基于韵母发音事件匹配与位置时延分析的音唇一致性判决方法

朱铮宇[1,2],廖丽平[1],杨春玲[2],王 泳[1],蔡 君[1],邱华愉[1]
(1. 广东技术师范大学网络空间安全学院,广东广州 510665; 2. 华南理工大学电子与信息学院,广东广州 510641)

摘 要: 针对传统一致性判决方法主要对整句(段)话进行分析,并无对分析内容加以筛选,存在字典规模过大、计算复杂度高及结果易受静音等弱关联片段影响等不足,本文以唇型变化显著的韵母为代表性发音事件,结合音唇初始时延分布范围的统计结果,提出基于韵母发音事件匹配与位置时延分析的一致性判决方法. 先利用提出的音视频结合韵母切分法对字典学习数据进行韵母段筛选,再通过学习所得的韵母字典分析韵母事件的音唇匹配度,并对各韵母出现位置的时延分布进行统计评分. 最后由韵母发音事件音唇匹配度得分与位置时延分析评分相融合的评分机制判决一致性. 实验结果表明,本文算法在识别性能上优于多种比较算法,且与传统字典法相比降低了一定的运算量.

关键词: 一致性分析; 声韵母切分; 字典学习
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112 (2021)01-0140-09
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.12263/DZXB.20190238

Lip Motion and Voice Consistency Recognition Based on Audio-Visual Matching of Vowel Pronunciation Events and Position Delay Analysis

ZHU Zheng-yu[1,2], LIAO Li-ping[1], YANG Chun-ling[2], WANG Yong[1], CAI Jun[1], QIU Hua-yu[1]
(1. *School of Electronics and Information, Guangdong Polytechnic Normal University, Guangzhou, Guangdong* 510665, *China*;
2. *School of Electronic and Information Engineering, South China University of Technology, Guangzhou, Guangdong* 510641, *China*)

Abstract: For the mainstream lip motion and voice coherence judgment method, the whole sentence (segment) is analyzed without screening the content. This leads to large dictionary size and high computational complexity, and the result is vulnerable to weak related segments such as mute. Considering the vowel with significant lip shape changes as a representative pronunciation event and combining with the statistical results of the audio-visual initial delay distribution range, a consistent decision method based on audio-visual matching of vowel pronunciation events and position delay analysis is proposed. Firstly, the dictionary learning data is selected by the proposed audio-visual vowel segmentation method, and then the vowel dictionary is used to analyze the matching of the vowel event, and the time delay distribution of each vowel position is statistically scored. A consistency judgment is made by a scoring mechanism in which the vowel pronunciation event lip matching score and the position delay analysis score are combined. Experimental results show that the proposed method is superior to compared algorithms in recognition performance and reduces the amount of computation compared with the traditional dictionary method.

Key words: coherence analysis; initial/final segmentation; dictionary learning

1 引言

语音唇动一致性判决(分析)是指通过发音过程中唇部运动与音频变化之间的联系来判断音视频是否同时录制、出自同一人、同一句话. 其在多说话人视频中的当前说话人定位或分割[1],身份认证中的活体检测[2],人眼注意机制中的显著性区域检测[3]等领域有着广泛的应用.

收稿日期: 2019-03-03; 修回日期: 2020-05-10; 责任编辑: 王天慧
基金项目: 国家自然科学基金(No. 61672173); 广东省普通高校青年创新人才(No. 2018KQNCX140)

现有的一致性分析方法主要可归纳为四类:互信息法(Mutual Information,MI)、多元统计分析(Multivariate Statistical Analysis,MSA)、相关系数法,双模态稀疏表示.其中,MI 基于最大信息子空间投影原则生成概率模型,使得音视频特征在该空间内的线性组合获得最大互信息,并以此衡量音视频的关联度,较典型的有二次互信息(Quadratic Mutual Information,QMI)算法[4].MSA 通过为音视频两类多维异构特征寻求不同的映射向量,使得两者投影后在相应的统计标准下获得最大值,再将待分析的音视频数据投影到解出的映射向量上获得两者的关联度,典型的 MSA 方法有典型相关分析(Canonical Correspondence Analysis,CCA)[5]和协惯量分析(Co-Inertia Analysis,CoIA)[6].相关系数法则通过分析各帧音频和视频特征各维间的相关系数,并以此为新特征结合分类模型等方式进行一致性判决,BLPM(Bimodal Linear Prediction Model)模型[7]和 NCC + SVM 法[8]便是基于上述思想.前三类方法把唇动过程拆分成帧为单位的相互独立个体进行分析,但唇动是时变的连贯过程,这样处理难以体现帧间联系,易丢失唇动的时变细节信息.针对以上问题,文献[9]在 CoIA 分析的基础上,引入唇形变化与音频幅度变化的时域相关度,并通过融合时空域相关度进行一致性分析.鉴于语言由不同音节组成,各音节的音频及口型序列会在不同词句中反复出现,有研究者将移不变稀疏表示中原子的概念引入到表征发相同音时音唇变化存在的共性及对应关系中[10,11],这些由连续帧组成的视频原子更好地保留了各音节唇形连续变化的细节信息.文献[12]基于此思想,通过改进的联合字典学习算法,无监督地训练出时空移不变的双模态字典,并以此作为表征不同音节发音时音唇同步变化关系的模板来进行一致性判决,但没对字典学习数据加以限制,使得原子表征范围过大且数量庞大,并出现部分无意义原子(如噪声、静音),也导致字典学习及一致性分析过程较为繁琐.综上所述,现有一致性分析方法主要是对整句(段)话进行分析,并无对分析的内容加以筛选,然而句子各部分提供信息的重要性各有不同,如由统计类算法获得音唇间的相关度,但句子不同成分的相关性会有所差异.静音、噪音以及部分辅音等嘴型变化不明显的片段,由于音频幅度和唇宽高的正比关联性[13],其音唇关联度并不显著,为弱关联片段,而这类似于信号中的噪声成分会给分析带来一定影响.同时,时延在一致性分析中的重要性也逐渐受到关注,文献[14]和文献[15]探讨了语音唇动间时延的主观感知范围,但都是基于对音视频信号加入不同时延后,让测试者进行评价获得.然而[-125ms, +45ms]范围内的时延,人的感知已难以觉察.现时,时延的相关研究主要针对感知范围开展,对于难以感知的音唇一致数据的初始时延分布却缺乏具体的界定,一致和不一致数据在时延上的差异也少有文献报告.

汉语单个字是一个音节,由声母、韵母和声调构成,若以音节为单位,原子数量大且缺乏代表性.考虑到汉语中韵母发音时长较大,同时音频能量高于声母,即使搭配不同声母,韵母的唇动规律也非常稳定[16].因此,针对前述问题,本文以汉语为研究重心,寻求以分析更具代表性的韵母发音事件代替整个句子实现一致性判决,结合对初始时延分布范围的统计分析,提出一种基于韵母发音事件匹配及其位置时延分析的语音唇动一致性判决方法.该方法以双模态移不变稀疏表示模型分析音视频信号,为生成更具针对性的移不变双模态韵母字典,先对训练数据进行韵母段切分,再进行字典训练.并针对韵母数据的自动切分问题,提出一种音视频结合的韵母发音事件切分方法.最后,利用学习所得韵母字典分析韵母事件的音唇匹配度,同时对各韵母出现位置的时延分布进行评分,提出韵母发音事件音唇匹配度得分与位置时延分析评分相融合的一致性评分机制,由融合后的最终得分判断一致性.

