科学通报 2011 年 第 56 卷 第 4-5 期：354 ~ 359
论 文 www.scichina.com csb.scichina.com
《中国科学》杂志社
SCIENCE CHINA PRESS

默声音符编码文本输入方法

滕鹏, 贾云得*

北京理工大学计算机学院, 智能信息技术北京市重点实验室, 北京 100081
* 联系人, E-mail: jiayunde@bit.edu.cn

2010-10-09 收稿, 2010-12-20 接受
国家自然科学基金资助项目(90920009, 60905006)

摘要 人们希望语音输入能够取代键盘, 成为计算机文本输入的最主要方式. 但语音输入在噪声环境中准确率低, 且会泄露输入信息, 也会影响他人. 本文提出默声音符编码文本输入方法, 该方法选择默声音中容易发声、容易精准识别且彼此间容易区分的少数几个音素作为默声音符(silent voice symbol), 对各种字符进行编码. 类似摩尔斯电码, 用户只需要记住编码方式, 就可以通过读默声音符准确地实现文本输入. 实验证明提出的方法有潜力替代键盘在各种环境下完成文本输入任务.

关键词
文本输入
默声音符
无声交流
无声语音接口

语音是人与人之间最自然、最重要的交流方式, 人们也希望语音输入能够取代键盘输入, 成为计算机文本输入的最主要方式. 目前已经成功开发出许多实用的语音输入系统, 这些系统在嘈杂环境中识别率会显著下降, 甚至不能工作; 在安静的环境下使用这些系统不仅会影响周围的人和环境, 也不利于保密输入内容[1,2]. 为了解决这些问题, 人们开始研究无声语音接口(silent speech interface, SSI)[3], 比如使用 EEG 传感器获取与无声语音相关的脑活动电位[4], 使用表面肌电传感器记录语音器官相关肌肉的体表电位并转换成对应的孤立词[5]等, 其基本思想是测量语音产生过程中某个环节的生理数据, 理解其中蕴含的语言信息. 这些SSI还处于概念研究阶段. 还有一类 SSI 是 Nakajima 等人[2,6,7]开发的 NAM (non-audible murmur)麦克风, 可以获取用户发出 NAM 时经身体组织传播的声道共振信号, 从中提取出 NAM 语音信息, 并对这种语音信息进行识别. Hirahara 等人[8]使用 NAM 麦克风将 NAM 加强为正常语音, 用于人与人的交流. 目前已经有商品化的 NAM 麦克风[3]. 相对于其他 SSI 所测量的生理数据, NAM麦克风所采集到的是经身体组织传导的声学信号, 更能直接、准确、稳定地反映最终的语音内容, 是目前最有希望被广泛应用的 SSI. 然而, NAM 识别面临着与正常语音识别同样的问题, 就是正确率不够理想[2], 不能很好地处理用户口语中的偶然性现象[1]. 此外, NAM 的信噪比低于正常语音信号, 并且NAM麦克风相对于普通麦克风采集到的信号又缺少高频信息[7,9], 这使得 NAM 识别更加困难.

针对语音输入存在的这些问题, 本文提出一种默声音符编码文本输入方法. 默声音(silent voice)是指声带不振动时, 由喉气管狭窄处的气流噪声及其在声道中的共鸣叠加而成的声音, 其响度类似低声耳语. 本文方法选择默声音中容易发声、容易精准识别且彼此间容易区分的少数几个音素作为默声音符(silent voice symbol), 对各种字符进行编码. 类似摩尔斯电码, 用户只需要记住编码方式, 就可以通过读默声音符实现文本输入. 该方法不需要手眼参与, 不易受到周围噪声的干扰, 也不会泄露输入信息或影响他人.

