汉语耳语音孤立字识别研究 *

杨莉莉 † 林 玮 徐柏龄

(南京大学声学所 近代声学重点实验室 南京 210093)

摘要 耳语音识别有着广泛的应用前景，是一个全新的课题。但是由于耳语音本身的特点，如声级低、没有基频等，给耳语音识别研究带来了困难。本文根据耳语音信号发音模型，结合耳语音的声学特性，建立了一个汉语耳语音孤立字识别系统。由于耳语音信噪比低，必须对其进行语音增强处理，同时在识别系统中应用声调信息提高了识别性能。实验结果说明了 MFCC 结合幅值包络可作为汉语耳语音自动识别的特征参数，在小字库内用 HMM 模型识别得出的识别率为 90.4%。

关键词 耳语音，语音识别，语音增强

Isolated word recognition in Chinese whispered speech

YANG Li-Li LIN Wei XU Bo-Ling

(*Key Laboratory of Modern Acoustics, The Institute of Acoustics, Nanjing University, Nanjing 210093*)

Abstract The whispered speech recognition is a new subject which has wide applications. However, the characteristics of whispered speech such as its low sound pressure level and the lack of fundamental frequency bring difficulty to the whispered speech recognition. In this paper, a Chinese isolated word recognition system is established based on the source-filter generation model combined with the acoustic characteristics of whispered speech. In addition, the speech enhancement algorithm is added to the system to improve the SNR of whispered speech, and the tone information is implemented to acquire better recognition performance. The experimental results demonstrate that the MFCC combined with the amplitude contour features can be used as efficient parameters for the Chinese whispered speech recognition. A recognition rate of 90.4% is obtained when a small Chinese isolated word database is tested using HMM approach.

Key words Whispered speech, Speech recognition, Speech enhancement

2004-12-17 收稿； 2006-01-16 定稿

* 国家自然科学基金项目 (60272037 和 60340420325)

作者简介：杨莉莉 (1981-)，女，福建漳州市人，硕士研究生，研究方向：语音与声信号处理。

林玮 (1977-)，男，博士生。徐柏龄 (1941-)，男，南京大学声学所教授，博士，博士生导师。

† 通讯联系人 Email: yanglili@nju.org.cn

1 引言

耳语是人们的一种语音交流方式。以往耳语音的研究主要为了语音基础研究和医学工作的需要 [1,2]。随着移动通讯工具的广泛使用，耳语音也成为通讯中的一种交流方式。在公共场合中，使用耳语音可以不影响他人并保证通话的保密性；对于喉部切除的失音患者，如能将其发出的气声自动识别出来，无需电子喉就能转换为正常音，对于每年上万人数量增长的失音患者来说，提供了一种更容易被接受的语言交流方式；此外，耳语音在公安、司法等部门也有重要的应用前景，识别耳语音可有助于公安部门语音的破译。

耳语音最主要的特征是声带不振动，没有基频，声级低，在公共场合中，信噪比更低。因此耳语音的研究特别是识别，比正常音要困难得多。目前耳语音方面的研究相对较少，在耳语音识别方面 [3,4] 也非常少。文献 [3] 对日语耳语音用 MFCC 参数 HMM 模型识别，正确率为 68%，用最大似然线性回归 (MLLR)[5] 自适应训练，可提高识别率十个百分点。文献 [4] 提出用英语耳语音共振峰偏移修正后，再转化为倒谱系数，进行 MLLR 自适应训练和识别，但文中没有给出具体的识别率。耳语音识别问题，无论在国内还是国外，都处于前期研究阶段，对于汉语这种有声调的语言，耳语音识别的难度就更大。

本文结合汉语耳语音的特点，建立了一个小字库的汉语耳语音孤立字识别系统，首先对耳语音前端处理进行了语音增强，然后用 MFCC 参数并结合汉语声调模型，同时采用幅值包络参数识别，得出的识别率是 90.4%，初步实现了汉语耳语音的自动识别，为国内此项研究填补了空白。