2　音视频结合的韵母发音事件分割

2.1　音节唇动序列的两步切分法

鉴于用唇动信息切分音节在鲁棒性和准确度上稍优于音频[17],本文两步切分法通过分析唇部视频帧序列实现音节切分:(1) 先对唇部序列进行“粗切分”,由视觉语音检测(Visual Speech Detection,VSD)找出唇动起止点;(2) 对起止点内的唇部序列进行“细切分”,切分出各音节的唇动序列,具体步骤如下.

步骤 1　由于唇内阴暗面积大,唇张开和闭合时低灰度值像素点的数量有较大差异,而灰度值统计的运算量较小,因此采用唇部区域像素统计分析法[18]进行 VSD,通过对逐帧帧移的滑动窗内视频进行灰度统计来判断是否唇动片段.设第 l 帧图像中低于某灰度阈值 J 的像素点数量为 v_l,则滑动窗内的 R 帧唇部图像可用 R 维随机向量 $\boldsymbol{V}=\{v_l\}_{l=1}^{R}$ 表示,v_l 定义为:

$$v_l = \sum_{j=0}^{J} \xi_l(j) \tag{1}$$

式(1)中 $\xi_l(j)$ $(j=1,2,\cdots,255)$ 为该帧的灰度直方图分布.若 v_l 和 w_l 分别为张嘴和闭嘴时灰度低于 J 的像素点数量,张闭嘴导致的像素点数量差异为 $\tilde{v}_l$,则 $v_l=\tilde{v}_l+w_l$.假定 $\tilde{v}_l$ 和 w_l 服从高斯分布,以 H_0 表示唇闭状态,H_1 表示唇张状态,则 $\boldsymbol{V}$ 的状态统计可写成以下似然比:

$$L(\boldsymbol{V})=\frac{p(\boldsymbol{V}\mid H_1)}{p(\boldsymbol{V}\mid H_0)}\begin{cases}\geqslant\varphi, \text{判为 } H_1\\ <\varphi, \text{判为 } H_0\end{cases} \tag{2}$$

根据奈曼皮尔逊准则,取误警率 $P_{FA}=p(H_1\mid H_0)=\gamma$

时,正确检测率最大的 φ 值为阈值. 由于 $p(\boldsymbol{V}|H_0)\sim N(\boldsymbol{0},\sigma^2\cdot\boldsymbol{I})$ 和 $p(\boldsymbol{V}|H_1)\sim N(\mu_{\hat{v}},(\sigma_{\hat{v}}^2+\sigma^2)\cdot\boldsymbol{I})$($\boldsymbol{I}$ 为单位向量,$\boldsymbol{0}$ 为全零向量),将以上分布函数代入式(2),取对数化简得:

$$L(\boldsymbol{V})=R\mu_{\hat{v}}\cdot\frac{1}{R}\sum_{l=0}^{R-1}v_l+\frac{\sigma_{\hat{v}}^2}{2\sigma^2}\sum_{l=0}^{R-1}v_l^2 \tag{3}$$

γ 确定时,由分布函数难以同时求解 $\mu_{\hat{v}},\sigma_{\hat{v}},\sigma$ 和 φ,可令:

$$L_1(\boldsymbol{V})=\frac{1}{R}\sum_{l=0}^{R-1}v_l \tag{4}$$

$$L_2(\boldsymbol{V})=\sum_{l=0}^{R-1}v_l^2 \tag{5}$$

则式(3)表示为:

$$L(\boldsymbol{V})=\alpha L_1(\boldsymbol{V})+\beta L_2(\boldsymbol{V}) \tag{6}$$

其中,$\alpha=R\mu_{\hat{v}},\beta=\sigma_{\hat{v}}^2/2\sigma^2$. 设阈值 φ_1 和 φ_2,当 $L_1(\boldsymbol{V})>\varphi_1$ 且 $L_2(\boldsymbol{V})>\varphi_2$ 时,$\boldsymbol{V}$ 为张嘴状态,两阈值可分别由以下两式确定:

$$\varphi_1=\sqrt{\frac{\sigma}{R}}Q^{-1}(P_{FA}) \tag{7}$$

$$\varphi_2=\sigma^2Q_{\chi_N^2}^{-1}(P_{FA}) \tag{8}$$

式(7)和式(8)中,$Q^{-1}(\cdot)$ 为右侧尾概率函数的反函数,$Q_{\chi_N^2}^{-1}(\cdot)$ 为 χ^2 分布. 由于 w_n 非零均值,因此须先估计 σ,可通过对视频前几帧的唇部闭合数据(无声段)采用"估值-插入"法[19]进行估算. 对于灰度阈值 J,先设为滑动窗(窗长 $R=10$)内首帧唇部图像的平均灰度值,再逐步调整 J 直到 $v_0>0$. 第 l 段窗内的唇状态 $V_{\mathrm{Lip}}(l)$ 为:

$$V_{\mathrm{Lip}}(l)=\begin{cases}1, & \text{第 } l \text{ 段为唇张开}\\ 0, & \text{第 } l \text{ 段为唇闭合}\end{cases} \tag{9}$$

设唇动帧计数器 counter,若 $V_{\mathrm{Lip}}(l)$ 为 1 则 counter = counter + 1,否则 counter = 0. 最后,通过判断 counter 是否小于最小无唇动片段长度来获得各唇动段起始和结束帧的位置 l_{start}、l_{end}.

步骤 2 步骤 1 虽可快速切分出唇动和无唇动的视频序列,但无法细分到每个音节,特别对 $l_{\mathrm{start}}-l_{\mathrm{end}}\gg 25$ 帧的唇动段,可能包含多个音节,因此须对"粗切分"后超过特定长度的唇动段作"细切分". 对超长片段 $[l_{\mathrm{start}},l_{\mathrm{end}}]$ 内的各帧唇部区域提取唇面积特征,并通过前后向滤波[20]对唇面积变化轨迹进行平滑处理,使得能更准确地提取波形的波峰和波谷位置(如图 1、图 2 所示). 其中,波峰位置对应音节内唇张开最大的时刻,而波谷位置为音节间的切换,按此方法进行搜索可分割出段内各音节的起始位置 $[l_{\mathrm{Cstart}},l_{\mathrm{Cend}}]$.

2.2 音视频结合的声韵母分割

在自动声韵母边界检测中,基于短时能量和过零率的方法[21],及基于听觉事件检测的方法[22],因高效而受到研究者的关注. 但前者易受噪声等环境因素影响;后者虽对噪声有较好的鲁棒性,但切分结果易受复合韵母韵腹或韵尾的音变(如变调、儿化等)所影响. 因此,在上述方法的基础上,本文结合唇动信息及汉语发音规律,对听觉事件检测范围作进一步界定,避开音变发生概率较大的区域,提出一种音视频结合的韵母分割方法,具体流程如下.