1 默声音符的选择

我们选择国际音标中的单元音音素/ʌ/, /e/, /i/,

英文引用格式: Teng P, Jia Y D. A text input method based on silent voice symbol encoding (in Chinese). Chinese Sci Bull (Chinese Ver), 2011, 56: 354–359, doi: 10.1360/972010-1864

/ɔ/, /u/作为基本的默声音符, 分别记作 a, e, i, o, u. 这 5 个音素广泛存在于众多语言当中, 无论是习惯于使用哪种语言的用户都能发出其对应的默声音. 这 5 个稳态单元音在发音空间中, 都处于发音的极端位置, 分别是低元音、前中元音、前高元音、后中元音、后高元音. 在读不同音符时, 语音器官(舌、腭、唇等)使声道表现出显著不同的共鸣特性[10], 产生的默声音信号差异较大, 彼此间易于区分. 再将上述 5 个音素对应的长音加入到默声音符集合中, 分别记作 A, E, I, O, U(需要说明的是, 这只是可行的默声音符选择方案之一). 我们将最终得到的默声音符集合记为 $V=\{V_k \mid k=1,\cdots,K\}$, 其中 V_k 表示第 k 个默声音符, K 为默声音符的数目.

默声音的声压波会由声道壁传入身体组织, 最后传播至体表. 使用接触式微型振动传感器, 即可获得表示默声音的振动信号. 图 1 是默声音符编码文本输入系统组成示意图, 图 2 为实验系统, 实验中我们将接触式微型振动传感器放置在下颚与颈部围

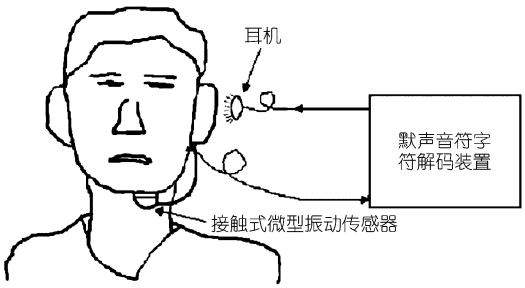

图 1　默声音符编码文本输入系统组成

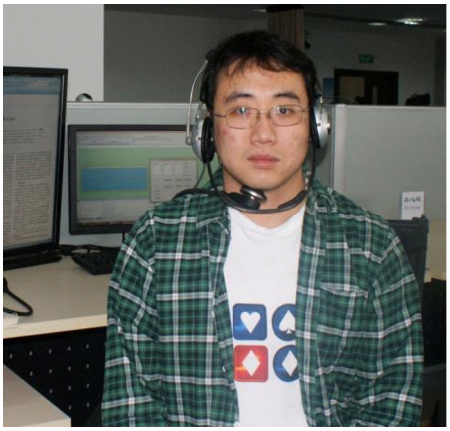

图 2　实验系统

成的平坦区域. 在实验者读某个默声音符 $v \in V$ 时, 采集到的默声音信号段即为 v 的一个观察样本, 用 $\boldsymbol{w}$ 表示. $\boldsymbol{w}$ 是一个实数序列, 其中每个实数表示振动信号在相应采样时刻的幅值. 图 3 是采集到的一段声道壁振动信号, 其中直线段对应的信号片段为默声音信号.

2　默声音符的识别

默声音符识别是由用户读默声音符 $v \in V$ 时的观察样本 $\boldsymbol{w}$, 得到对 v 的预测. 一种方法是提取 $\boldsymbol{w}$ 的特征 $\boldsymbol{y}$, 并将每个 $V_k \in V$ 分别描述为一个特征类, 计算并比较 $\boldsymbol{y}$ 属于各个特征类的可能性, 选择可能性最大的类别作对 $\boldsymbol{w}$ 所代表默声音符的预测[11]. 对观察样本 $\boldsymbol{w}$ 的特征提取过程分为两个阶段: 参数提取和特征选择. 参数提取主要完成 $\boldsymbol{w}$ 中与识别任务相关的信息提取, 用 p 维参数矢量 $\boldsymbol{x}$ 表示; 特征选择阶段对 $\boldsymbol{x}$ 进行分析, 得到用于描述 $\boldsymbol{w}$ 的 m 维特征矢量 $\boldsymbol{y}$.

