(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 113707136 B
(45)授权公告日 2021.12.31

(21)申请号 202111258776.X
(22)申请日 2021.10.28
(65)同一申请的已公布的文献号
申请公布号 CN 113707136 A
(43)申请公布日 2021.11.26
(73)专利权人 南京南大电子智慧型服务机器人研究院有限公司
地址 210019 江苏省南京市建邺区白龙江东街8号科技综合A区1幢14层
专利权人 南京大学
江苏南大电子信息技术股份有限公司
(72)发明人 雷桐 卢晶 刘晓峻 狄敏 吴宝佳
(74)专利代理机构 南京瑞弘专利商标事务所(普通合伙) 32249
代理人 陈建和
(51)Int.Cl.
G10L 15/06(2013.01)
G10L 15/14(2006.01)
G10L 15/20(2006.01)
G10L 15/25(2013.01)
G10L 25/84(2013.01)
(56)对比文件
CN 112735460 A,2021.04.30
CN 111599371 A,2020.08.28
CN 110931036 A,2020.03.27
CN 113030862 A,2021.06.25
审查员 庞秋婵

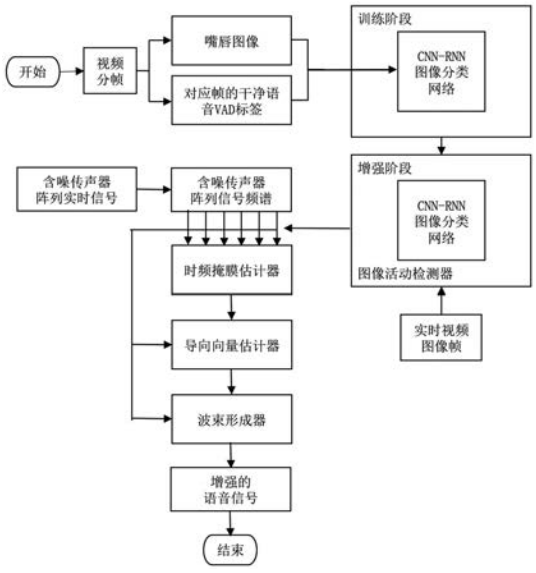
权利要求书4页 说明书9页 附图3页

(54)发明名称

服务型机器人语音交互的音视频混合语音前端处理方法

(57)摘要

本发明公开了一种服务型机器人语音交互的音视频混合语音前端处理方法，具体步骤如下：(1)通过视频处理手段捕获期望说话人嘴部动作信息；(2)根据期望说话人嘴部动作信息获得准确的语音激活检测结果；(3)根据语音活动检测结果，优化机器人传声器阵列的波束算法；(4)通过阵列传声器实现语音增强，抑制环境噪声，提升机器人采集语音的信噪比。本发明在机器人所处复杂声场环境中可以有效提升机器人采集语音的信号质量。



CN 113707136 B

1.一种服务型机器人语音交互的音视频混合语音前端处理方法,其特征在于,包括以下步骤:

步骤1,模型训练:采集训练音视频样本,将训练音视频样本中视频部分按帧分成图像,将训练音视频样本中语音部分按对应帧图像进行标签,得到对应帧的干净语音VAD标签;将图像和对应帧的干净语音VAD标签导入CNN-RNN图像分类网络中,对图像中的唇动状态和应帧的干净语音VAD标签进行训练,得到训练好的CNN-RNN图像分类网络;

步骤2,采集目标说话人嘴部动作视频和对应的含噪语音;嘴部动作视频用卷积神经网络方法标记出目标说话人人脸五官定位,并裁出嘴唇区域图像;嘴唇区域图像逐帧进行灰度图重塑得到嘴唇区域灰度图像,将嘴唇区域灰度图像输入到图像活动语音检测器;

步骤3,图像活动语音检测器根据输入的嘴唇区域灰度图像检测到目标说话人正在说话,则将嘴唇区域灰度图像输入到训练好的CNN-RNN图像分类网络中,得到此帧嘴唇区域灰度图对应的图像语音VAD概率;