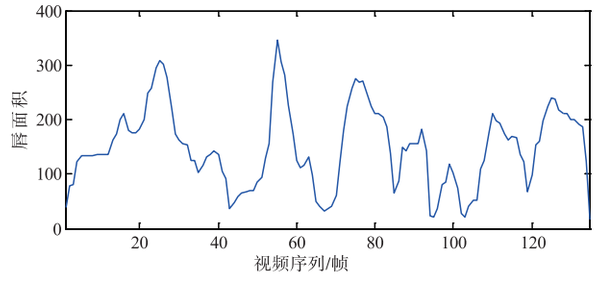


图1 滤波平滑前唇动序列

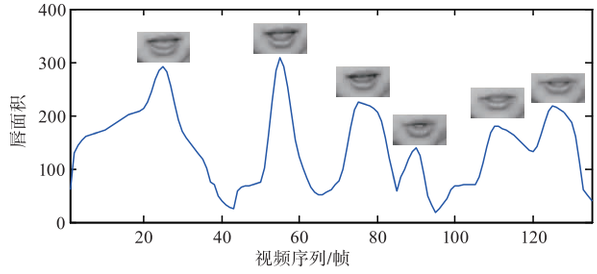


图2 滤波平滑后唇动序列

步骤 1 确定语音段各音节起止位置. 两步切分法定位出各语音段的起始点 $[l_{\mathrm{start}},l_{\mathrm{end}}]$,以及段内各音节的位置 $[l_{\mathrm{Cstart}},l_{\mathrm{Cend}}]$,音频采样点 n 和视频帧 l 之间的对应关系为:

$$n\in[\lambda\cdot(l-1)+1,\lambda\cdot l] \tag{10}$$

$$l=\lceil n/\lambda\rceil \tag{11}$$

式(10)和式(11)中 λ 为音频与视频采样率的比值,"$\lceil\ \rceil$"表示正向取. 由以上关系可得音频各音节的起止范围 $[n_{\mathrm{Cstart}},n_{\mathrm{Cend}}]$,为避免漏检,对起始点取式(10)范围的上限,而结束点取下限.

步骤 2 对语音信号进行浊音检测. 先对唇动序列 $[l_{\mathrm{start}},l_{\mathrm{end}}]$ 对应的 $[n_{\mathrm{start}},n_{\mathrm{end}}]$ 内的语音信号进行浊音检测. 由于浊音帧和清音帧在不同频段的能量分布有显著差异:前者的能量主要集中在低频,而后者则集中在高频. 本文通过短时能量,短时平均过零率和梅尔标度映射频域能量分析相结合进行清浊音切分. 先利用各帧短时能量 E_m 和短时平均过零率 Z_m 对语音帧(m 为帧索引)进行浊音帧初筛选,对满足 $E_m>E_T$ 且 $Z_m<Z_T$(E_T 和 Z_T 分别为短时能量和过零率阈值)的语音帧利用梅尔标度映射频域能量分析进行检测,判别出浊音和清音帧并根据结果修正 Z_T 和 E_T,流程如图 3 所示. 由于不同频点上能量的贡献不同,结合梅尔标度定义,第 i 点

频率处能量的贡献度为:

$$\psi_i=\frac{\lg\left(i*\frac{F_S/O}{1000}+1\right)-\lg\left((i-1)*\frac{F_S/O}{1000}+1\right)}{\lg\left(\frac{F_S/O}{1000}+1\right)} \tag{12}$$

其中,F_S为音频采样率,O 为单帧采样点数. 设语音信号 i 点频率对应的能量为 P_i,则低频和高频部分的能量分别为:

$$E_{\text{Low}}=\sum_{j=1}^{z}P_i\psi_i \tag{13}$$

$$E_{\text{High}}=\sum_{j=z+1}^{N}P_i\psi_i \tag{14}$$

式(13)和(14)中,z 为高低频分界点,可通过令步骤(1)切分的无声段中 $E_{\text{Low}}=E_{\text{High}}$ 求出. $E_{\text{Low}}>\zeta E_{\text{High}}(\zeta>1)$时,低频分量占主导地位,当前帧判为浊音帧. 由于短的非浊音段不可能出现在连续浊音段之间,可据此进行分类后处理[23]确定浊音段的起止范围$[n_{\text{Sstart}},n_{\text{Send}}]$,本文取 $\zeta=1.7$.

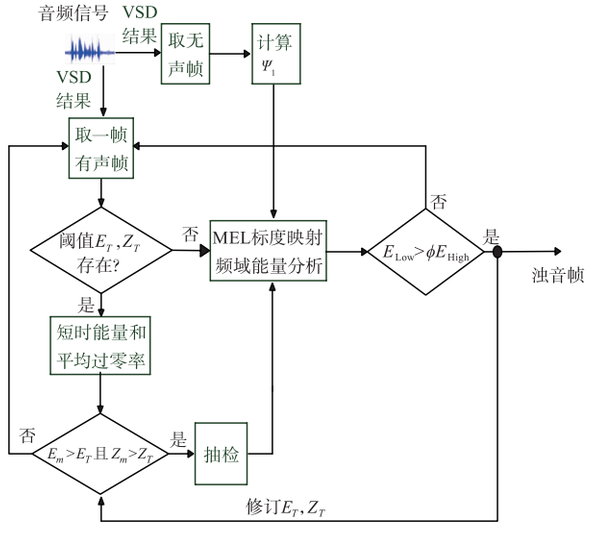


图3　浊音检测流程图

步骤 3　结合步骤 1 和步骤 2 的结果定位声韵母分割点. 由于汉语韵母为浊音,而声母有清音、浊音([m],[n],[l],[r]为浊音声母)和无声段三类. 无声段在 step1 中已去除,但仍须对步骤 2 检出的浊音段$[n_{\text{Sstart}},n_{\text{Send}}]$作成分分析. 当声母发清音时,浊音段只有韵母,声母在浊音段前;而发浊音时,声母在浊音段内,浊音段同时包含声母和韵母. 根据以上两种情况,先界定韵母分界点的搜索范围,再利用听觉事件检测法检测分界点,切分出韵母段. 声韵母分界点检测包含三个子步骤.

子步骤 1　清音声母检测. 在浊音段$[n_{\text{Sstart}},n_{\text{Send}}]$的 n_{Sstart}处往时轴反方向找最近的音节起始点 n_{Cstart},并在$[n_{\text{Cstart}},n_{\text{Sstart}}]$内检测是否存在上一段浊音段的结束点. 若无,在 0kHz ~ 0.4kHz 频段上检测正突变事件发生点,定位清浊音分界;若有,在上一段浊音结束点到当前浊音段的起始点内进行上述正突变事件检测. 索引为 c 的频段内,语音信号 $s(n)$ 出现正突变听觉事件 e_{on} 定义为:

$$e_{\text{on}}(n)=\begin{cases}1, & \text{if}\quad s(n)>\theta_{\text{on}}(c)\\ 0, & \text{else}\end{cases} \tag{15}$$

阈值 $\theta_{\text{on}}(c)$由下式确定:

$$\theta_{\text{on}}(c)=\mu(c)+para\times\sigma(c) \tag{16}$$

其中,$para$ 为待定系数,取值 0.7,$\mu(c)$和 $\sigma(c)$分别为差分信号 $s(n+1)-s(n)$的均值和方差. 最后,选时轴正向最靠后的正突变事件出现位置 n_{Boundary}为分界点,则$[n_{\text{Sstart}},n_{\text{Send}}]$为韵母段,而$[n_{\text{Boundary}},n_{\text{Sstart}}]$为声母段.