2.1　参数提取

参数提取的一种常见方式是: 假设观察样本 $\boldsymbol{w}$ 的信号类型已知, 建立信号的产生模型, 然后求取它对应于该模型的系数作为 $\boldsymbol{x}$ 的元素. 默声音与正常语音都属于语音相关的声学信号, 在语音识别领域广泛使用的声学信号参数提取方法同样适用于默声音信号, 如从发声机理角度提出的线性预测系数(linear prediction coefficient, LPC)[12]和从听觉系统机理出发提出的 Mel 倒谱系数(mel-frequency cepstral coefficient, MFCC)[13]等表示方法. 将所选择的参数提取过程记作 $PA(\cdot)$, 即

$$\boldsymbol{x}=PA(\boldsymbol{w}). \tag{1}$$

2.2　特征选择

对参数矢量 $\boldsymbol{x}$ 进行分析, 得到用于分类的特征

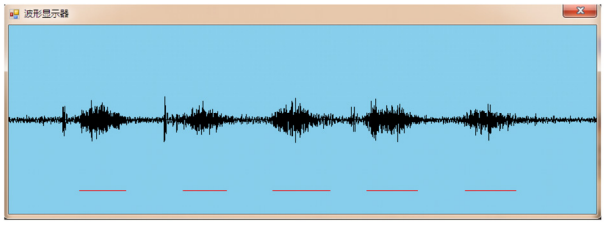

图 3　一段声道壁振动信号

矢量 $\boldsymbol{y}$, 即

$$\boldsymbol{y} = F(\boldsymbol{x}). \tag{2}$$

$F(\cdot)$是一个由 $\boldsymbol{x}$ 所在的参数空间到 $\boldsymbol{y}$ 所在的特征空间的映射, 每一个默声音符观察样本都对应于特征空间的一个矢量. 我们期望在特征空间中不同音符样本的特征间差异较大, 同一音符样本的特征间差异较小, 即具有良好的可分性. 假设 $F(\cdot)$是满足以上期望的线性映射, 也就是存在大小为 $p\times m$ $(p \geqslant m)$ 的线性变换矩阵 $\boldsymbol{T}$ 使得

$$\boldsymbol{y} = F(\boldsymbol{x}) = \boldsymbol{T}^{\mathrm{T}}\boldsymbol{x}, \tag{3}$$

则可以利用线性判别分析(linear discriminant analysis, LDA), 通过监督学习由各个默声音符的观察样本集合求得 $\boldsymbol{T}$ 和 $F(\cdot)$.

将各个默声音符的样本作为训练集. 设得到的默声音符 V_k 的观察样本集合 W_k 中包含 N_k 个观察样本, $\boldsymbol{w}_{ki}$ 为 W_k 中第 i 个观察样本, 其中 k=1,⋯,K; i=1,⋯,N_k, N 是所有默声音符观察样本的总数, $\boldsymbol{x}_{ki}$ 为 $\boldsymbol{w}_{ki}$ 的参数矢量, $\boldsymbol{y}_{ki}$ 为 $\boldsymbol{w}_{ki}$ 的特征矢量. 由式(1)~(3)得

$$\boldsymbol{x}_{ki} = PA(\boldsymbol{w}_{ki}),\ \ \boldsymbol{y}_{ki} = F(\boldsymbol{x}_{ki}) = \boldsymbol{T}^{\mathrm{T}}\boldsymbol{x}_{ki}. \tag{4}$$