2 耳语音的特点

在《语音学和音系学字典》[6] 中有这样定义：耳语音是一种单一发音类型，声门前部 (韧带) 完全靠拢，后部 (杓状软骨) 有一个宽三角裂隙。气流通过开放区产生摩擦噪声，形成耳语音。

2.1 声学特性

耳语音与正常音有着不同的发音方式，导致它们有着不同的声学特性。耳语音的清擦音、塞擦音和塞音等清辅音部分与正常音的发音方式没有大的区别，但是元音却有较大的不同。正常元音的声源为周期性脉冲，声带振动的周期即为基音周期，声带振动的频率即为基频。而耳语音的声源为噪声源，即耳语发音的元音 (和浊辅音) 没有基频。由于耳语音是气流发声，声级较低，它比正常发音约低 20dB。并且耳语发音时需大量的气流及其较低的肺活量和气管压力，使得耳语音语速较慢，音长较长 [7]。此外，正常音具有浊音能量大于清音能量和浊音过零率小于清音过零率的规律，但是由于耳语音的各音素都是噪声激励，故耳语音没有此规律。另外，由于耳语发音时，假声带区域变窄、声门保持半开的状态，使得声道增加了气管和肺部分，因此声道传输函数发生改变，耳语音元音共振峰的位置和带宽都发生变化 [8]，还使得耳语元音 500Hz 以下的谱被衰减，200Hz~2000Hz 的谱较之正常元音的谱更平坦。

2.2 耳语音信号产生模型

由耳语音的声学特性分析，我们可以建立一个离散时域的耳语音信号产生模型，如下图 1 所示：

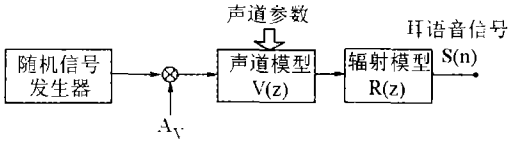

图 1 耳语音信号发生模型

此模型与正常音信号产生模型的区别为：(1) 正常音激励源分浊音和清音两个分支，浊音情况下激励信号由一个周期脉冲发生器所产生，清音则由一个随机噪声发生器产生，而耳

语音的各音素都是噪声激励，所以它的激励源只有随机信号发生器一个，其中乘系数 A_v 的作用是调节耳语音信号的幅度或能量；(2) 声道模型 $V(z)$ 在正常音情况下是一个全极点函数，而耳语音声道传输函数产生了附加的零极点。

3　预处理过程

3.1　端点检测

耳语音的信噪比低，而通常用于端点检测的能零积法 [9] 抗噪声性能不高，正确率低。本课题组先前提出的基于听觉模型法 [10] 和信息熵法 [11] 可准确地判断出语音段，其中，信息熵法鲁棒性强，运算量小，因此本文采用信息熵值法进行耳语音的端点检测。

3.2　语音增强

语音识别系统通常是将在安静环境下训练得到的参数应用于实际环境中，当实际环境中有噪声存在时语音识别系统性能急剧下降。而耳语音由于本身的特性导致声级低，安静环境下信噪比就很低 (约 3~5dB)，实际环境下信噪比下降得更多，因此为了使耳语音识别系统得到满意的工作性能，需要对耳语音进行语音增强以提高信噪比。耳语音的语音增强处理可由信号放大 (归一化)、高通滤波和功率谱相减法组成。

信号放大：由于耳语音的能量很小，给辨认和处理带来了困难，因此可将其信号放大，进行归一化处理，以汉语 “三” 为例，结果如图 2(b) 所示，信号放大后的耳语音幅度增加约 20 倍，人耳可以比较清晰地辨认出语音。