步骤4,对含噪语音做短时傅里叶变换得到短时傅里叶频谱;

对含噪语音做短时傅里叶变换得到短时傅里叶频谱的方法:

$k \in \{1, \ldots, K\}$是源索引,K表示源信号个数,$m \in \{1, \ldots, M\}$ 是传声器索引,M表示传声器个数;在时域中,第 m 个传声器的语音信号 $y_m(t)$写为:

$$y_m(t)=\sum_k\sum_\tau h_m^{(k)}(\tau)s^{(k)}(t-\tau)+n_m(t)$$

其中,$y_m(t)$表示第 m 个传声器的语音信号,$s^{(k)}(t)$表示第 k 个源信号的噪声信号,$n_m(t)$表示第 m 个传声器采集到的噪声信号,$h_m^{(k)}(\tau)$表示对应于第k 个源和第 m 个传声器之间的脉冲响应,τ是图像的时间帧索引,t表示时刻;

第 m 个传声器的语音信号 $y_m(t)$通过应用短时傅立叶变换在频域中表示为:

$$y_m(f,t)=\sum_k h_m^{(k)}(f)s^{(k)}(f,t)+n_m(f,t)$$

其中,$y_m(f,t)$为$y_m(t)$的频域表示,$h_m^{(k)}(f)$为$h_m^{(k)}(\tau)$的频域表示,$s^{(k)}(f,t)$为$s^{(k)}(t)$的频域表示,$n_m(f,t)$为$n_m(t)$的频域表示;

脉冲响应的长度远小于 STFT 窗口的长度,因此,脉冲响应和源信号在时域中的卷积表示为时不变频率响应和时变源信号在频域中的乘积,引入矢量符号,将应用短时傅立叶变换在频域中表示改写为:

$$\mathrm{y}(f,t)=\sum_k \mathrm{r}^{(k)}(f)s^{(k)}(f,t)+\mathrm{n}(f,t)$$

其中:

$$\mathrm{y}(f,t)=\left[y_1(f,t),y_2(f,t),\ldots,y_M(f,t)\right]^T$$

$$\mathrm{r}^{(k)}(f)=\left[h_1^{(k)}(f),h_2^{(k)}(f),\ldots,h_M^{(k)}(f)\right]^T$$

$$\mathbf{n}(f,t)=\left[n_1(f,t),n_2(f,t),\ldots,n_M(f,t)\right]^T$$

其中，$\mathbf{y}(f,t)$表示含噪语音的观测信号，$\mathbf{r}^{(k)}(f)$表示第 k 个信号源和各个传声器之间的频率响应，$s^{(k)}(f,t)$表示源信号的短时傅立叶变换，$\mathbf{n}(f,t)$表示噪声信号的短时傅立叶变换，T 表示非共轭转置；

步骤5，将图像语音VAD概率通过非线性的映射函数得到映射后图像语音概率，映射后图像语音概率与相应帧的所对应音频信号的短时傅里叶频谱进行时域上的加权操作，进行图像VAD和传声器阵列信号的多模融合，得到图像VAD加权后的传声器阵列信号频谱；

$$t(x(\tau))=\frac{exp(10x(\tau))-exp(-10x(\tau))}{exp(10x(\tau))+exp(-10x(\tau))}$$

$$\mathbf{y}_{f,t}\leftarrow\mathbf{y}_{f,t}*t(x(\tau))$$

其中，$t(x(\tau))$表示映射后图像语音概率，$x(\tau)$是图像语音VAD概率，τ是图像的时间帧索引，$\mathbf{y}_{f,t}$表示短时傅里叶频谱，f表示频域，t表示时刻；

步骤6，将得到的图像VAD加权后的传声器阵列信号频谱输入基于复数高斯混合模型CGMM的时频掩模估计器，然后用最大似然法估计CGMM 参数，得到时频掩膜序列；然后对于所有频域点数，依次在线递归更新空间相关矩阵，含噪语音和噪声的协方差矩阵，以及聚类的混合权重；最后更新所有源的期望协方差矩阵并作时间平滑，分离它们的特征向量作为对应源导向矢量的估计，用MVDR波束的空间最优权矢量滤波器得到目标方向增强的语音信号；