子步骤 2　浊音声母检测. 若上述范围内无检测到正突变事件,可推断声母发浊音,此时检测浊音段起始点 n_{Sstart}与时轴正向最近的音节结束点 n_{Cend}内是否存在当前浊音段的结束点,若无,则为浊声母范围;若有,对上述范围在 0.4kHz ~ 4kHz 频段检测正突变事件发生点,定位浊辅音声母和韵母的分界,并选取时轴正方向最靠后的正突变事件位置 n_{Boundary}为分界点,则$[n_{\text{Sstart}},n_{\text{Boundary}}]$为声母段,$[n_{\text{Boundary}},n_{\text{Send}}]$为韵母段.

子步骤 3　零声母音节检测. 若上述情况均无检测到正突变事件,则浊音段$[n_{\text{Sstart}},n_{\text{Send}}]$内无声韵母分界点,判为零声母音节,整段为韵母.

3　基于韵母发音事件匹配及位置时延分析的一致性判决

3.1　韵母发音事件匹配

移不变稀疏表示(Shift Invariant Sparse Representation,SISR)以包含时序及空间信息的移不变原子为片段模板,能更好地反映唇动过程的时空特性,且训练出的成对音视频联合原子能更准确地描述语音唇动间的对应关系[11]. 故本文采用 SISR 模型描述音唇信号,并通过韵母分割算法从音唇一致训练数据中切分出韵母段,实现对训练数据的筛选,将训练范围限制为韵母发音单元. 为增强音频原子对韵母特性的表征能力,以时频域分析语音信号,筛选后的音唇信号为:

$$\boldsymbol{S}'=(\boldsymbol{S}'_a(m,\omega),\boldsymbol{S}'_v(x,y,l)) \tag{17}$$

式(17)中 $\boldsymbol{S}'_a(m,\omega)$为切分后语音信号的短时傅里叶变换,$m$ 为音频帧索引,ω 为第 m 帧频谱的频率索引,$\boldsymbol{S}'_v(x,y,l)\in\Re^{X\times Y\times L}$($X$、$Y$ 和 L 分别为视频帧尺寸和总帧数)为切分后的唇动信号,x 和 y 为像素坐标,l 为视频帧索引. 音频信号沿时轴平移时,与 m 相关的频率成分也发生时移,但不发生频移. 因此,$\boldsymbol{S}'_a(m,\omega)$可简化为

$S'_a(m)\in\Re^{M\times W}$($M$ 为总帧数,W 为频率范围),则音唇信号可写成以下移不变稀疏表示形式:

$$\begin{pmatrix}\boldsymbol{S}'_a(m)\\ \boldsymbol{S}'_v(x,y,l)\end{pmatrix}=\sum_{k=1}^{K}\begin{pmatrix}\sum_{\hat{m}} c'^{(a)}_{k\hat{m}}T'^{(a)}_{(\hat{m})}\boldsymbol{d}'^{(a)}_k\\ \sum_{\hat{x},\hat{y},\hat{l}} c'^{(v)}_{k\hat{x}\hat{y}\hat{l}}T'^{(v)}_{\hat{x}\hat{y}\hat{l}}\boldsymbol{d}'^{(v)}_k\end{pmatrix}=\sum_{k=1}^{K}\begin{pmatrix}\sum_{\hat{m}} c'^{(a)}_{k\hat{m}}\boldsymbol{d}'^{(a)}_k(m-\hat{m})\\ \sum_{\hat{x},\hat{y},\hat{l}} c'^{(v)}_{k\hat{x}\hat{y}\hat{l}}\boldsymbol{d}'^{(v)}_k(x-\hat{x},y-\hat{y},l-\hat{l})\end{pmatrix}\tag{18}$$

式(18)中,而 $c'^{(a)}_{k\hat{m}}$ 和 $c'^{(v)}_{k\hat{x}\hat{y}\hat{l}}$ 分别为信号中$(\hat{m},\hat{x},\hat{y},\hat{l})$位置的第 k 个音频和视频原子的稀疏系数,$\boldsymbol{d}'^{(a)}_k\in\Re^{\tilde{M}\times W}$($M>>\tilde{M},k=1,2,\cdots,K$)为音频原子,$\tilde{M}$ 为原子包含的帧数. 而 $\boldsymbol{d}'^{(v)}_k\in\Re^{\tilde{X}\times\tilde{Y}\times\tilde{L}}$($X>>\tilde{X},Y>>\tilde{Y},L>>\tilde{L}$)为视频原子,$T'^{(a)}_{(\hat{m})}$ 和 $T'^{(v)}_{(\hat{x}\hat{y}\hat{l})}$ 分别为第 k 个音视频原子的平移变换算子,$\hat{x}$ 和 $\hat{y}$ 为视频原子在像素空间上的平移量,$\hat{m}$ 和 $\hat{l}$ 分别为音频和视频原子在时轴上的平移量,m 和 l,及 $\hat{m}$ 和 $\hat{l}$ 之间的关系由式(10)和式(11)求得. 式(18)的音唇信号可由多模态移不变 K-SVD 算法[10,24]训练出韵母字典 $\boldsymbol{d}'_k=(\boldsymbol{d}'^{(a)}_k(m),\boldsymbol{d}'^{(v)}_k(x,y,l))$.

由于待测数据的音唇一致性未知,这以音频部分为参考模态进行处理. 先利用 2.3 中的步骤 2 对待测信号的音频进行清浊音检测,再由改进的移不变稀疏表示匹配追踪(SISR Matching Pursuit,SISR-MP)算法[12]对浊音段音频信号进行分解,得到各选中音频韵母原子 $\boldsymbol{d}'^{(a)}_{g,k}$(下标表示第 g 个选中原子的索引 $k,g=1,\cdots,G$)在待测信号中的位置 $\hat{m}_{g,k}$. 若是音唇一致数据,则与视频原子 $\boldsymbol{d}'^{(v)}_{g,k}$ 相似的唇部运动过程会出现在待检测句子视频部分的第 $\hat{l}_{g,k}$ 帧(与 $\hat{m}_{g,k}$ 对应)附近位置,将唇动序列的时域位置检索范围定义在视频部分的$[\hat{l}_{g,k}-10,\hat{l}_{g,k}+10]$内,并将人脸下半部分作为像素空间上唇动序列的搜索范围,寻找该范围内与视频原子 $\boldsymbol{d}'^{(v)}_k$ 相似度的最大值. 前 G 个发音事件的匹配度得分 δ_1 可由下式求得:

$$\delta_1=\frac{1}{G}\sum_{g}^{G}\underset{l\in[\hat{l}_{g,k}-10,\hat{l}_{g,k}+10]}{\mathrm{Max}}\langle T_{-(\hat{x}\hat{y}\hat{l})}\boldsymbol{S}'_v,\boldsymbol{d}'^{(v)}_{g,k}\rangle\tag{19}$$

式(19)中,"$\langle\cdot\rangle$"为内积.