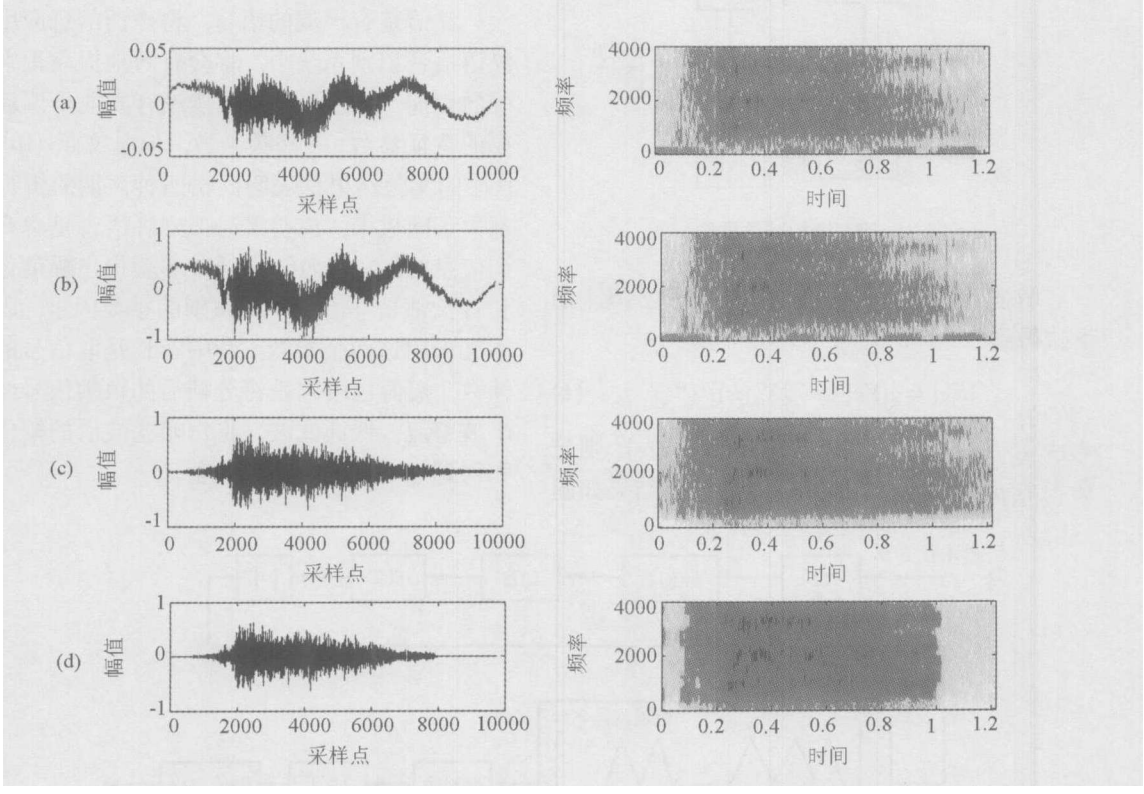

图 2　汉语 “三” (\san\) 语音增强结果

(a) 原始语音　(b) 信号放大后　(c) 高通滤波后　(d) 谱相减法去噪声后

高通滤波：由于耳语音没有基频，共振峰向高频偏移，使得其 500Hz 以下的能量很小，而机器噪声多在此频段，故可进行截止频率为 500Hz 的高通滤波，滤波后的语音对可懂度和识别没有影响。结果如图 2(c) 所示，可见原始语音中 500Hz 以下没有了语音能量，并且低频噪声被明显抑制了。

谱相减法去噪声：功率谱相减法及其改进算法 [12,13] 是一种常见的消除噪声的技术，其基本原理是从含噪语音的功率谱中减去噪声的功率谱，基本框图如图 3 所示。图中 $y(n)$ 表示含噪语音，$s(n)$ 是纯净的语音，$d(n)$ 为加性噪声，$\lambda_n(k)$ 是噪声功率谱系数，通常在语音的无声段可以估得 (本文采用信息熵端点检测法检测语音中的噪声段)。