基于复数高斯混合模型CGMM的时频掩模估计器采用结合图像信息的CGMM-MVDR在线方法：

初始化协方差矩阵$\mathcal{R}_{f,0}^{(v)}\leftarrow 0$，掩膜和$\Lambda_{f,0}^{(v)}\leftarrow 0$，聚类的混合权重$\alpha_f^{(v)}=\frac{1}{K+1}$，$v\in\{k+n,n,k\}$分别表示含噪语音、噪声、干净语音；

首先通过基于复数高斯混合模型CGMM的时频掩模估计器进行CGMM的EM方法掩膜估计，在掩膜估计期望步骤中后验概率用以下式子计算：

$$\lambda_{f,t}^{(v)}\leftarrow\frac{\alpha_f^{(v)}p\left(y_{f,t}\mid v,\Theta'\right)}{\sum_v\alpha_f^{(v)}p\left(y_{f,t}\mid v,\Theta'\right)}$$

其中，$\lambda_{f,t}^{(v)}$表示v类的掩膜，$\alpha_f^{(v)}$表示v类的混合权重，p表示条件概率，v表示含噪语音、噪声、干净语音中的任一一类，Θ表示一系列CGMM参数；

步骤5得到的图像VAD加权后的传声器阵列信号频谱得到混合权重为$\alpha_f^{(v)}$的复高斯混合模型：

$$\mathbf{y}_{f,t}\sim\sum_v\alpha_f^{(v)}\mathcal{N}_c\left(0,\phi_{f,t}^{(v)}\mathbf{R}_f^{(v)}\right)$$

其中，$\mathcal{N}_c$表示复数高斯混合分布，$\phi_{f,t}^{(v)}$表示时频点的信号方差，$\mathbf{R}_f^{(v)}$表示v类的空间相关矩阵；

具有均值μ和协方差矩阵Σ的多元复高斯分布为：

$$\mathcal{N}_c(\mathrm{x}\,|\,\mu,\sigma)=\frac{1}{|\pi\Sigma|}exp\left(-(\mathrm{x}-\mu)^H\,\Sigma^{-1}(\mathrm{x}-\mu)\right)$$

其中，$\mathcal{N}_c(\mathrm{x}\,|\,\mu,\sigma)$表示随机变量为X均值为$\mu$方差为$\sigma$的复数高斯混合分布，$\mathrm{x}$表示随机变量，$\mu$表示均值，$\sigma$表示方差，$H$表示共轭转置；

在掩膜估计最大化步骤中，CGMM 参数用以下式子更新：

$$\varphi_{f,t}^{(v)}\leftarrow\frac{1}{\mathrm{M}}\mathrm{tr}(\mathbf{y}_{f,t}\mathbf{y}_{f,t}^{\mathrm{H}}\mathbf{R}_f^{(v)-1})$$

其中，$\varphi_{\mathrm{f,t}}^{(\mathrm{v})}$表示$v$类时频点的信号方差，$\mathrm{M}$表示传声器个数，$\mathrm{tr}$表示取矩阵的迹，$\mathbf{y}_{f,t}$表示含噪语音的观测信号的时频点，$\mathbf{R}_f^{(v)-1}$表示空间相关矩阵取逆；

