新的全参考音视频同步感知质量评价模型

魏耀都，谢湘，匡镜明，韩辛璐
（北京理工大学 信息与电子学院，北京 100081）

摘 要：提出一种利用协惯量分析构建的全参考音视频同步感知质量模型。通过对齐得到待测音频与视频的同步误差。将音视频内容分为纯净语音、无语音和有背景语音 3 类。将纯净语音类分为视频中有说话人和无说话人 2 个子类。分别对各类选取多维特征，利用协惯量分析从特征中获得音视频最相关的特征映射和相关程度。通过参考音视频得到相关程度曲线并得到同步误差到感知质量的映射关系。结果表明该模型评测结果与主观实验结果有较好相关性。

关键词：信息处理技术；音视频质量评价；协惯量分析；同步
中图分类号：TN912 **文献标识码**：A **文章编号**：1000-436X(2012)02-0182-08

Novel full reference perceptual quality metric for audio-visual asynchrony

WEI Yao-du, XIE Xiang, KUANG Jing-ming, HAN Xin-lu
(School of Information and Electronics, Beijing Institute of Technology, Beijing 100081, China)

Abstract: A full reference model was proposed to evaluate the perceptual quality of audiovisual asynchrony. A standard synchronization process was used to determine the time difference between audio and video. The mapping between the time difference and the perceptual quality was derived by co-inertia analysis. The co-inertia analysis extracted the most related component from audio and video features, and then formed a mapping for each audiovisual sequence. Audiovisual contents were divided into three categories: clean speech, non speech and mixed speech. The clean speech category was further split into two subcategories. Audio and video features were chosen separately for each category. Subjective test results showed that the proposed model conforms well with subjective results.

Key words: signal processing technique; audiovisual quality assessment; co-inertia analysis; synchrony

1 引言

进入 21 世纪，多媒体通信的蓬勃发展使沟通和交流变得更加轻松方便，然而多媒体业务的质量却参差不齐，在信道不稳定时常常无法提供让人满意的质量。对多媒体质量进行准确的评价能够规范多媒体服务水平，促进行业健康发展。现有的多媒体质量评价方法集中在单独对视频或者音频的评价上，然而在大多数多媒体业务中用户都会同时使用音频和视频，因此对音视频质量进行综合评价能够更准确地描述用户感知体验。

目前，国际上对音视频质量综合评价已经有一些研究，Andrew Rimell 等分析了音视频质量间的相互影响[1]，Hands 提出了一种基本的音视频质量评价模型[2]，该模型在音视频同步的假设下提出，当音视频不同步时没有给出评价方法。对音视频同

收稿日期：2010-12-31；**修回日期**：2011-10-15
基金项目：国家科技重大专项基金项目资助（2010ZX03004-003）
Foundation Item: The Important National Science Technology Specific Project (2010ZX03004-003)

步感知质量的研究可以完善现有音视频质量评价模型。

在对音视频同步的研究方面，Steinmetz 定义了同步、失步和暂态区间，用于描述感知质量在不同同步误差范围内的性质[3]。随后针对几种音视频研究了同步区间的宽度，结果表明同步区间的宽度受到音视频内容的影响。但 Steinmetz 没有指出如何在内容和感知质量之间建立联系。Nishibori 等利用格式塔心理学中的同时性和同向性作为判断音频与视频事件是否出自同一事件的准则[4]。Bredin 等总结了音频和视频相关性的衡量方法[5]。在音视频相关性的基础上，Gillet 和 Liu 等分别提出了自动恢复音视频同步的方法[6,7]。Enrique 等利用隐马尔科夫模型和协惯量分析(CoIA, co-inertia analysis)对失步的音视频进行自动同步[8]。Eveno 等通过对同步的检测设计了一种活性评分机制，用以检测语音是否由视频中的人物实时说出[9]。Kumar 等研究了脸部正面图像与语音的同步检测方法[10]。虽然目前已经有多种自动恢复同步的方法，但实际业务中仍然常会出现音视频同步误差，所以仍然需要在质量评价模型中加入同步质量评价指标。然而目前对音视频同步的研究并没有对同步误差与感知质量之间的关系进行分析。