3.2 位置时延分析

由于发音时唇动稍先于语音,以及采集设备、环境等因素影响[25],即使音唇一致数据的音视频也并非完全同步,这种同一时轴上音频和视频信号在对应时间点上出现的合理偏移称为初始时延. 本文时延估计主要针对短时长、少样本的音视频数据,而 CoIA 算法[6]在这方面有更好的鲁棒性,故采用此法进行估计. 设语音和唇动数据的样本均为 N 帧,两者时延为 τ_{av} 帧,时延搜索范围为$[-D_{\mathrm{left}},+D_{\mathrm{right}}]$. 以音频为参考模态,去掉其前 D_{left} 帧和后 D_{right} 帧数据,对余下 $N\text{-}D_{\mathrm{left}}-D_{\mathrm{right}}$ 帧提取特征构成矩阵 $\boldsymbol{F}_a$,从视频中每次取相同帧数的数据,提取特征构成视频特征集合 $\boldsymbol{F}=\{\boldsymbol{F}_v^{(\tau)}\}(\tau=1,2,\cdots,D_{\mathrm{left}}+D_{\mathrm{right}}+1)$. 由 CoIA 算法可得 $\boldsymbol{F}_a$ 和各 $\boldsymbol{F}_v^{(\tau)}$ 之间的相关度,从而求得关联性系数曲线 $\rho_{\mathrm{av}}(\tau)$,时延 τ_{av} 为:

$$\tau_{\mathrm{av}}=\underset{\tau\in[1,D_{\mathrm{left}}+D_{\mathrm{right}}+1]}{\mathrm{argmax}}\{\rho_{\mathrm{av}}(\tau)\}-(D_{\mathrm{left}}+1)\tag{20}$$

精确到 ms 为单位的子帧时延 $\hat{\tau}_{\mathrm{av}}$:

$$\hat{\tau}_{\mathrm{av}}=T_{\mathrm{v}}\cdot\left(\tau_{\mathrm{av}}+\frac{0.5\cdot(\rho(\tau_{\mathrm{av}}-1)-\rho(\tau_{\mathrm{av}}+1))}{\rho(\tau_{\mathrm{av}}-1)-2\cdot\rho(\tau_{\mathrm{av}})+\rho(\tau_{\mathrm{av}}+1)}\right)\tag{21}$$

其中 T_{v} 为视频帧间间隔(ms),时延估算过程如图 4 所示.

本文位置时延分析以韵母原子在句中的出现位置 $\hat{m}_{k_g}$ 为参考点,同时截取$[\hat{m}_{k_g},\hat{m}_{k_g}+\tilde{M}]$范围内的音视频数据,通过时延估算获得各韵母音频原子对应位置分段的时延 τ_{k_g}. 对于音唇一致的数据,各 $\hat{m}_{k_g}$ 点附近的时延值 τ_{k_g} 会非常接近且处于合理时延范围内,而音唇不一致数据则恰恰相反. 最后,G 个韵母发音事件的时延分布得分 δ_2 定义为:

$$\eta(g)=\begin{cases}1, & \tau_{k_g}\in[-\zeta_{\mathrm{Left}},\zeta_{\mathrm{Right}}]\\ 0, & \tau_{k_g}\notin[-\zeta_{\mathrm{Left}},\zeta_{\mathrm{Right}}]\end{cases}\tag{22}$$

$$\delta_2=\frac{1}{G}\sum_{g=1}^{G}\eta(g)\tag{23}$$

式(22)合理时延范围 ζ_{Left} 和 ζ_{Right} 的取值由时延统计分析结果确定.

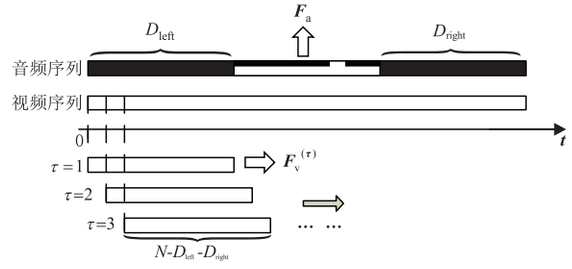


图4 时延估计过程

3.3 一致性判决

语音唇动一致的音视频信号属于同一时轴,其各位置上的唇动和语音变化相匹配,同时不同位置的音视频初始时延也应该非常接近,而且均属于一致性合理时延范围内. 基于以上思想,本文通过基于高斯混合模型(Gaussian Mixture Model,GMM)的贝叶斯融合法,将音频韵母发音事件匹配程度得分 δ_1 和各相应位置的时延分布得分 δ_2 相融合,由融合后的最终得分来判决一致性. 假设语音唇动事件的状态为 $\Lambda_r(r=0,1)$,其中 Λ_0 表示语音唇动不一致,Λ_1 表示语音唇动一致,$\boldsymbol{\delta}=(\delta_1,\delta_2)$ 为得分向量. 结合贝叶斯准则,由条件概率 $P\{\boldsymbol{\delta}|\Lambda_0\}$

和 $P\{\boldsymbol{\delta}|\Lambda_1\}$ 可得最终的一致性判决标准为:

$$\frac{P\{\boldsymbol{\delta}|\Lambda_1\}}{P\{\boldsymbol{\delta}|\Lambda_0\}}\begin{cases}\geqslant\varepsilon, & \text{判为 } \Lambda_1\\ <\varepsilon, & \text{判为 } \Lambda_0\end{cases} \tag{24}$$

分别采用两个高斯模型描述上述两个条件分布:

$$P\{\boldsymbol{\delta}\mid\Lambda_r\} = \sum_{i=1}^{N} w_i^{(\Lambda_r)} \frac{\exp\left\{-\frac{1}{2}(\boldsymbol{\delta}-\mu_i^{(\Lambda_r)})^T(\Sigma_i^{(\Lambda_r)})^{-1}(\boldsymbol{\delta}-\mu_i^{(\Lambda_r)})\right\}}{(2\pi)^{\mathrm{DI}/2}\left|\sum_i^{(\Lambda_r)}\right|^{1/2}} \tag{25}$$

其中,DI 为特征向量维度,而 $w_i^{(\Lambda_r)}$,$\mu_i^{(\Lambda_r)}$ 和 $\Sigma_i^{(\Lambda_r)}$ 分别为模型 Λ_r 第 i 个高斯分量的权重系数、均值和协方差矩阵,确定以上三个参数便可以确定此 GMM 模型. 可先利用 LBG 算法对模型进行初始化,再由期望最大化(Expectation Maximization,EM)算法训练相应的 GMM 模型. 本文整个一致性分析流程如图 5 所示. 时延估计时,对唇部区域(56×32)进行分块二维离散余弦变换,各块取 Zig-Zag 排序后最大系数并拼接上其一二阶差分组成 84 维视觉特征. 音频帧长取 20ms,帧间重叠 10ms,各帧提取对数能量及 13 维梅尔频率倒谱系数(Mel Frequency Cepstrum Coefficient,MFCC)同样拼接其一二阶差分组成 42 维音频特征.