被最大化的 Q 函数为：

$$\begin{aligned}Q(\Theta\,|\,\Theta')&=\mathbb{E}\left[log\,p(\mathrm{y}|\Theta,\mathrm{v})\right]_{\mathrm{v}}\\&=\sum_{f,t}\sum_{v}\lambda_{f,t}^{(v)}log\,\alpha_f^{(v)}\mathcal{N}_c\left(\mathrm{y}_{f,t}\,|\,0,\phi_{f,t}^{(v)}\mathrm{R}_f^{(v)}\right)\\&=\sum_{f,t}\sum_{v}\lambda_{f,t}^{(v)}\Big\{log\,\alpha_f^{(v)}-M\,log\,\phi_{f,t}^{(v)}-log\,det\,\mathrm{R}_f^{(v)}\\&\quad-\frac{1}{\phi_{f,t}^{(v)}}tr\left(\mathrm{y}_{f,t}\mathrm{y}_{f,t}^{H}\mathrm{R}_f^{(v)-1}\right)\Big\}\end{aligned}$$

直至EM方法迭代达到指定次数。

2.根据权利要求1所述服务型机器人语音交互的音视频混合语音前端处理方法，其特征在于：

EM方法迭代指定次数后，第B_l批处的空间相关矩阵由下式递归估计：

$$\begin{aligned}\mathrm{R}_{f,l}^{(v)}&\leftarrow\frac{\Lambda_{f,l-1}^{(v)}}{\Lambda_{f,l-1}^{(v)}+\sum_{t\in B_l}\lambda_{f,t}^{(v)}}\mathrm{R}_{f,-1}^{(v)}\\&\quad+\frac{1}{\Lambda_{f,l-1}^{(v)}+\sum_{t\in B_l}\lambda_{f,t}^{(v)}}\sum_{t\in B_l}\lambda_{f,t}^{(v)}\frac{1}{\phi_{f,t}^{(v)}}\mathrm{y}_{f,t}\mathrm{y}_{f,t}^{H}\end{aligned}$$

含噪语音和噪声的协方差矩阵被在线递归更新为：

$$\mathcal{R}_{f,l}^{(v)} \leftarrow \frac{\Lambda_{f,l-1}^{(v)}}{\Lambda_{f,l-1}^{(v)} + \sum_{t \in B_l} \lambda_{f,t}^{(v)}} \mathcal{R}_{f,l-1}^{(v)} + \sum_{t \in B_l} \lambda_{f,t}^{(v)} \sum_{t \in B_l} \lambda_{f,t}^{(v)} \mathbf{y}_{f,t} \mathbf{y}_{f,t}^{H}$$

递归更新混合权重：

$$\alpha_{f,l}^{(v)} \leftarrow \frac{1}{T} \sum_{t} \lambda_{f,t}^{(v)}$$

更新所有源的期望协方差矩阵。

3.根据权利要求2所述服务型机器人语音交互的音视频混合语音前端处理方法，其特征在于：通过导向向量估计器进行导向向量估计：

先计算含噪语音$k+n$和噪声n的协方差矩阵估计：

$$\mathcal{R}_{f,l}^{(v)} = \frac{1}{\sum_t \lambda_{f,t}^{(v)}} \sum_t \lambda_{f,t}^{(v)} \mathbf{y}_{f,t} \mathbf{y}_{f,t}^{H}$$

得到k-th语音信号协方差矩阵估计：

$$\mathcal{R}_{f,l}^{(k)} = \mathcal{R}_{f,l}^{(k+n)} - \mathcal{R}_{f,l}^{(n)}$$

然后对$\mathcal{R}_{f,l}^{(k)}$执行特征向量分解，提取最大特征值相关联的特征向量作为导向向量$\mathbf{r}_{f,l}^{(k)}$的估计；

最后进行MVDR波束形成，得到增强语音；

MVDR波束的k-th源的滤波器系数：

$$\mathbf{w}_{f,l}^{(k)} = \frac{\mathcal{R}_{f,l}^{(n)-1} \mathbf{r}_{f,l}^{(k)}}{\mathbf{r}_{f,l}^{(k)H} \mathcal{R}_{f,l}^{(n)-1} \mathbf{r}_{f,l}^{(k)}}$$

