭代过程可以不断地提高训练集的似然得分，从而导致整个过程不收敛. 当然，将训练集分为两部分也并没有从理论上保证上述过程一定收敛，但实验表明这种做法可以使整个迭代过程收敛.

2.3　识别过程

和训练过程相同，识别过程也需要估计话者的频率归正系数 α_i，但识别过程中估计 α_i 不需要很多语音数据，一般有一个词的发音就够了，我们将其记作为 S_T. α_i 的估计方法可以分为两种，一种是已知 S_T 的脚本，一种是不知其脚本. 如果脚本已知，可以直接利用(2)式估计话者的频率归正系数 $\hat{\alpha}_i$. 但如果脚本不知，α_i 的估计过程分为两步：① 利用 λ_N 对 S_T 进行识别，识别结果记为 W^U；② 用 λ_N 代替(9)式中的 λ，用 W^U 代替(2)式中的 W 来估计 $\hat{\alpha}_i$. 当 $\hat{\alpha}_i$ 确定以后，对于同一话者的语音，先利用 $\hat{\alpha}_i$ 进行频率归正后，再送入识别器进行识别即可.

2.4　信号带宽变化的处理

频率归正会导致语音信号频谱的带宽发生变化. 例如，如果采样频率为 16kHz，那么语音信号的最大带宽为 8kHz. 但当频率归正系数在 0.88～1.12 之间变化时，归正后的语音频谱其最大带宽将不再是 8kHz，而是在 7.04～8.96kHz 之间变化. 但是我们估计最佳频率归正系数时所用模型其语音信号的带宽是 8kHz，因此，频率归正引入了信号带宽之间的不匹配.

为了解决因频率归正引起的带宽变化问题，我们一般采用分段线性频率归正方法，例如，采用以下的频率变换函数

$$G(f)=\begin{cases}\alpha f & 0\leqslant f\leqslant f_0,\\ \dfrac{f_{\max}-\alpha f_0}{f_{\max}-f_0}(f-f_0)+\alpha f_0, & f_0\leqslant f\leqslant f_{\max},\end{cases}\tag{3}$$

式中 f 为频率，$f_{\max}$ 为信号的最高频率，f_0 为一经验值，一般我们取大于第三共振峰频率的某一频率. 实验表明，基于分段线性频率变换的频率归正技术比用线性频率变换的频率归正技术性能更好[7].

3　基于 Mellin 变换的与说话声道长度无关的语音新特征

前面我们介绍了基于最大似然估计的声道归正自适应技术，用以克服由于话者声道长度的不同而引起的说话人之间的差异. 但是由于声道长度本身随着语境和发音的脚本变化也产生一些小的变化，加之环境因素的影响，因此很难准确地估计出频率归正系数，从而影响了频率归正自适应技术的性能.

在文[9]中，我们提出了一种基于 Mellin 变换的语音新特征，由于 Mellin 变换具有尺度不变性，所以这种特征是与话者声道长度无关的，因此既避开了归正系数的估计不准这一问题，又消除了由于声道长度不同而引起的说话人之间的差异. 这种新特征的提取框图如图 1 所示.

和倒谱特征参数的提取过程相类似，我们首先也要对语音信号进行分帧、加窗和预加重等预处理，并利用 FFT 来估计语音信号的频谱，然后对频谱的幅度取对数变换，对数操作在这里有两个用途，其一是压缩频谱幅度的动态范围，其二是将频谱域中的乘性成分变成对数谱域中的加性成分，以便后级滤波能够滤除时域中的卷积噪声. 紧接着对语音信号的对数谱进行修正的 Mellin 变换，最后对 Mellin 变换的结果进行离散余弦变换(DCT). DCT 的作用也有两点，其一是将 Mellin 变换的结果压缩成低阶的倒谱系数，其二是对不同维数的特征进行解相关处理，以便后面的 HMM 模型可以使用对角矩阵. 从上面可以看出，这种新的特征实际上是语音信号对数谱的修正 Mellin 变换，所以我们将其简称为 MMTLS (Modified Mellin Transform of the Log-Spectrum). 可以证明，MMTLS 是与话者的声道长度无关的[9].