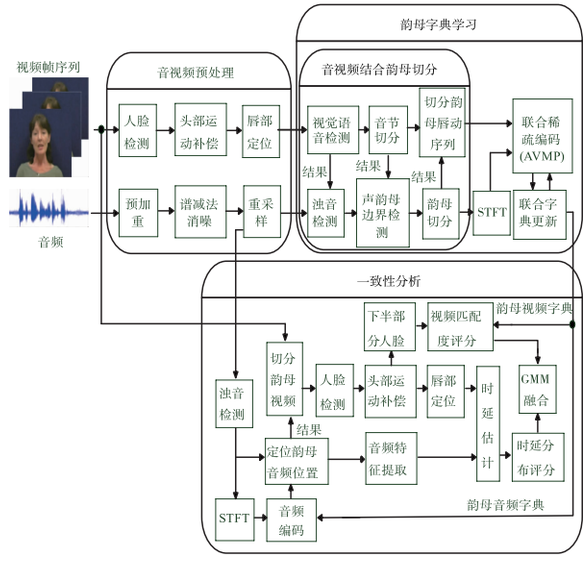


图5　整个一致性判决算法的流程框图

4　实验结果及分析

4.1　时延分布实验

实验通过对不同设备录制的多个音视频数据库进行时延统计,分析初始时延的合理分布范围,同时对多种不一致数据的时延分布进行分析,时延估算采用 3.2 节方法. 音唇一致数据分别来自 VidTIMIT,CUAVE,中文通用库以及实验室自建库,共 2600 句,样本分布见表 1 所示. 本文结合实际应用,总结出四类不一致数据(见表 2 所示),并由数据库中不同句子的音频和视频交叉组合模拟出这四类数据,每类 1000 个样本. 其中,第四类数据最接近一致数据,一致性判别难度最大. 根据音视频描述内容是否相同将四类不一致数据分两种情形进行统计:(1)音视频描述内容不同(表2 中的第一、二类)(2)音视频描述内容相同(表 2 中的第三、四类). 实际上,对于情形一的两类不一致数据由于其音视频描述的内容已不一致,因此它们的"时延"主要指时延估算的结果,实验中时延搜索边界设为 $D_{\text{left}}=D_{\text{right}}=400\text{ms}$.

表 1　语音唇动一致数据样本分布

数据库	样本数量/句	单句时长/s	数据总时长/min	语言
VidTIMIT	400	4s ~ 6s	27	英文
CUAVE	600	2s ~ 4s	35	
自建数据库	200	1.5s ~ 6s	15	
中文通用数据库	1400	4s ~ 7s	161	中文

表 2　语音唇动不一致数据分类

不一致据种类	说明
第一类	语音及唇动数据来自不同的人,描述内容亦非同一句话
第二类	语音及唇动数据来自同一人,描述内容非同一句话
第三类	语音及唇动数据来自不同的人,描述内容为同一句话
第四类	语音及唇动数据来自同一人,描述内容亦为同一句话,但非同一时刻录制

一致数据的统计结果如图 6(a)所示,其时延集中分布在[-50ms, +50ms]范围内,处于人主观感知范围以外,基本很难觉察,且取负值较多,即音频滞后于唇动的情况居多. 不一致数据的时延分布如图 6(b)和(c)所示. 由图可知,不一致数据的时延分布虽没有一致数据集中,但两种情形的不一致数据均多数分布在[-50ms, +50ms]的两边,其中以情形一更为突出,由于该情形下音视频描述的内容并不一致,因此其相关系数峰值很少出现在零时延位置附近. 而情形二由于音视频描述的内容一致,因此音频与唇动之间存在一定的联系,特别是第四类数据,由于个人发音的相对稳定性,虽是不同时刻录制的数据,但同一人同一句话的唇部运动过程有很高的相似性. 有研究认为音唇间在同一时轴上约只有 3 ~ 4 帧的偏移[26],这也解析了第二种情形下时延在[-230ms, -140ms]和[+100ms, +180ms]内较为集中的原因. 根据以上结果,韵母位置时延分析中的搜索范围 D_{left} 和 D_{right} 均取 100ms,而 ζ_{Left} 和 ζ_{Right} 均取 50ms.

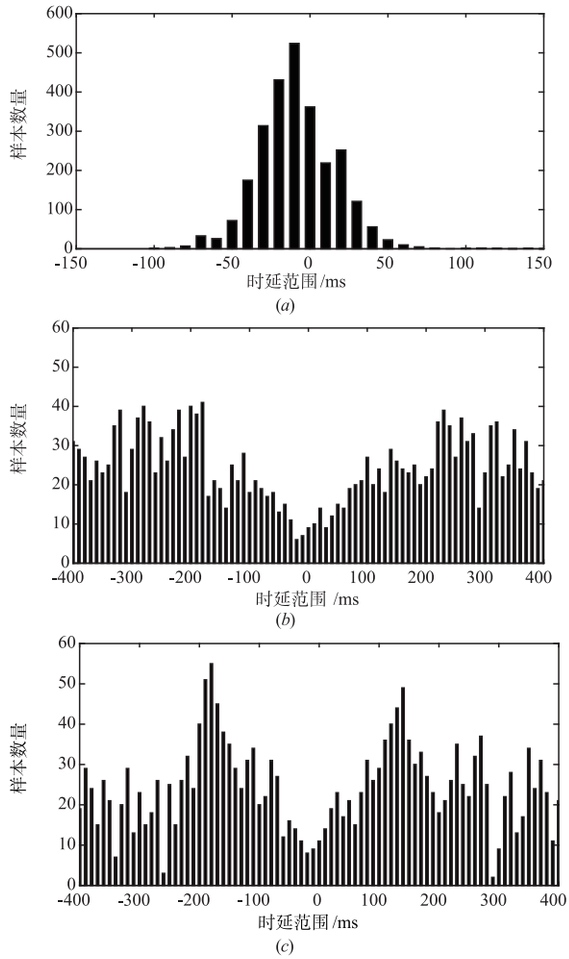


图6 时延范围统计分布图

4.2 一致性判决算法性能分析

通过本文方法与相关法中的 BLPM[7] 和 NCC + SVM[8] 法,QMI 算法[4],文献[9]的时空融合法,多元统计分析法中的 CCA[5] 算法和 CoIA[6] 算法结合文献[9]时延评分机制后的结果,以及文献[12]的传统联合字典法进行比较分析,验证本文采用韵母字典并结合位置时延分析在降低运算量和提高分析准确度上的有效性. 本实验的数据来自中文通用数据库,库中一致数据分为测试集 1200 句和训练集 600 句,并合成出四类不一致数据各 2000 句,训练集主要用于字典及 CCA 和 CoIA 算法的映射矩阵训练. 本文方法联合字典设置见表 3 所示. 从训练数据中根据语料挑选包含韵母较多的 190 句,共 35386 帧,作为本文方法的字典训练数据. 音频数据先降采样到 16kHz,视频采样率 30Hz,将音频帧长设为 512(32ms),帧移为 133(8.3ms),则音频采样率为 120Hz(帧/秒),$W=521/2+1=257$,联合字典初始化数据选自韵母发音单元库. 虽然韵母的总体平均时长为 167.3ms,但其中带鼻尾类韵母的平均时长较大为 225.3ms[27],因此将视频原子长度 $\tilde{L}$ 设为 8 帧. 实验以检测错误折中(Detection Error Tradeoff,DET)曲线和等误识率(Equal Error Rate,EER)作为算法性能评价标准.