在特征空间中, 样本协方差矩阵的行列式可以作为数据分散程度的一种度量. 假设采集的样本能较好地反映各个默声音符所有可能样本的真实分布, 定义类内协方差矩阵 $\tilde{\boldsymbol{S}}_{\mathrm{W}}$ 与类间协方差矩阵 $\tilde{\boldsymbol{S}}_{\mathrm{B}}$ 分别为

$$\tilde{\boldsymbol{S}}_{\mathrm{W}} = \sum_{k=1}^{K}\frac{1}{N_k}\sum_{i=1}^{N_k}(\boldsymbol{y}_{ki}-\tilde{\boldsymbol{\mu}}_k)(\boldsymbol{y}_{ki}-\tilde{\boldsymbol{\mu}}_k)^{\mathrm{T}},$$
$$\tilde{\boldsymbol{S}}_{\mathrm{B}} = \sum_{k=1}^{K}N_k(\tilde{\boldsymbol{\mu}}_k-\tilde{\boldsymbol{\mu}})(\tilde{\boldsymbol{\mu}}_k-\tilde{\boldsymbol{\mu}})^{\mathrm{T}}, \tag{5}$$

其中, $\tilde{\boldsymbol{\mu}}_{\boldsymbol{k}} = 1/N_k\sum_{i=1}^{N_k}\boldsymbol{y}_{ki}$ 为第 k 类模式的均值, $\tilde{\boldsymbol{\mu}} = 1/N\sum_{i=1}^{N}\boldsymbol{y}_i$ 为所有观察样本模式的均值. 数据的可分性可以用下式度量:

$$J(\boldsymbol{T}) = \frac{\left|\tilde{\boldsymbol{S}}_{\mathrm{B}}\right|}{\left|\tilde{\boldsymbol{S}}_{\mathrm{W}}\right|}, \tag{6}$$

能使(6)最大化的 $\boldsymbol{T}$ 即为所需要的线性变换矩阵. 由式(4)~(6)可得

$$J(\boldsymbol{T}) = \frac{\left|\tilde{\boldsymbol{S}}_{\mathrm{B}}\right|}{\left|\tilde{\boldsymbol{S}}_{\mathrm{W}}\right|} = \frac{\left|\boldsymbol{T}^{\mathrm{T}}\boldsymbol{S}_{\mathrm{B}}\boldsymbol{T}\right|}{\left|\boldsymbol{T}^{\mathrm{T}}\boldsymbol{S}_{\mathrm{W}}\boldsymbol{T}\right|}, \tag{7}$$

其中

$$\boldsymbol{S}_{\mathrm{W}} = \sum_{k=1}^{K}\frac{1}{N_k}\sum_{i=1}^{N_k}(\boldsymbol{y}_{ki}-\boldsymbol{\mu}_k)(\boldsymbol{y}_{ki}-\boldsymbol{\mu}_k)^{\mathrm{T}}.$$
$$\boldsymbol{S}_{\mathrm{B}} = \sum_{k=1}^{K}N_k(\boldsymbol{\mu}_k-\boldsymbol{\mu})(\boldsymbol{\mu}_k-\boldsymbol{\mu})^{\mathrm{T}},$$

分别为参数空间中观察样本的类内协方差矩阵与类间协方差矩阵($\boldsymbol{\mu}_{\boldsymbol{k}} = 1/N_k\sum_{i=1}^{N_k}\boldsymbol{x}_{ki}$ 是第 k 类样本参数矢量的均值, $\boldsymbol{\mu} = 1/N\sum_{i=1}^{N}\boldsymbol{x}_i$ 是所有参数矢量的均值). 求使式(7)最大化的 $\boldsymbol{T}$ 等同于求解广义特征方程

$$\boldsymbol{S}_{\mathrm{B}}\boldsymbol{T} = \lambda\boldsymbol{S}_{\mathrm{W}}\boldsymbol{T}, \tag{8}$$

则 $\boldsymbol{T}$ 的列向量即为式(8)中最大的前 m 个特征值对应的特征向量, 通常取 m = min(K–1, p) [11]. 至此, 对于每一个观察样本 $\boldsymbol{w}_{ki}$ 或 $\boldsymbol{w}$, 可以使用式(4)提取对应的特征矢量 $\boldsymbol{y}_{ki}$ 或 $\boldsymbol{y}$.