在图 1 给出的 MMTLS 特征中，我们假设声道传递函数 $V(\alpha\omega)$ 中的 α 在每一帧语音信号的频谱内部是不变的，而实际上 α 也是变化的[1]，为此本文对 MMTLS 特征提取过程进行了改进，其框图如图 2 所示. 在改进的 MMTLS 特征提取过程中，我们将语音信号对数谱沿频率轴分成若干段，在每一段内，假设 α 是不变的，然后对每一段进行修正的直接 Mellin 变换.

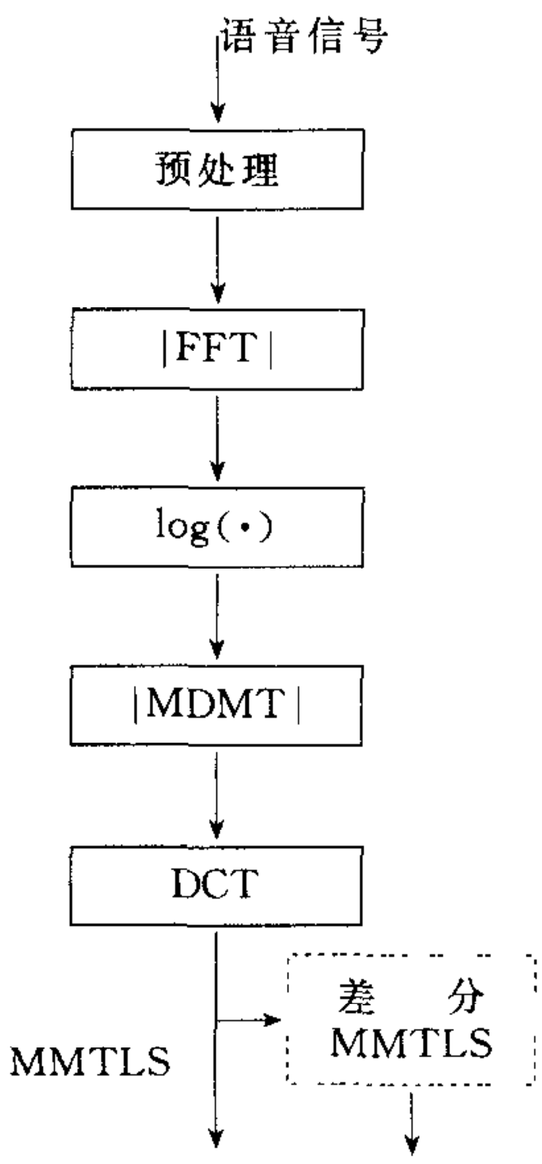

图 1 MMTLS 特征

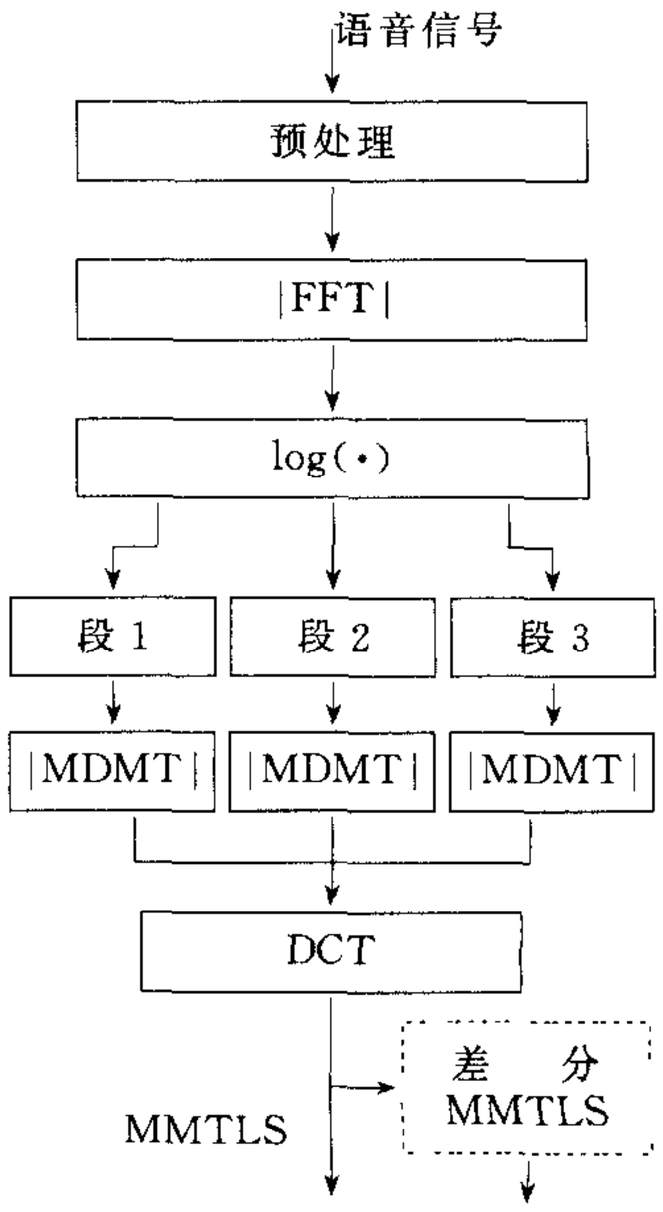

图 2 改进的 MMTLS

实验表明，改进后的 MMTLS 特征，其性能远远好于 MMTLS 特征. 本文后面给出的实验结果均是利用改进的 MMTLS 作为特征参数，因此，在后面的实验中，如无特殊申明，MMTLS 均代表改进的 MMTLS.

4 实验

实验中使用的语音库共包括 21 个男声的发音，词表为 174 个孤立词，每个人对每个词各发音一遍. 语音由 Sound-Blaster 加上普通的话筒录制，采样频率为 16kHz，有效位为 16bit.

实验中使用的语音识别系统为基于 CDHMM 的非特定人孤立词语音识别系统. 我们利用语音库中 21 位男声中的 16 位作为训练集(随机选取),其余 5 位作为测试. 识别器以整词建模,模型采用从左到右无跳转的 HMM 模型,每个模型共有八个状态.

语音帧长为 48ms(每帧含有 384 个语音样本点), 重叠率为 50%. 在 MFCC 特征参数估计过程中,窗函数采用长度为 48ms 的 Hamming 窗,预加重系数为 0.97. 特征有两种选法.

1)MFCC: 由 12 阶 MFCCs+12 阶一阶差分 MFCCs 组成的 24 维特征;

2)MMTLS: 由 12 阶 MMTLSs+12 阶一阶差分 MMTLSs 组成的 24 维特征.

实验结果如表 1 所示.