表 3 联合字典参数设置

相关参数	字典大小 K	音频原子 $\tilde{M}\times W$	视频原子 $\tilde{X}\times\tilde{Y}\times\tilde{L}$	迭代次数 I	最大投影次数 P
参数设置	40:10:100	32×257	32×48×8	100	2300

实验表明,文献[12]方法取 $K=225$ 时检测效果较优,而本文方法取 $K=80$ 时效果最优,并以此结果进行比较. 本文方法及各比较方法的总体 DET 曲线如图 7 所示,各类不一致数据的单独统计结果见表 4 所示.

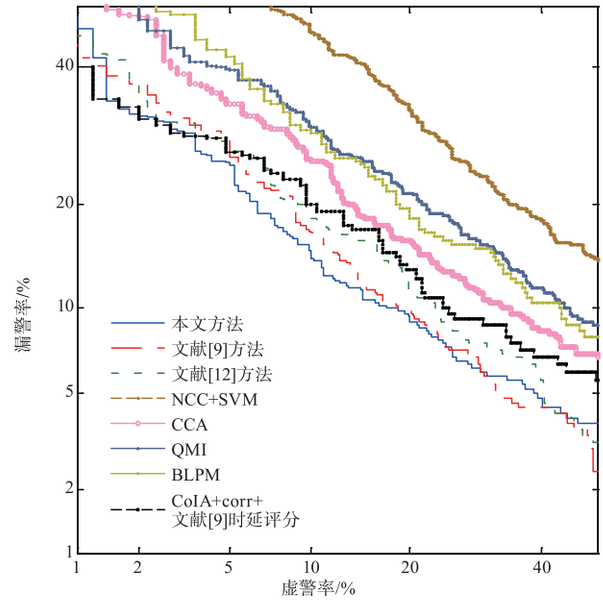


图7 不同算法总体DET曲线图

图 7 结合表 4 可知,对于各类不一致数据,本文方法性能优于其它算法;MSA 类和稀疏表示类方法总体上优于 MI 类和相关系数类算法. CoIA + corr + 文献[9]时延评分机制,CCA 和 BLPM 三者性能较为接近. 与其它方法相比,融合位置时延分布得分后对第三、第四类不一致数据的识别提升较为明显,尤其是第四类数据,这表明对于第四类不一致数据,时延是重要的区分因素,除文献[9]外其他文献并无考虑时延的影响. 文献[9]虽引入时延因素,但与本文方法不同的是其在整句话的评分机制中引入. 本文方法与文献[9]和文献[12]方法相比,总体 EER 较以上两者分别下降了 2.1% 和 4.6%. 本文方法采用韵母字典,对四类数据的性能比同采用多模态移不变字典的文献[12]方法分别有 2.4% ~7.6% 的提升,但原子数量只有文献[12]方法的 1/3 左右. 由于汉语是单音节结构,无声调音节有 408 个,若加上声调则有 1300 个以上,以音节为原子,对于大词汇

量数据集字典的规模会很大;同时,不同音节出现频率并不一致,部分音节可能在训练集中出现过少,而没在原子中得以表达;且训练集往往难以覆盖所有音节,导致字典只对训练集内包含的音节有较好的表达效果,当待分析语句包含集外音节时,一致性分析的结果必然会受到影响. 韵母数量不多且固定,多数数据库语料均覆盖所有韵母,且韵母在音节中也占有较大的时长,基于韵母切分的可行性,采用韵母字典可减少集外情况出现提高原子的代表性,从实验结果也得到了证明. 同时,韵母通常占整句话时长的一半左右,本文方法在字典训练前先进行韵母段筛选,去掉了静音和声母部分,实验中实际用于训练的数据只有 23000 帧左右,约占总帧数的 65%,一致性分析时也进行了类似处理. 虽引入韵母切分增加了运算成本,但本文方法在 CPU 型号为 I7 7700K,32G 内存的主机上运行,对 5s 时长语句分析的耗时约为 7.48s,其中音频韵母切分耗时为 0.137s,而其他方法的耗时介于 6.89s(CoIA + corr + 整句时延分析)与 12.46s(文献[12])之间. 同等条件下,与文献[12]采用整句话进行字典训练及分析相比,降低了整体的运算量. 其他比较算法也是整句进行分析,一致性分析时须逐帧提取唇部特征,而本文方法只须对韵母部分进行处理.

表 4　不同算法各类数据 EER 结果比较

不一致数据种类	EER(%)							
	本文($K=80$)	文献[9]	文献[12]($K=225$)	CoIA + corr + 文献[9]时延评分	CCA	BLPM	QMI	NCC + SVM
第一类	8.7	9.2	11.3	13.2	13.1	14.1	16.7	20.3
第二类	8.4	10.5	12.2	14.3	12.9	14.3	15.1	19.2
第三类	12.1	14.4	14.7	16.4	18.2	18.8	21.5	24.2
第四类	20.4	24.6	29.2	25.9	30.3	29.5	33.6	35.1
总体	13.3	15.4	16.9	17.2	17.5	19.9	21.7	27.9

5　结论

针对整句分析进行一致性判决所存在的不足,本文结合初始时延分布范围的统计分析结果,以分析代表性的韵母发音事件代替整个句子实现一致性判决,提出一种基于韵母发音事件匹配与位置时延分析的语音唇动一致性判决方法. 该方法利用双模态移不变稀疏表示模型描述音视频信号,通过提出的音视频结合韵母发音事件切分法对训练数据进行韵母段切分. 然后由学习所得韵母字典分析韵母事件的音唇匹配度,同时对各韵母出现位置的时延分布进行评分. 最后将韵母音唇匹配度得分与位置时延评分相融合进行一致性判决. 实验结果表明本文方法与现有主流方法相比有更好的识别效果,也在一定程度上降低了运算量.

参考文献

[1] Han Y,Song S,Zhao W. Retrieval of TV talk-show speakers by associating audio transcript to visual clusters [J]. IEEE Access,2017,5:20512 – 20523.

[2] Schonherr L,Zeiler S,et al. Spoofing detection via simultaneous verification of audio-visual synchronicity and transcription [A]. 2017 IEEE Automatic Speech Recognition and Understanding Workshop [C]. Okinawa:IEEE,2017. 591 – 598.