2.3 分类

在特征空间中,V_k 的特征类由它所有观察样本的特征矢量的集合 $Y_k = \{y_{ki} \mid i = 1,\cdots,N_k\}$ 表示. 使用最小距离分类器, 如果一个未见观察样本的特征矢量 $\boldsymbol{y}$ 到第 j 个特征类的距离 $d_j(\boldsymbol{y})$是它到所有特征类的距离 $d_k(\boldsymbol{y})$, k=1,⋯,K 中最小的, 那么则将它分类为 V_j. $\boldsymbol{y}$ 到 Y_k 的距离 $d_k(\boldsymbol{y})$使用 Mahalanobis 距离度量:

$$d_k(\boldsymbol{y}) = (\boldsymbol{y}-\boldsymbol{\mu}_k)^{\mathrm{T}}\sum_k^{-1}(\boldsymbol{y}-\boldsymbol{\mu}_{\mathrm{k}}), \tag{9}$$

其中$\boldsymbol{\mu}_k$ 是 $\boldsymbol{Y}_k$ 中所有元素的均值, $\boldsymbol{\Sigma}_k$ 是它们的协方差矩阵.

3 实验

实验数据采集自一位女生和五位男生, 共 6 个数据集. 默声音符集合为{a,e,i,o,u,A,E,I,O,U}. 在日常办公室环境下, 从下颚与颈部间平坦区域采集默声音信号(图 2), 每个实验者读每个音符各 50 遍, 10 个音符共得到 500 个样本. 将来自同一个人的 500 个样本分为 50 组, 每组包含每个默声音符的样本各 1 个. 对于每个样本提取 22 维的参数矢量, 其中 20 维为 20 阶 Mel 倒谱系数, 1 维为样本持续时长系数, 1 维平均能量系数. 然后使用 LDA 方法进行特征选择, 取特征向量维数 m = 9. 采用 Leave-One-Out 交叉验证的方法, 即对于来自同一个

人的实验数据, 每次选择 1 组样本作为测试集, 其余所有 49 组样本作为训练集, 共进行 50 次识别, 然后综合统计识别率. 得到每个人数据的识别率如表 1 所示, 混淆矩阵见图 4.

表 1 和图 4 所示的实验结果表明, 所有数据集上的默声音符识别实验都得到了很高的识别率, 平均为 98.87%. 说明在我们选择的默声音符集下, 同一个人所读出默声音符信号之间具有较好的可分性. 如果采用商品化的采集设备并对软件进行优化, 或者对默声音信号做更深入的分析(例如, 考虑声学信号在身体组织中的传播特性[14]), 将会得到更高的识别率. 个人的发音习惯可能使得一些音符信号的模式较为接近, 导致识别结果出现轻微的混淆. 实验结果中出现的 V_i 单向地被误识为 V_j 的单向混淆现象, 如图 4(f)中 e 被误识为 a 而 a 未被误识为 e, 其原因是在特征空间中 V_i 与 V_j 各自样本均值间的欧式距离较近, 而在样本集的分散程度上 V_j 大于 V_i, 所以在使用式(9)比较一个未见样本属于 V_i 和 V_j 的可能性时, 相对于 V_i 来说, 更倾向将其预测为 V_j. 在我们的实验中, 长符的识别率显著高于短符, 这表明就我们的算法而言, 信号持续时间变长, 会使在整个信号段上提取的样本特征更稳定、更有可判别性.

4 默声音符编码

我们建立了默声音符编码文本输入实验系统. 受摩尔斯电码方法的启发, 我们使用第 1 节中选定的 10 个默声音符(长、短音符的时长区别, 由实验者在训练阶段按个人习惯决定)对 26 个英文字母、数字 0~9、若干标点符号、退格/确认等字符和控制命令进行无重码编码, 码长不超过 2. 为了方便记忆, 应该从字符的形状、读音、出现频率来考虑编码方案. 表 2 是本文采用的一种编码方案.