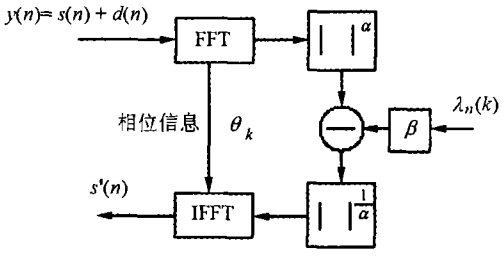

图 3 功率谱相减法框图

增强后的语音 $s'(n)$ 的谱幅度系数 $|S'_k|$ 由下式得到

$$|S'_k| = [|Y_k|^\alpha - \beta\lambda_n^\alpha(k)]^{1/\alpha} \qquad (1)$$

$\alpha = 2$, $\beta = 1$, Y_k 和 S_k $(k = 0, 1, \cdots)$, 则分别表示 $y(n)$ 和 $s(n)$ 的频谱系数。其结果如图 2(d) 中所示，语音中所含噪声被明显抑制。语音增强后，信噪比比原始含噪语音大约提高了 15dB。

4 特征参数的选取

4.1 MFCC 参数

Mel 频率倒谱参数 (Mel-Frequency Cepstral Coefficients)[14]，简称为 MFCC，是一种能够比较充分利用人耳这种特殊的感知特性的参数，近年来，得到了广泛的应用。MFCC 参数的提取过程如图 4 所示。

本文采用 12 阶的 MFCC，最后计算描述动态特性的一阶差分 MFCC。这样每帧语音就可由一个 24 维的观测向量 (12 维 MFCC 和 12 维 ΔMFCC) 表示。

4.2 声调信息参数

汉语是有声调的语言，将声调信息应用在汉语语音识别系统中，将会有效地提高识别系统的性能 [15]。耳语音虽然没有表征声调最主要的特征参数 — 基频参数，但是文献 [16] 和 [17] 的实验结果都表明汉语四种声调的辨听率都大于随机率，说明汉语孤立耳语音是含有声调信息的，文献 [16] 中通过实验得出幅值包络和音长都是耳语音声调识别的重要因素，因此本文采用这两个参数。其中音长是取信号的总帧数，幅值包络则是将分帧后的每帧信号经过半波整流、低通滤波，将每帧滤波后的幅值求和得到每一帧的幅值包络值。

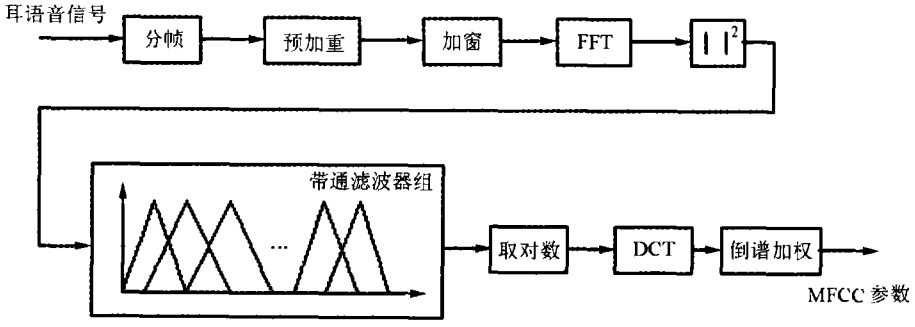

图 4 MFCC 参数提取过程

关于声调信息的应用方式，目前文献中较少涉及。比较常用的方法是采用 Multi-Stream 模型 [15,18]，即将声调特征和识别系统原有的特征分别作为特征流，分别训练码本。在识别算法中计算距离时，分别计算声调特征和原有特征的距离，然后加权求和。另外一个有效的方法是将声调特征和系统原有特征拼接起来作为一个特征向量处理，即 Single-Stream 模型。因为 MFCC 参数是按帧计算的，所以可将幅值包络参数与其拼接。音长则是取信号的总帧数，应用音长特征时只能采用 Multi-Stream 模型。