从表 1 中不难看出,MMTLS 特征的性能比 MFCC 的要好得多. 使用 MMTLS 特征时,系统对 5 位集外话者的平均误识率为 3.1%, 而使用 MFCC 时,系统对 5 位集外话者的平均误识率为 4.6%, 可见,同 MFCC 相比,使用 MMTLS 特征使系统的误识率下降了 33%.

表 1　采用不同方法系统的误识率

特征 / 误识率(%) / 话者	MFCC	基于最大似然估计的频率归正 MFCC	MMTLS
mstone	4.0	2.9	2.9
mstong	8.6	7.5	4.6
mwwren	6.3	5.7	4.0
mqwang	1.7	1.7	2.3
mshishi	2.3	1.7	1.7
均　值	4.6	3.9	3.1
标准离差	2.6	2.3	1.1

同 MFCC 特征参数相比,并不是对每位话者,MMTLS 特征都可以提高系统的识别率,例如对于 mqwang, 利用 MFCC 特征时系统的识别性能反而好于利用 MMTLS 时系统的性能,但从表 1 中可以看出,利用 MFCC 特征时,系统对不同讲话者的误识率其标准离差是利用 MMTLS 特征时的两倍,我们认为产生这一现象的主要原因在于,尽管 MFCC 特征是目前用于语音识别的最有效的特征,但它受话者声道长度的影响却很大,如果测试话者的声道长度与训练集中某位话者的声道长度非常接近,那么系统对该话者的识别率将会比较高,反之,系统对该测试者的识别率将会明显下降. 但对于 MMTLS 特征,由于 Mellin 变换的尺度不变性,因此该特征消除了由话者声道长度不同而引起的说话人之间的差异,从而提高了系统对话者的鲁棒性,减小了系统性能随话者的变化.

从表 1 中还可以看出,基于最大似然估计的声道长度归正方法是一种有效的说话人自适应技术,同 MFCC 特征相比,经过归一化处理后,系统对集外说话人的平均误识别率降低了 15.2%, 而误识率的标准离差下降了 11.5%,可见,声道归一化技术不仅可以降低系统对每位话者的误识率,而且误识率的标准离差也有所下降. 但即使加了声道归正技术,系统对不同话者的误识率其标准离差仍比 MMTLS 特征的要高得多,这进一步说明 MMTLS 特征较好地取掉了说话人之间的差异.