本文针对 QVGA 分辨率的视频进行研究。首先将待测序列的音视频分别与参考序列进行对齐，利用对齐结果得到待测序列的同步误差。由于感知质量受到音视频内容的影响，所以根据音频内容将音视频分为纯净语音、无语音和有背景语音 3 类，纯净语音类进一步划分为视频中出现说话人和不出现说话人 2 个子类。对各类分别提取不同的特征。利用 CoIA 寻找使音频和视频特征协方差最大的映射，并将该映射结果的协方差系数作为相关程度参数。将参考序列的音频进行小范围的移位，每次移位后均进行 CoIA 计算，从而得到相关程度参数曲线。利用该曲线估计同步误差与感知质量之间的映射关系，从而在主观质量与同步误差之间建立质量评价模型。模型中各参数由主观实验结果确定。为验证模型有效性，选择不同类型的序列进行了验证实验，验证结果表明本模型与主观质量有较好相关性。

2　评价模型

人类对音视频是否同步的判断主要依靠视频事件和对应的音频事件是否同时发生，因此本模型通过计算音频和视频特征在时间上的关联对感知质量进行估计。评价模型包括 2 部分，第 1 部分获得同步误差，第 2 部分通过对音视频内容的分析获得同步误差与感知质量之间的映射关系，从而通过同步误差对感知质量进行估计。

计算音视频的同步误差需要使用参考序列。假设参考序列的音视频完全同步，根据 ITU-T P.931 标准建议的方法将待测序列的视音频分别与参考序列进行对齐，从而得到视频延迟的帧数 f_{video} 和音频延迟的帧数 f_{audio}。若相邻视频帧间隔时间为 t_{video}，音频帧长度为 t_{audio}，则待测序列的音视频同步误差 t_{skew} 为

$$t_{\text{skew}} = f_{\text{video}} \times t_{\text{video}} - f_{\text{audio}} \times t_{\text{audio}} \tag{1}$$

其中，t_{skew} 为负值时表示音频的播放领先于视频，为正值时表示视频的播放领先于音频。

感知质量通常采用平均意见分(MOS，mean opinion score)进行定量描述。MOS 的评分范围一般为 5 等级。然而主观评价中测试人对评分表两端的使用较为慎重，导致实际评分的可区分度不高。因此本评价模型采用 9 等级评分，在获得评价数据之后对 MOS 分值进行去除隐含参考条件操作(HRR, hidden reference removal)，得到 ACR-HRR 分值。ACR-HRR 分值能够提供与使用失真等级评定（DCR，differential category rating）方法进行实验相同的信息，同时使测试时间仅为 DCR 实验的一半[11]。

音视频的特征通常为多维异构特征，对多维异构特征的关联性计算方法有典型相关方法(CANCOR, canonical correlation)和 CoIA 方法。CANCOR 方法可以从多维特征中找到相关系数最大的映射，CoIA 可以找到协方差最大的映射，二者都可以分析音频与视频的关联。CoIA 由 Doledec 和 Chessel 在关于物种与环境关系的研究中提出，但直到近年才被引入到多媒体分析中[12]。CoIA 能够对 2 个具有不同维数的多元随机变量 X、Y 寻找到矩阵 $\boldsymbol{A}$ 和 $\boldsymbol{B}$，使得 X 和 Y 分别在 $\boldsymbol{A}$ 和 $\boldsymbol{B}$ 上的投影具有最大的协方差。Enrique 等人给出了 CoIA 的详细计算过程，随后比较了 CANCOR 与 CoIA 2 种方法在音视频同步分析中的效果，结果表明 CoIA 更适于分析音视频之间的关联性[8]。另一方面，CANCOR 的计算中需要多次对音视频特征的协方差矩阵求逆，在协方差矩阵不可逆的情况下无法求

得结果。而 CoIA 则不需要求逆矩阵，适用性更好。因此本模型采用 CoIA 进行音视频关联性的计算。

由于音视频内容对同步感知质量有明显的影响，所以需要对内容进行分类，根据各类的性质分别构建评价模型[3]，分类方法如图 1 所示。视频的内容非常灵活，分类方法和所分类别众多；对视频内容进行自动识别和归类所需要的计算资源较多，同时可识别的种类很少，因此很难根据视频内容对音视频进行有效的分类。然而音频信号的分析、识别与归类则较为容易。同时，语音在人类感知中有着特殊的作用，人类对语音和唇型的同步感知比其他内容更为敏感。因此根据音频内容将音视频分为 3 类：纯净语音类、无语音类和有背景语音类，其中纯净语音类根据视频内容中是否出现与语音对应的嘴进一步分为有说话人和无说话人 2 个子类。

图 1 音视频内容分类

2.1 无说话人子类

由于语音只能由嘴的运动发出，嘴部视频与其所发出的语音具有很强的关联性，所以唇型与语音的不同步很容易被察觉。而当视频中没有出现与语音对应的唇型时，语音与视频内容的关联性则较弱，用户对同步误差也较为不敏感。因此纯净语音类根据视频中是否出现与语音相对应的嘴分为 2 个子类：有说话人和无说话人。由于视频镜头切换可能会造成视频中人物的变化，所以对纯净语音类的评价以一个视频镜头为单位进行。