图 5 所示的是系统工作流程框图. 系统从采集的信号中检测出含有默声音符的片段, 识别出默声音符, 对默声音符流进行切分得到默声音符序列, 并解码为字符, 再通过视觉/听觉反馈, 实现文本确认. 本文实验中, 我们要求实验者读完一个字符编码后稍作停顿后再读下一个字符编码, 停顿的时间大约

表 1 默声音符识别实验结果

实验者	性别	口音	识别率
L C	女	山东	99%
L J	男	山东	98.6%
D Y	男	山西	99%
G Z	男	湖南	98.2%
Y M	男	江西	99%
T P	男	黑龙江	99.4%
平均			98.87%

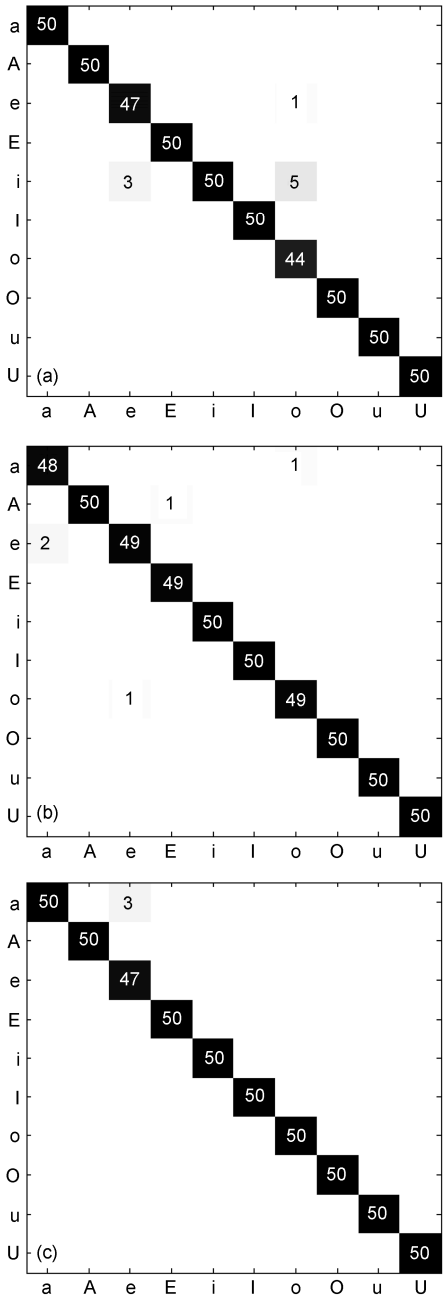

图 4 默声音符识别的混淆矩阵
(a) GZ; (b) YM; (c) TP

表 2 默声音符编码方案

字符	编码	字符	编码
a	a	u	u
b	io	v	ui
c	ii	w	uu
d	oi	x	UU
e	e	y	ue
f	eu	z	iA
g	oo	0	OO
h	eo	1	II
i	i	2	aa
j	ei	3	oe
k	Iu	4	iE
l	Ii	5	uo
m	uU	6	iO
n	Ou	7	aI
o	O	8	oa
p	Ia	9	iu
q	Ai	,	ee
r	A	.	OU
s	E	(空格)	O
t	I	(退格/确认)	U

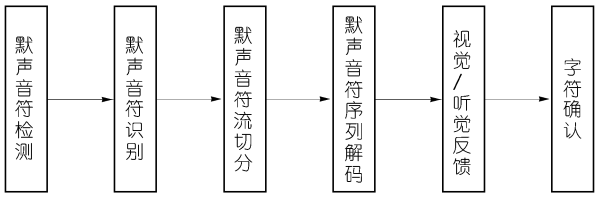

图 5 默声音符编码文本输入系统框图

1 s; 使用该停顿切分缓存的默声音符流, 得到默声音符序列; 把默声音符序列解码为字符, 并将字符对应的合成语音(无效编码对应错误提示音)通过耳机反馈给实验者, 由实验者确认. 我们采用如上方法, 输入“9876543210”耗时约 35 s, 输入“the quick brown fox jumps over a lazy dog.”, 共耗时约 170 s.