得到增强的k-th源信号估计：

$$\widehat{s}_{f,t}^{(k)} = \mathbf{w}_{f,l}^{(k)H} \mathbf{y}_{f,t}$$

$\widehat{s}_{f,t}^{(k)}$表示增强的k-th源信号估计。

4.根据权利要求3所述服务型机器人语音交互的音视频混合语音前端处理方法，其特征在于：由于只针对某一批次的每个时间点$t \in B_l$，结束这一批次以后，需要更新掩膜和：

$$\Lambda_{f,l}^{(v)} \leftarrow \Lambda_{f,l-1}^{(v)} + \sum_{t \in B_l} \lambda_{f,t}^{(v)}$$

然后进行下次批次的更新，直到音频结束。

5.根据权利要求4所述服务型机器人语音交互的音视频混合语音前端处理方法，其特征在于：步骤4中使用清晰视频数据集的音频、多通道噪声数据集，并根据相应的传声器空间位置以及随机声源位置模拟出相应传声器采集到的含噪语音。

服务型机器人语音交互的音视频混合语音前端处理方法

技术领域

[0001] 本发明属于语音信号处理的技术领域,具体涉及一种复杂环境中使用传声器阵列的语音前端,用于提升服务型机器人的语音采集质量。

背景技术

[0002] 语音交互系统,作为最快捷有效的智能人机交互系统,在我们的生活中无处不在。语音交互系统需要在不同的场景下捕捉使用者的说话音频,在语音增强与分离等预处理步骤后进行自动语音识别(automatic speech recognition, ASR)。在远场、嘈杂等声学环境恶劣的情况下,识别准确率迅速下降。为了提高系统的鲁棒性,需要利用各种算法进行语音增强以提高语音的质量和可靠度。语音增强主要包括:语音分离、语音去混响和语音降噪,三者要解决的干扰分别来源于其他说话人的声音信号、空间环境对声音信号反射产生的混响和各种环境噪声。语音增强通过有效抑制这些噪声或人声来提高语音质量,现已应用于语音识别、助听器以及电话会议等。

[0003] 传声器阵列指两个或以上的传声器单元以特定空间位置排列组成的声学系统,配合信号处理方法,能够达到声源定位、盲源分离、声全息和语音增强等目的。此技术在传统的通信、生物医学工程等领域以及最近热门的虚拟现实(VR)、增强现实(AR)和人工智能(AI)领域皆有广泛的应用前景。基于阵列的增强方案包括阵列波束形成(beamforming)与盲源分离(HIGUCHI T, ITO N, YOSHIOKA T, et al. Robust MVDR beamforming using timefrequency masks for online/offline ASR in noise[C] // 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2016 :5210 - 5214.)等。

[0004] 传声器阵列波束形成,即按照阵列和声源的相关空间位置的导向矢量(steering vector)设计空间滤波器。按照空间滤波器参数可变与否,分为固定波束形成和自适应波束形成。固定波束由于滤波器参数不可调整,具有相对自适应波束更差的抗干扰能力和分辨率。当声源位置时变时,固定波束性能显著下降。但是其运算量较小、易于实现、且对传声器和声源位置的准确性有更好的鲁棒性。

[0005] 固定波束设计的设计目标是使波束主瓣指向目标声源,达到增强声源信号,抑制其他方向噪声信号的目的。延时求和(delay and sum, DS)波束(BRANDSTEIN M, WARD D. Microphone arrays: signal processing techniques and applications[M]. [S.l.] : Springer Science & Business Media, 2013.)是最常用的固定波束算法,它对于扰动鲁棒性好,但是主瓣随频率升高而变窄,即频率越高指向性越强,导致信号低通畸变。另外,延时求和波束要获得好的指向,需要足够多的单元数量。固定波束算法难以设计具有任意指向性的波束,而宽带波束的方法可以根据不同的代价函数和滤波求和结构,设计满足空间特征的波束:最小二乘法(least square, LS)、特征滤波器法(eigenfilter method)、基于阵列特征参数的方法、非线性优化波束(DOCLO S. Multimicrophone noise reduction and dereverberation techniques for speech applications[J], 2003.)等。