5　结论

对基于 HMM 的非特定人语音识别系统来说,说话人之间的差异主要是声道形状、尤其是声道长度之间的差异. 为了减小由于话者声道长度不同而引起的识别系统性能的

下降,本文研究了两种方法,一种是基于最大似然估计的声道归正说话人自适应方法,一种是基于 Mellin 变换的与说话人声道长度无关的语音信号新特征. 实验结果表明,这两种方法不仅可以提高系统对集外话者的识别性能,而且使系统对不同话者的误识率其离散程度也大大下降. 但比较之下,基于 Mellin 变换的 MMTLS 特征其性能要好得多,这说明,MMTLS 特征比 MFCC 更适合作为非特定人语音识别的特征参数,而且不再需要话者归正自适应处理.

参 考 文 献

1 Hisashi Wakita. Normalization of vowels by vocal-tract length and its application to vowel identification. *IEEE Transactions on Acoustics, Speech and Signal processing*, April 1997, ASSP-**25**(2): 183～192

2 Li Lee, Richard C. Rose. Speaker normalization using efficient frequency warping procedures. In: Proceeding of International Conference on Acoustics, Speech and Signal Processing (ICASSP'96), IEEE, Atlanta, USA, May 1996, 353～356

3 Wegmann S, McAllaster D, Orloff J, Peskin B. Speaker normalization on conversational telephone speech. In: Proceeding of International Conference on Acoustics, Speech and Signal Processing (ICASSP'96), IEEE, Atlanta, USA, May 1996, 339～341

4 Cohen J, Kamm K Andreou A. G. Vocal tract normalization in speech recognition: compensating for systematic speaker variability. *The Journal of the Acoustical Society of America*, 1995, **97**(2): 3246～3247

5 Kamm T, Andreou A, Cohen J. Vocal tract normalization in speech recognition: compensating for systematic speaker variability. In: Proceeding of the Fifteen Annual Speech Research Symposium, CLSP, Baltimore, Md: Johns Hopkins University, June, 1995, 175～178

6 Eide E, Gish H. A parametric approach to vocal-tract-normalization. In: Proceeding of the Fifteen Annual Speech Research Symposium, CLSP, Baltimore, Md: Johns Hopkins University, June, 1995, 161～167

7 Li L, Richard R. A frequency warping approach to speaker normalization. *IEEE Transactions on Speech and Audio Processing*, January 1998, 49～60

8 Ellen E , Gish H. A parametric approach to vocal tract length normalization. In: Proceeding of ICASSP'96, IEEE, Atlanta, USA, May 1996, 346～348

9 Chen Jingdong, Xu Bo, Huang Taiyi. A novel robust feature of speech signal based on the Mellin transform for speaker-independent speech recognition. In: Proceedings of ICASSP'98, Seattle, USA, May 1998, 629～732

陈景东 1993 年和 1995 年分别于西北工业大学获应用电子及水下信号处理专业学士及硕士学位. 1998 年于中国科学院自动化研究所模式识别国家实验室获语音识别专业博士学位. 曾于日本国际先进通信技术研究所(ATR)音声翻译通信实验室进行语音合成方面的研究工作. 现为澳大利亚 Griffith 大学微电子系的 Senior researcher fellow. 已在国际、国内学术刊物、会议上发表论文二十余篇. 主要学术研究方向是语音识别、高分辨力特征提取、高鲁棒性算法、自适应算法、语音合成及数字信号处理等.

徐 波 1992 年和 1997 年在中国科学院自动化研究所分别获硕士和博士学位. 1988 年起一直从事智能信息处理的研究,重点在汉语语音信号处理、识别和合成,自然语言理解等方面的基础研究和系统开发工作. 目前正负责承担国家"973"、国家"863"和重大国际合作等多个项目的研究和开发. 现任该所研究员,博士生导师,中国声学学会和中国自动化学会委员.

黄泰翼 1956 年毕业于上海交通大学电机工程系. 现任中国科学院自动化研究所研究员,长期从事信息科学的研究,重点是人机语音通讯方法及应用的研究,是国内人机语音通讯领域的开拓者之一,支持了十多个国家重大项目及自然科学基金课题,在国际、国内刊物和会议上发表论文数十余篇.