[3] Dov D,Talmon R,et al. Sequential audio-visual correspondence with alternating diffusion kernels [J]. IEEE Transactions on Signal Processing,2018,66(12):3100 – 3111.

[4] Liu Y,Sato Y. Recovery of audio-to-video synchronization through analysis of cross-modality correlation [J]. Pattern Recognition Letters,2010,31(8):696 – 701.

[5] Izadinia H,Saleemi I,et al. Multimodal analysis for identification and segmentation of moving-sounding objects [J]. IEEE Transactions on Multimedia,2013,15(2):378 – 390.

[6] EA Rúa,H Bredin,et al. Audio-visual speech asynchrony detection using co-inertia analysis and coupled hidden markov models [J]. Pattern Analysis and Applications,2009,12(3):271 – 284.

[7] Kumar K,Navratil J,et al. Audio-visual speech synchronization detection using a bimodal linear prediction model [A]. IEEE Computer Society Conference on Computer Vision & Pattern Recognition Workshops [C]. Florida:IEEE,2009. 53 – 59.

[8] S Kumagai,K Doman,et al. Detection of inconsistency between subject and speaker based on the co-occurrence of lip motion and voice towards speech scene extraction from news videos [A]. 2011 IEEE International Symposium on Multimedia [C]. California:IEEE,2011. 311 – 318.

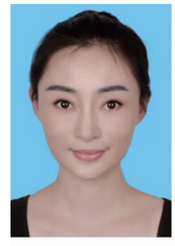
[9] 朱铮宇,贺前华,奉小慧,等. 基于时空相关度融合的语音唇动一致性检测算法 [J]. 电子学报,2014,42(4):779 – 785.
ZHU Zheng-yu,HE Qian-hua,FENG Xiao-hui,et al. Lip motion and voice consistency algorithm based on fusing spatiotemporal correlation degree [J]. Acta Electronica Sinica,2014,42(4):779 – 785. (in Chinese)

[10] Monaci G,Vandergheynst P,et al. Learning bimodal structure in audio-visual data [J]. IEEE Transactions on

Neural Networks,2009,20(12):1898 – 1910.
[11] Qingju Liu,Wenwu Wang,et al. Source separation of convolutive and noisy mixtures using audio-visual dictionary learning and probabilistic time-frequency masking [J]. IEEE Transactions on Signal Processing,2013,61(22):5520 – 5535.
[12] 贺前华,朱铮宇,奉小慧. 基于平移不变字典的语音唇动一致性判决方法 [J]. 华中科技大学学报(自然科学版),2015,43(10):69 – 74.
HE Qian-hua,ZHU Zheng-yu,FENG Xiao-hui. Lip motion and voice consistency analysis algorithm based on shift-invariant dictionary [J]. Journal of Huazhong University of Science and Technology (Nature Science Edition),2015,43(10):69 – 74. (in Chinese)
[13] El-Sallam A A,Mian A S. Correlation based speech-video synchronization [J]. Pattern Recognition Letters,2011,32(6):780 – 786.
[14] Eg R,Griwodz C,et al. Audiovisual robustness:exploring perceptual tolerance to asynchrony and quality distortion [J]. Multimedia Tools & Applications,2015,74(2):345 – 365.
[15] Staelens N,Meulenaere J D,et al. Assessing the importance of audio/video synchronization for simultaneous translation of video sequences [J]. Multimedia Systems,2012,18(6):445 – 457.
[16] 孙金城,倪宏,莫福源,等. 普通话声母和韵母的统计特性[J]. 应用声学,1995,14(3):35 – 41.
SUN Jin-cheng,NI Hong,MO Fu-yuan,et al. The statistical distribution of standard chinese initials and finals [J]. Journal of Applied Acoustics,1995,14(3):35 – 41. (in Chinese)
[17] Song T,Lee K,et al. Visual voice activity detection via chaos based lip motion measure robust under illumination changes [J]. IEEE Transactions on Consumer Electronics,2014,60(2):251 – 257.
[18] Siatras S,Nikolaidis N,et al. Visual lip activity detection and speaker detection using mouth region intensities [J]. IEEE Transactions on Circuits and Systems for Video Technology,2009,19(1):133 – 137.
[19] Wang Q,Shi G,et al. Analysis and design of an optimum detector for weak sinusoidal signals [A]. International Conference on Signal Processing [C]. Beijing:IEEE,2002. 1608 – 1611.
[20] Gustafsson F. Determining the initial states in forward-backward filtering[J]. IEEE Transactions of Signal Processing,1996,44(4):988 – 992.
[21] 钱博,李燕萍,唐振民,等. 基于频域能量分布分析的自适应元音帧提取算法 [J]. 电子学报,2007,35(2):279 – 282.
QIAN Bo,LI Yan-ping,TANG Zhen-min,et al. Self-adaptive vowel-frame detection algorithm based on energy distribution analysis in frequency domain [J]. Acta Electronica Sinica,2007,35(2):279 – 282. (in Chinese)
[22] 李皓,唐朝京. 采用损失函数和声学特征切分声韵母的方法 [J]. 声学学报,2012,37(3):339 – 345.
LI Hao,TANG Chao-jing. Initial/final segmentation using loss function and acoustic feature [J]. Acta Acustica,2012,37(3):339 – 345. (in Chinese)
[23] 胡瑛,陈宁. 基于小波变换的清浊音分类及基音周期检测算法 [J]. 电子与信息学报,2008,30(2):353 – 356.
HU Ying,CHEN Ning. Voiced/unvoiced classification and pitch period detection algorithm based on wavelet transform [J]. Journal of Electronics and Information Technology,2008,30(2):353 – 356. (in Chinese)
[24] Yang B,Liu R,Chen X. Fault diagnosis for wind turbine generator bearing via sparse representation and shift-invariant K-SVD [J]. IEEE Transactions on Industrial Informatics,2017,13(3):1321 – 1331.
[25] Ragnhild Eg,Dawn Behne,Carsten Griwodz. Audiovisual temporal integration in reverberant environments [J]. Speech Communication,2015,66:91 – 106.
[26] Takahashi T,Kageyama Y,Ariuntsengel B,et al. Analysis of lip motion due to the influence of vocalization [A]. SICE Annual Conference 2012 [C]. Akita:IEEE,2012. 973 – 978.
[27] 邵健,赵庆卫,颜永红. 基于鼻韵尾分离的汉语声韵母识别模型[J]. 声学学报,2010,35(5):587 – 592.
SHAO Jian,ZHAO Qing-wei,YAN Yong-hong. Initial/final acoustic model based on separating nasal coda in chinese putonghua speech recognition [J]. Acta Acustica,2010,35(5):587 – 592. (in Chinese)

作者简介

朱铮宇 男,1984 年出生,广东广州人,博士后,讲师. 主要从事音视频多模态信号处理方面的研究工作.
E-mail:zhuzhengyu0701@163.com



廖丽平 女,1981 年出生,福建厦门人,教授,硕士生导师,广东省系统工程学会理事. 主要从事软件定义、智能路由、大数据处理及应用等方面的研究工作.
E-mail:liping1110@hotmail.com