造成视频中没有说话人的原因有 2 种：①视频中的人物在听镜头外其他人说的话，例如正在听记者问题的被采访对象；②视频中不存在人物或者没有清晰可辨的嘴，例如视频为风光或者体育节目中的远景镜头。在这 2 种情况下语音与视频内容都不存在严格的时间关联，音视频同步误差 t_{skew} 的增大对主观质量分值(*ACR_HRR*)的影响较小。所以无说话人子类的评价模型采用一条较为平坦的高斯曲线来进行描述。

$$ACR\_HRR = 7 - 7 \times \mathrm{e}^{-\left(\frac{t_{\text{skew}}}{\sigma}\right)^2} \tag{2}$$

其中，σ 为高斯曲线的标准差。

2.2 有说话人子类

嘴部特征可以分为形状特征和整体特征，形状特征包括嘴的高度、宽度、面积和轮廓等，整体特征包括 DCT 系数等。Bredin 等比较了 2 种特征与音频的相关性，结果表明采用整体特征优于形状特征[5]。因此对于有说话人子类，首先进行嘴部检测找到视频中嘴部的位置。对每个镜头分别用一个包含嘴部的固定尺寸矩形作为嘴部范围，对嘴部矩形范围内的亮度图形进行二维 8×8 DCT 变换后按照 Z 字型扫描顺序选取前 30 个 DCT 系数，这 30 个系数和它们的一、二阶差分一起作为视频特征。

音频按照 20ms 的长度进行分帧，帧间叠接 10ms。对每个音频帧计算短时能量和梅尔倒谱系数(MFCC, mel-frequency cepstral coefficients)，采用短时能量和 MFCC 前 12 个系数及其一、二阶倒数作为音频特征。由于音频帧和视频帧的长度不同，所以对每个视频帧长度范围内的音频特征求平均，使得音频特征与视频特征的长度相同。

将缩短后的音频特征在±20 视频帧范围内进行逐帧移位。每次移位后均对视频特征与移位后的音频特征进行协惯量分析。协惯量分析的 2 个第一维特征为使音频与视频特征协方差最大的映射结果，将 2 个第一维特征的协方差作为音视频的相关程度参数。通过音频的移位可以得到相关程度参数曲线，对该曲线进行归一化。如果音频与视频内容存在关联，相关程度曲线会在靠近中心的位置出现凸起，在两侧逐渐下降；而当音视频内容不存在关联时则不会出现明显的凸起[9]。因此本模型通过相关程度曲线的凸起程度估计主观质量随 t_{skew} 的上升程度。为描述相关程度曲线，采用高斯曲线对其进行拟合，拟合公式为

$$y = \mathrm{e}^{-\left(\frac{t_{\text{skew}} - \beta}{\sigma_{\text{obj}}}\right)^2} \tag{3}$$

其中，β为相关程度曲线峰值位置，σ_{obj} 为标准差，描述了相关程度曲线的凸起程度。随后再次利用高斯曲线对主观质量与 t_{skew} 之间的映射关系进行建模，高斯曲线为

$$ACR\_HRR = 7 - 7\mathrm{e}^{-\left(\frac{t_{\text{skew}}}{\sigma_{\text{sub}}}\right)^2} \tag{4}$$

其中，σ_{sub} 由σ_{obj} 通过线性或者非线性映射得到。最后通过待测序列的音视频同步误差 t_{skew} 得到该序列的感知质量 ACR-HRR 评分。模型中各参数通过主

观实验结果确定。

2.3　无语音类

无语音类的评价以一个或数个无语音类镜头为单位进行。评价模型的构造方法与有说话人子类的方法在音视频特征选取上有明显的区别，其余部分完全相同。

在无语音类音视频当中，用户通常只在出现镜头切换、冲击音、节奏变换等音频或者视频内容有显著变化的时刻才会感受到音视频的不同步。所以在视频特征的选取中需要选取能够反映视频运动状态变化的特征。运动矢量的统计量可以较好地反映视频内容的状态[13]。光流法可以提供与运动矢量类似的结果，同时能够反映视频中光线的变化，利用光流矢量的统计结果可以更全面地描述视频中的运动状态。因此本模型对于待测序列首先进行镜头切换检测，得到镜头切换时刻。随后通过 Horn & Schunck 方法得到每帧视频的光流矢量图。最后统计光流幅度的方差、光流幅度非零块的个数、非零光流的幅度方差、光流幅度最大 2 个块间的距离。将光流的统计量与镜头切换点一起作为视频特征。