5 结论

本文提出了默声音符编码文本输入方法, 该方法选择默声音中容易发声、容易精准识别且彼此间容易区分的少数几个音素作为默声音符, 对各种字符进行编码. 类似摩尔斯电码, 用户只需要记住编码方式, 就可以通过读默声音符准确地实现文本输入. 本文选择国际音标中的单元音音素/ʌ/, /e/, /i/, /ɔ/, /u/以及它们对应的长音作为默声音符, 进行了真实场景下的默声音符识别实验和文本输入应用. 实验结果表明, 同一人读出的默声音符具有很好的可区分性, 使用简单的分类器即可准确地识别默声音符并解码出字符.

默声音符编码文本输入方法追求高度可靠地完成文本输入任务. 和语音输入一样, 它与周围声学环境互不影响也无需手眼参与, 它还能在环境嘈杂、需要保持安静等多种场景中正常工作. 同时, 该方法对计算资源要求较低, 尤其适合应用于下一代便携式个人计算终端.

参考文献

1 Deng L, Huang X D. Challenges in adopting speech recognition. Comm ACM, 2004, 1: 69–75
2 Nakajima Y, Kashioka H, Shikano K, et al. Nonaudible murmur recognition input interface using stethoscopic microphone attached to the skin. In: Proc. IEEE ICASSP, 2003. 708–711
3 Denby B, Schultz T, Honda K, et al. Silent speech interfaces. Speech Comm, 2010, 52: 270–287
4 Wester M, Schultz T. Unspoken speech-speech recognition based on electroencephalography. Master’s Thesis. Karlsruhe: Universität Karlsruhe (TH), 2006
5 Maier-Hein L, Metze F, Schultz T, et al. Session independent non-audible speech recognition using surface electromyography. In: IEEE Workshop on Automatic Speech Recognition and Understanding, 2005. 331–336
6 Nakajima Y, Kashioka H, Campbell N, et al. Nonaudible murmur (NAM) recognition. IEICE Trans Inform Systems, 2006, E89-D: 1–8
7 Heracleous P, Kaino T, Saruwatari H, et al. Unvoiced speech recognition using tissue-conductive acoustic sensor. EURASIP J Adv Signal Process,2007, 1: 1–11
8 Hirahara T, Otani M, Shimizu S, et al. Silent-speech enhancement system utilizing body-conducted vocal-tract resonance signals. Speech Comm, 2010, 52: 301–313
9 Otani M, Hirahara T, Shimizu S, et al. Numerical simulation of transfer and attenuation characteristics of soft-tissue conducted sound originating from vocal tract. Appl Acous, 2009, 70: 469–472
10 于水源. 汉语元音激励模式及其语音性质. 中国科学 F 辑: 信息科学, 2009, 39: 886–895

11 Duda R, Hart P, Stork D. Pattern Classification. New York: Wiley, 2000
12 Makhoul J. Linear Prediction: A Tutorial Review. Proc IEEE, 1975, 63: 561–580
13 Rabiner L, Juang B. Fundamentals of Speech Recognition. London: Prentice-Hall, 1993
14 徐泾平, 程敬之. 功率倒谱技术在呼吸系统声传播特性分析中的应用. 科学通报, 1995, 40: 1518–1521

A text input method based on silent voice symbol encoding

TENG Peng & JIA YunDe

Beijing Laboratory of Intelligent Information Technology, School of Computer Science, Beijing Institute of Technology, Beijing 100081, China

Speech input is expected to become the principal method of computer text input and replace the keyboard. But speech input is seriously degraded in nosiy environment, and often makes a leakage of information as well. To avoid these problems, this paper presents a text input method based on silent voice symbol encoding. We select several easy-to-pronounced and easy-to-recognized phonemes in silent voice as silent voice symbols (SVSs) to encode characters. Similar to Morse code, memorizing the encoding scheme, users can input text accurately by speaking relevant SVSs. Experimental results demonstrate that the proposed method is promising to be a reliable method of text input for many applications.

text input, silent voice symbol, silent communication, silent speech interface

doi: 10.1360/972010-1864