5　实验系统构架

综上所述，我们可以建立一个汉语耳语音孤立字识别系统，如图 5 所示。

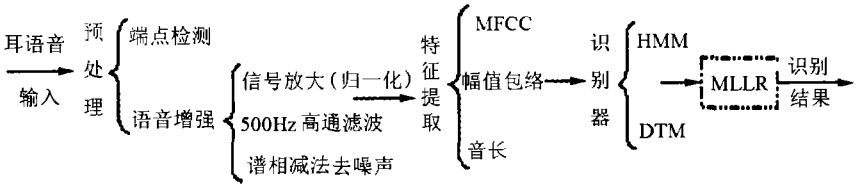

图 5　汉语耳语音孤立字识别系统

图 5 中，识别算法采用 DTW (Dynamic Time Warping, 动态时间弯折) 和 HMM(隐马尔可夫模型) 两种常见的算法进行比较。MLLR (Maximum Likelihood Linear Regression, 最大似然线性回归) 自适应训练可提高识别率，是系统的可选部分。

6　实验研究

6.1　实验数据

耳语音识别系统无论是在手机通讯上的应用，还是对于失音者语音恢复的应用，系统都可针对特定人进行训练和识别。因此，本文针对特定人的耳语音进行实验。同时，实验采用了通讯中常用字，一共有 40 个 (小字库)，如表 1 所示。所用语音由一女生发音，每个字发 24 遍，采样频率为 8000Hz 。考虑到普适性，样本的采集历经 3 个月，采集的耳语音中含有计算机等机器的背景噪声，信噪比约为 3~5dB 。取每类音的前 4 个样本为训练样本，后 20 个样本共 800 个字作为识别样本。下面的实验都是对这些数据进行处理的。

表 1　实验所用孤立字

零	一	二	三	四	五	六	七	八	九
你	我	他	是	有	男	女	电	话	喂
好	请	讲	了	在	家	找	忙	哪	里
谢	说	天	没	到	声	上	班	开	会

6.2　实验结果

6.2.1　语音增强效果

为了验证本文中提出的语音增强方法的有效性，进行了三组实验：第一组，训练和识别音都没有经过语音增强；第二组，训练用音进行语音增强；第三组，训练和识别音都进行了语音增强。所得识别结果如图 6 所示，其中特征参数都只取了 MFCC，识别算法分别采用了 DTW 和 HMM 。如图 6，由于训练与识别环境不匹配，第二组语音的识别率比第一组明显下降。而第三组语音的识别率均高于第一组，其中 DTW 提高了 7.8%,HMM 提高了 12%，说明在耳语音识别系统中增加语音增强处理是有效的。

6.2.2　应用特征参数比较

为了比较上文中所提到三种特征参数的识别效果，也进行了三组实验：第一组，只采用 MFCC 参数；第二组，采用 MFCC 加音长的 Multi-Stream 模型；第三组，采用 MFCC 加幅值包络的 Single-Stream 模型。所得识别结

果如图 7 所示。第二组与第一组的识别率相差很小，提高得十分有限，音长参数在汉语耳语音识别中的作用不大，并且 Multi-Stream 模型需增加存储空间，运算量大了一倍，因此这组参数并无实际应用意义。第三组的识别率与第一组相比有了一定的提升，最佳识别率达到了 90.4%，同时 Single-Stream 模型不需要增加额外的运算量，所以在汉语耳语音识别系统中幅值包络参数是 MFCC 参数的有益补充。