在音频中引入 Gillet 等提出的段落相似程度指标[6]。该指标比较当前时刻的前后 2 段音频，将 2 段音频的各种特征映射到再生核希尔伯特空间中，在该空间中各特征均可被认为服从正态分布。随后计算 2 段音频特征分布的 Kullback-Leibler 距离作为 2 段音频的段落相似程度。该指标在音频中旋律和段落变化的时刻会出现峰值，从而反映音频性质的改变情况。得到段落相似程度指标后进行能量检测，获得冲击音发生的时刻，同时进行基音检测以提取旋律特征。最后计算音频响度。将段落指标、冲击音发生时刻、基音以及响度作为音频特征。

2.4　有背景语音类

在有背景语音类中同时存在着纯净语音和非语音的背景，因此音频与视频的相关性既可以存在于语音与图像之间又可以存在于非语音的音频和图像之间。即有背景语音类可以看作纯净语音类与非语音类的叠加，2 种类别的音频与视频之间的相关都可以使有背景语音类的音频和视频之间产生相关性。因此，有背景语音类可以利用纯净语音类与无语音类的评价模型进行评价。

首先对待测序列的视频内容进行判断，根据判断结果分别使用不同的评价模型。如果待测序列中没有出现清晰的嘴，则采用无语音类的方法进行评价；如果待测序列存在清晰的嘴，但嘴没有说话，则使用纯净语音类中无说话人子类的方法；如果序列中存在说话的嘴则使用纯净语音类中有说话人子类的方法。

各类评价模型的流程如图 2 所示。

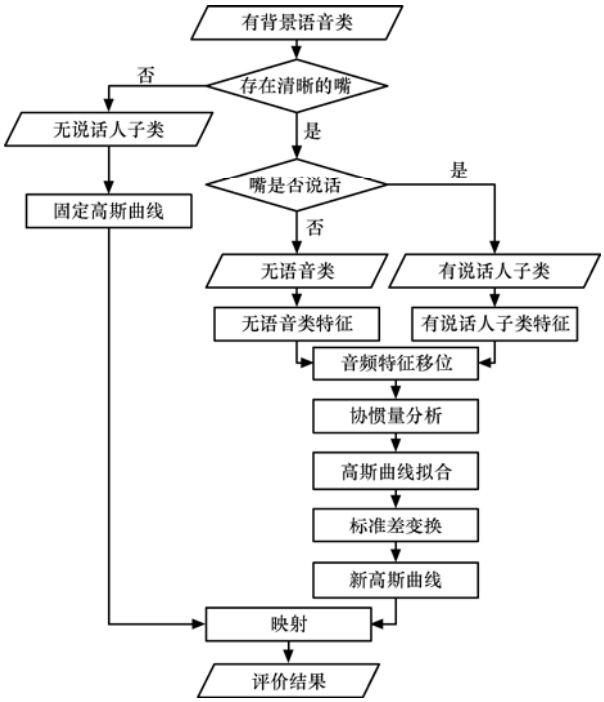


图 2　评价模型流程

3　确定模型参数的主观实验

3.1　实验设计

主观实验的测试环境符合 ITU-T P.911 标准的规定。实验采用 4 台三星 T220P（1920×1200）液晶显示器进行视频播放。音频播放使用 4 个 Sennhesier HD25 耳机进行。视频序列采用 QVGA 分辨率在显示器中央进行显示，显示器其余部分显示中灰色作为背景。每组测试由 4 名测试人同时进行，对各组分别使用不同的随机播放顺序。根据 ITU-T P.911 的建议，测试人可以在视频高度的 1~8 倍距离内自行调节观看距离。对纯净语音类的测试使用 32 名测试人，对无语音类的测试使用 20 名测试人，男女测试人员各占测试人数的一半。测试人年龄在 22~29 岁之间，全部具有正常听力、视力或矫正视力且均不是音频或视频方面的专家。由于实验规模较大，测试分为 3 阶段进行，阶段之间均间隔两周以上。

实验采用 ITU-T P.911 建议的 9 等级绝对等级评分(ACR, absolute category rating)获得平均意见分，评分准则如图 3 所示。在获得评价数据之后对 MOS 分值进行去除隐含参考条件操作得到 ACR-HRR 分值。

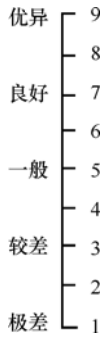


图 3　ACR 评分准则

测试序列均由高质量源视频通过双 3 次插值转换为 320×240 的无压缩 avi 格式视频。音频转换为 48kHz 采样，16 比特量化的 PCM 单声道音频。实验采用 10 条有说话人序列、3 条无说话人序列以及 5 条无语音序列。测试序列名称及内容在表 1 中详细列出。有背景语音类由于可以利用纯净语音类与无语音类的评价模型，故在本阶段实验中没有进行测试，只在验证实验中进行测试。

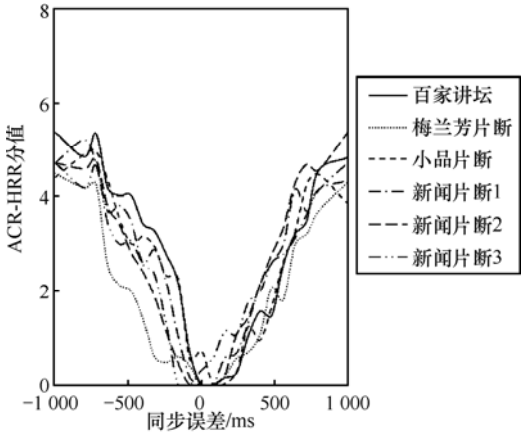
将测试序列的音频进行移位，对移位后的序列进行主观质量的评分，音频的移位时间量分别为±1 000ms、±800ms、±720ms、±640ms、±560ms、±480ms、±400ms、±320ms、±240ms、±160ms、±80ms 和 0ms。

3.2　实验结果与评价模型参数的确定

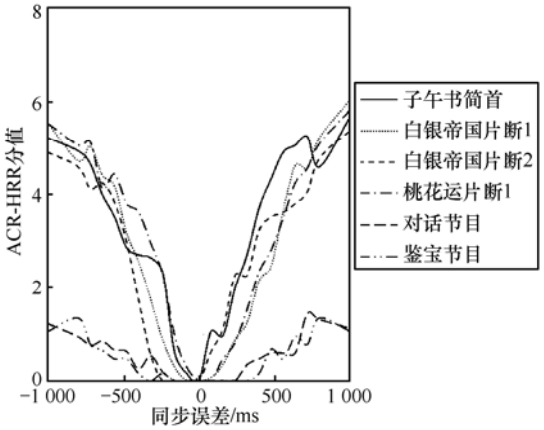
各序列的 ACR-HRR 实验结果如图 4 所示。从图 4 可以看出，当同步误差增加时，所有序列的主观质量均下降，下降趋势与高斯曲线基本吻合，但下降的速度各不相同，表明内容对评价结果有明显的影响。所以主观质量曲线可以利用具有不同标准差的高斯函数进行拟合，拟合方法采用非线性最小均方误差法。拟合结果在表 2 中列出，拟合为式（4）。纯净语音类拟合的 R^2>0.8，无语音类 R^2>0.7。同时可以看出“对话节目”、“鉴宝节目”与“足球 2”3 条序列曲线都非常平坦，如果将 3 条曲线通过平移使得中心一致的话，3 条曲线的变化趋势非常接近，可以用一条固定的高斯曲线对它们进行统一描述。

表 1　　**测试序列**

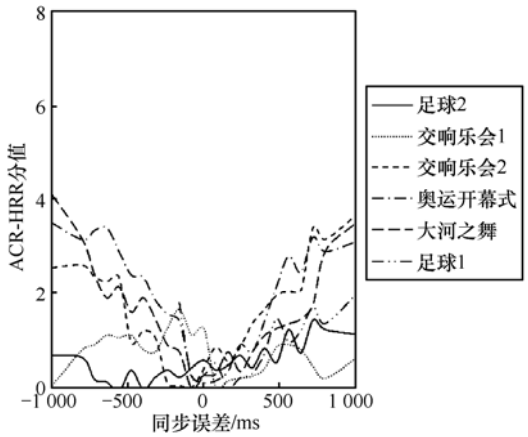
序列名称	视频内容	音频内容	序列类别	序列长度/s
百家讲坛	嘉宾演讲，半身镜头	嘉宾的语音	有说话人	8
梅兰芳片断	主角对大家讲话，广角镜头	主角的语音	有说话人	7
小品片断	表演者说台词，广角镜头	表演者的语音	有说话人	8
新闻片断 1	2 名主播中 1 名致开场白，广角镜头	主播的语音	有说话人	8
新闻片断 2	受访者接受采访，面部特写	受访者的语音	有说话人	8
新闻片断 3	1 名主播进行播报，半身镜头	主播的语音	有说话人	8
子午书简首片断	主持人致开场白，半身镜头	主持人语音	有说话人	6
白银帝国片断 1	主角念台词，面部特写镜头	主角语音	有说话人	8
白银帝国片断 2	主角念台词，半身镜头	主角语音	有说话人	8
桃花运片断 1	主角念台词，半身镜头	主角语音	有说话人	8
交响乐会 1	音乐大厅与观众，广角镜头	乐队演奏的乐曲	无语音	8
交响乐会 2	从小提琴演奏者半身镜头切换到鼓手半身镜头	小提琴乐曲切换到鼓声	无语音	8
奥运开幕式	击缶表演，广角镜头	敲击声	无语音	8
大河之舞	从踢踏舞广角镜头切换到观众鼓掌	踢踏舞声切换到鼓掌声	无语音	8
足球 1	球员冲撞犯规，广角镜头	助威声和裁判鸣哨声	无语音	7
足球 2	正常比赛，广角镜头	解说员语音	无说话人	8
对话节目	听观众发言的嘉宾	观众语音	无说话人	8
鉴宝节目	1 件玉器	解说员对玉器进行描述的语音	无说话人	8