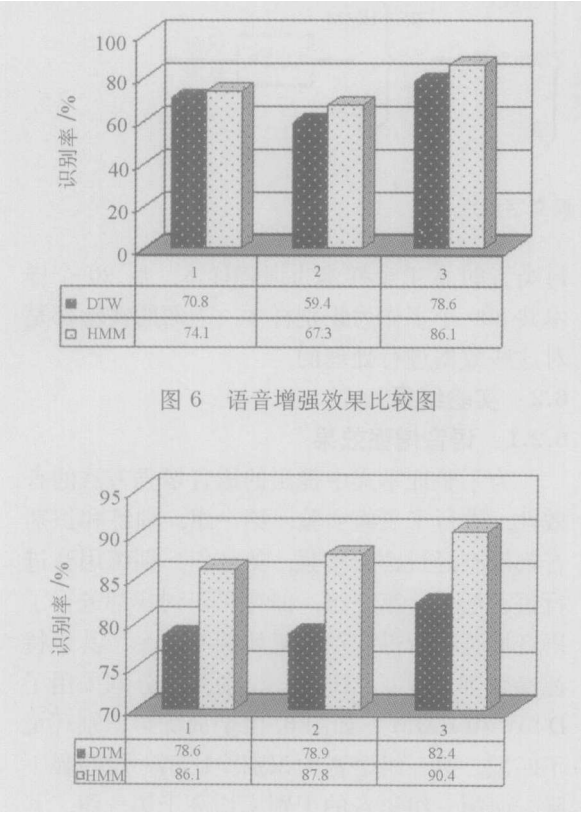

图 6 语音增强效果比较图

图 7 特征参数应用结果比较图

7 结论

本文分析了耳语音不同于正常音的声学特性，根据耳语音信号产生模型建立了汉语耳语音孤立字识别系统。通过实验比较分析，验证了语音增强处理在耳语音识别系统中的有效性，得出 MFCC 结合汉语声调模型中的幅值包络参数可作为汉语耳语音自动识别的特征量，它利用 Single-Stream 模型将汉语耳语音的声调信息应用在识别系统中取得了较好的效果，最佳识别率达到了 90.4%。

汉语耳语音孤立字识别离实用还有一段距离。我们知道，人们可用耳语音进行正常的交流，人脑对耳语的辨认还依赖于对上下文语义信息的理解，因此研究连续耳语音识别有可能提高耳语音的识别率，下一步，我们将研究连续耳语音的识别，继续推进耳语音识别从研究走向实用。

参 考 文 献

1 Schwartz M F, Rine H E. *J. Acous. Soc. Am.*, 1968, **44**(6):1736~1737.

2 于华. 中央民族大学学报, 1996, **5**(2):163~166.

3 Itoh T, Takeda K and Itakura F. *Proc. ICASSP*, Orlando, Florida, USA, 2002, 389~392.

4 Morris R W. [PhD Thesis], Georgia Institute of Technology, USA, 2002.

5 Leggetter C J, Woodland P C. Proc. of the ARPA Spoken Language Technology Workshop, Barton Creek, 1995.

6 R L 特拉斯克编, 《语音学和音系学字典》 (A dictionary of Phonetics and Phonology), 《语音学和音系学字典》编译组译, 北京: 语文出版社, 2000. 286.

7 Gao M. Tones in Whispered Chinese: Articulatory and Perceptual Cues. [Master], University of Victoria, Canada, 2002.

8 Jovicic S T. *Acustica-Acta Acustica*, 1998, **84**(4):739~743.

9 陈韬, 李昌立, 莫福源. 声学学报, 1993, **18**(3):161~171.

10 丁慧, 栗学丽, 徐柏龄. 应用声学, 2004, **23**(2):20*sim*25.

11 栗学丽, 丁慧, 徐柏龄. 声学学报, 2005, **30**(1):69~75.

12 Boll S F. *IEEE Transaction on Acoustic, Speech, and Signal Processing*, 1979, **ASSP-27**(2):113~120.

13 Hu Hwai-Tsu, Kuo Fang-Jang, Wang Hsin-Jen. *Speech Communication*, 2002, **36**(3-4):205~218.

14 Young S J. The HTK Book. Version 2.1, 1997. 72~75. http://svr-www.eng.cam.ac.uk.

15 Hank Chang, Han Huang, Frank Seide. *Proc. ICASSP*, 2000, 1523~1526.

16 沙丹青, 栗学丽, 徐柏龄. 电声技术, 2003, **11**:4~7.

17 梁之安. 生理学报, 1963, **26**(2):85~91.

18 Sharlene Liu, Sean Doyle, Allen Morris. *Proc. ICSLP*, 1998, **6**:2647~2650.