(a) 第 1~6 条序列



(b) 第 7~12 条序列



(c) 第 8~13 条序列

图 4　测试结果

对有说话人子类和无语音类，分别对各测试序列进行移位和协惯量分析，得到相关程度曲线，百家讲坛、交响乐会 1 和大河之舞 3 条序列的相关程度曲线如图 5 所示。由于测试使用的参考序列本身的音视频并不是完全准确同步，所以相关程度曲线的凸起中心有一定偏移。随后用式（3）对相关程度曲线中的凸起部分进行拟合。纯净语音类与无语音类的σ_{obj}与σ_{sub}的映射关系如图 6 所示，对 2 类分别进行映射关系的曲线拟合，纯净语音类的拟合式为

$$\sigma_{sub}=8.4\times\sigma_{obj}+618.5 \tag{5}$$

无语音类的拟合公式为

$$\sigma_{sub}=e^{\frac{\sigma_{obj}+25.85}{6.912}} \tag{6}$$

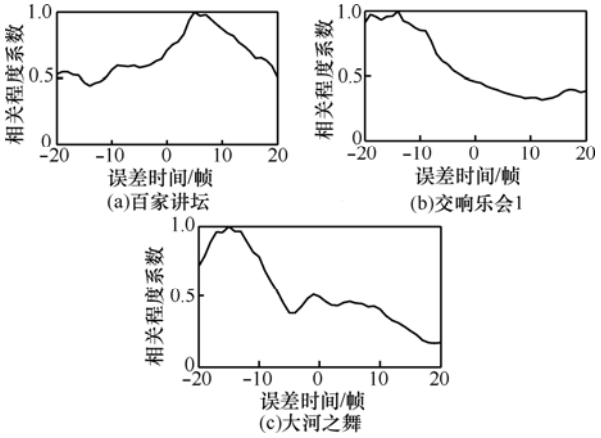
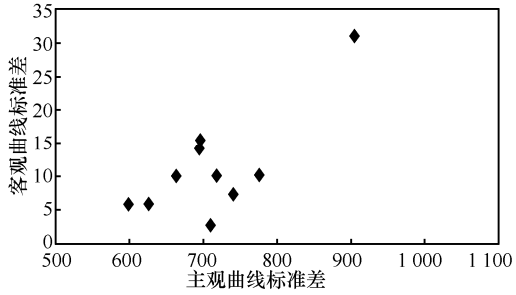
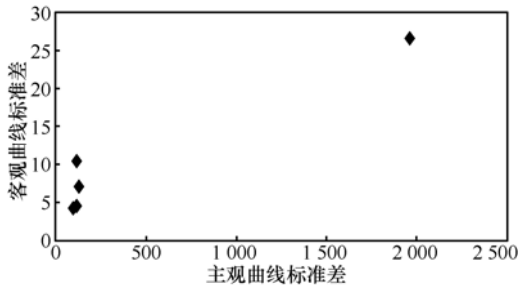


图 5　相关程度曲线



(a) 纯净语音类



(b) 无语音类

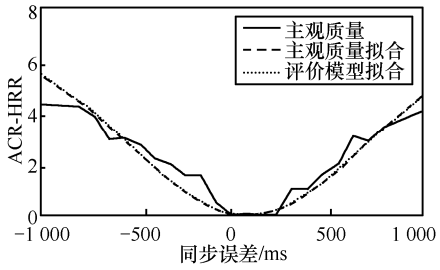
图 6　σ_{obj}与σ_{sub}的映射关系

无说话人子类各序列的主观质量曲线非常接近，因此采用同样的高斯曲线对其进行回归分析，并将拟合得到的曲线直接作为同步误差与主观质量间的映射曲线。拟合结果为

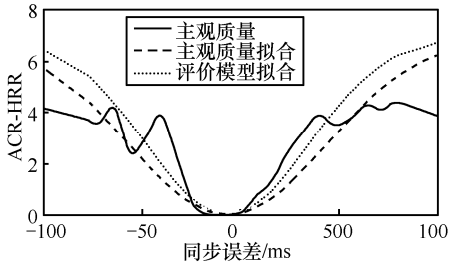
$$ACR\_HRR=7-7\times e^{-\left(\frac{f_{skew}}{2047}\right)^2} \tag{7}$$

表 2　　验证实验测试序列

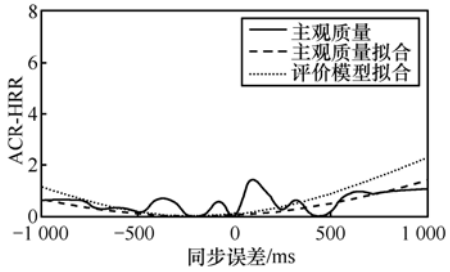
序列名称	视频内容	音频内容	序列类别	序列长度/s
电视剧片断	主角在饭桌上说话，广角镜头	主角的语音	有说话人	8
桃花运片断 2	主角在念报纸，半身镜头	主角的语音	有说话人	8
今日说法	主持人播报案情，半身镜头	主持人的语音	有说话人	8
英雄片断	镜头在前行的马匹和草原风景间来回切换	随镜头内容在马蹄声和寂静间切换	无语音	8
神话片断	2 人打斗场面，广角镜头	刀剑碰撞声	无语音	8
打鼓片断	主角有节奏的打鼓，广角镜头	敲击声	无语音	8
演唱会片断 1	男歌手歌唱，面部特写镜头	歌手歌声和伴奏声	有背景语音	8
演唱会片断 2	女歌手歌唱，面部特写镜头	歌手歌声和伴奏声	有背景语音	8
记录片	2 辆卡车先后从摄像机前开过，广角镜头	解说员解说语音和卡车由远到近的声音	有背景语音	8
阅兵式	领导乘车通过列队，半身镜头	解说员语音和阅兵式背景音乐	有背景语音	8
武打片断	2 位师傅示范拳术，广角镜头	画外音讲解和挥拳声	有背景语音	8



(a) 电视剧片断



(b) 打鼓片断



(c) 纪录片

图 7　验证实验结果

回归分析的 R^2>0.9。

因此，对于待测的音视频序列，可以首先计算其相关程度曲线，然后通过相关程度曲线得到σ_{obj}对σ_{sub}进行估计，从而得到主观质量与同步误差时间之间的映射关系。

4　模型性能验证

为了验证所提出模型的性能，另外选择了 5 条有背景语音序列、3 条无语音序列以及 3 条纯净语音序列使用同样的测试人员进行了验证实验。验证实验的实验设计除测试序列外与上一次实验相同。验证实验的测试序列内容如表 2 所示。

图 7 显示了部分序列的高斯曲线拟合结果，可以看出高斯曲线可以较好地描述主观测试结果。分别利用有说话人子类、无说话人类和无说

话人子类对验证实验各序列的主观质量曲线进行客观估计，估计结果如图 7 所示。在有背景语音类型序列中，演唱会片断 1 和演唱会片断 2 按照有说话人子类方法处理，纪录片和武打片断按照无语音类方法处理，阅兵式按照无说话人子类方法处理。由图 7 可以看出估计结果与主观曲线本身的拟合结果很接近，也能较好地描述主观质量曲线。

为了比较评价模型的结果与主观实验结果，分别计算主观质量拟和曲线与评价模型拟和曲线对主观质量曲线之间的均方根误差。各序列的均方根误差比较结果如表 3 所示。

表 3　　均方根误差比较

序列名称	主观拟和曲线 RMSE	客观拟和曲线 RMSE
电视剧片断	0.693	0.700
桃花运片断 2	0.707	0.714
今日说法	0.877	0.877
英雄片断	0.781	0.787
神话片断	0.742	0.742
打鼓片断	1.044	1.114
演唱会片断 1	0.794	1.034
演唱会片断 2	0.866	0.975
纪录片	0.640	0.735
阅兵式	0.671	0.671
武打片断	0.775	0.975

从表 3 可以看出，用高斯曲线对主观评价结果进行拟和可以获得较小的均方根误差，说明高斯曲线可以较好地拟和主观质量随同步误差时间增加而下降的趋势。客观拟和曲线与主观实验结果的均方根误差与主观拟和曲线非常接近，表明所提出的模型可以较好地对各种类型的音视频序列进行同步质量估计。

5　结束语

对音视频质量进行评价需要对各种导致质量下降的因素进行定量分析。音视频不同步作为当前音视频业务常见的失真方式之一，将导致用户的感知体验受到损伤，从而降低音视频质量。已有的研究集中在如何对音视频质量进行融合上，对同步带来的损伤没有进行定量的分析与描述。

本文提出一种利用协惯量分析进行的全参考音视频同步质量感知评价模型，模型通过对齐算法获得音视频同步的时间误差，随后针对不同类型音视频内容，通过协惯量分析获得同步误差与主观质量之间的映射关系从而得到质量评分。实验结果表明，采用 9 等级评分制时，11 条测试序列中的 9 条序列由本模型获得的评价结果与主观实验结果的均方根误差小于 1，其余 2 条序列的均方根误差略大于 1，说明本模型的评价结果与主观结果的偏离程度较小，可以较好地描述主观感知质量。

由于采用高斯曲线对同步误差与主观质量的映射关系进行建模，本模型对音频和视频存在周期性的音视频内容尚无法给出较准确的质量估计结果。另外，如果音视频分类出现错误，会在质量估计结果中引入很大的误差，因此对于音视频的自动分类方法还需要进行进一步的研究。为实现本模型，还需要引入有效的嘴部区域检测算法以支持纯净语音类和有背景语音类中的特征提取方法。

参考文献：

[1] RIMELL A, OWEN A. The effect of focused attention on audio-visual quality perception with applications in multi-model codec design[A]. ICASSP 2000[C]. Istanbul, Turkey, 2000. 2377-2380.

[2] HANDS D S. A basic multimedia quality model[J]. IEEE Transactions on Multimedia,2004 ,12 (6): 806-816.

[3] STEINMETZ R. Human perception of jitter and media synchronization[J]. IEEE Journal on Selected Areas in Communications, 1996, 14(1): 61-72.

[4] NISHIBORI K, TAKEUCHI Y, MATSUMOTO T, *et al*. Finding the correspondence of audiovisual events by object manipulation[J]. Electronics and Communications, 2009, 92(5): 1-13.

[5] BREDIN H, CHOLLET G. Audiovisual speech synchrony measure: application to biometrics[J]. Eurasip Journal on Advances in Signal Processing, 2007, (3): 1-11.

[6] GILLET O, ESSID S, RICHARD G. On the correlation of automatic audio and visual segmentations of music videos[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2007,3(17): 347-355.

[7] LIU Y Y, SATO Y. Recovery of audio-to-video synchronization

through analysis of cross-modality correlation[J]. Pattern Recognition Letters, 2010 ,31 (8): 696-701.

[8] ENRIQUE A R, BREDIN H, GARCIA M C, *et al*. Audio-visual speech asynchrony detection using co-inertia analysis and coupled hidden markov models[J]. Pattern Analysis & Applications, 2009,9(12):271-284.

[9] EVENO N, BESACIER L. Co-inertia analysis for “liveness” test in audio-visual biometrics[A]. Proceedings of the 4th International Symposium on Image and Signal Processing and Analysis[C]. Zagreb, Croatia, 2005. 257-261.

[10] KUMAR K, NAVEATIL J, MARCHERET E, *et al*. Audio-visual speech synchronization detection using a bimodal linear prediction model[A]. 2009 IEEE Conference on Computer Vision and Pattern Recognition[C]. 2009. 53-59.

[11] QUAN H T, GHANBARI M. A comparison of subjective video quality assessment methods for low-bit rate and low-resolution video[A]. The 7th IASTED International Conference on Signal & Image Processing[C]. 2005.70-76.

[12] DOLEDEC S, CHESSEL D. Co-inertia analysis: an alternative method for studying pecies-environment relationships[J]. FreshwaterBiology, 1994,31: 277-294.

[13] JEANNIN S, DIVAKARAN A. MPEG-7 visual motion descriptors[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2001, 6(11):720-724.

作者简介：

魏耀都（1984-），男，四川成都人，北京理工大学博士生，主要研究方向为音视频质量评价方法与模型。

谢湘（1976-），男，湖南衡阳人，博士，北京理工大学副教授，主要研究方向为语音及音频信号处理、人机交互技术和移动通信。

匡镜明（1943-），男，湖南益阳人，博士，北京理工大学教授，主要研究方向为语音及音频信号处理、人机交互技术和移动通信。

韩辛璐（1987-），女，陕西兴平人，北京理工大学硕士生，主要研究方向为音频质量评测。