[0006] 自适应波束设计结合了波束指向性和空间信息自适应的特点,通过一定的迭代方式使实际响应接近期望响应。自适应波束根据不同的策略,如线性约束最小化方差(linearly constrained minimum variance, LCMV)策略、广义旁瓣抵消(generalized sidelobe cancellation, GSC)策略等。其中,LCMV的应用之一最小方差无失真响应(minimum variance distortionless response, MVDR)波束是应用得最广泛的自适应波束之一,也是本发明阵列的波束形成策略。

[0007] 常用的语音增强算法一般将处理重点放在音频信号本身。而人脑在处理别人传达的信息时,往往是将多种模态的信息,例如肢体语言、嘴唇动作和面部表情等,融合在一起处理的。与之类似,在设计语音增强解决方案时,若能充分关注这些多模特征,有望进一步提升系统性能。另外,在机器人人机交互、车载交互、视频会议等语音交互系统中,信息的传入设备同时包含传声器(阵列)和摄像头,这也为结合视频信息处理语音增强问题提供了基本的硬件条件。

[0008] 图像序列的行为识别任务有一个通用的框架,即用卷积神经网络(convolutional neural networks, CNN)提取特征,再通过几层循环神经网络(recurrent neural network, RNN)以方便利用帧与帧之间的关联信息(DONAHUE J, ANNE HENDRICKS L, GUADARRAMA S, et al. Longterm recurrent convolutional networks for visual recognition and description[C]// Proceedings of the IEEE conference on computer vision and pattern recognition.2015 : 2625 - 2634.)。本发明也采取类似的网络设置来对唇部图像的VAD判决进行预测,以期望达到图像唇读VAD的SOTA方案的准确性。

发明内容

[0009] 发明目的:传统只依赖音频信息的语音增强方法在对低信噪比、非稳态噪声、强混响环境下的语音进行增强时往往难以去除噪声成分,这对后续机器人的语音识别和语义理解造成了巨大的困难,本发明提供一种服务型机器人语音交互的音视频混合语音前端处理方法,本发明提出结合图像和视频分析的多模语音增强方案,其具有不错的鲁棒性,并且在低信噪比时对语音识别效果提升很明显。

[0010] 技术方案:为实现上述目的,本发明采用的技术方案为:

[0011] 一种服务型机器人语音交互的音视频混合语音前端处理方法,包括以下步骤:

[0012] 步骤1,模型训练:采集训练音视频样本,将训练音视频样本中视频部分按帧分成图像,将训练音视频样本中语音部分按对应帧图像进行标签,得到对应帧的干净语音VAD标签。将图像和对应帧的干净语音VAD标签导入CNN-RNN图像分类网络中,对图像中的唇动状态和应帧的干净语音VAD标签进行训练,得到训练好的CNN-RNN图像分类网络。

[0013] 步骤2,采集目标说话人嘴部动作视频和对应的含噪语音。嘴部动作视频用卷积神经网络方法标记出目标说话人人脸五官定位,并裁出嘴唇区域图像。嘴唇区域图像逐帧进行灰度图重塑得到嘴唇区域灰度图像,将嘴唇区域灰度图像输入到图像活动语音检测器。

[0014] 步骤3,图像活动语音检测器根据输入的嘴唇区域灰度图像检测到目标说话人正在说话,则将嘴唇区域灰度图像输入到训练好的CNN-RNN图像分类网络中,得到此帧嘴唇区域灰度图对应的图像语音VAD概率。

[0015] 步骤4,对含噪语音做短时傅里叶变换得到短时傅里叶频谱。

[0016] 使用清晰视频数据集的音频、多通道噪声数据集,并根据相应的传声器空间位置以及随机声源位置模拟出相应传声器采集到的含噪语音。

[0017] 步骤5,将图像语音VAD概率通过非线性的映射函数得到映射后图像语音概率,映射后图像语音概率与相应帧的所对应音频信号的短时傅里叶频谱进行时域上的加权操作,进行图像VAD和传声器阵列信号的多模融合,得到图像VAD加权后的传声器阵列信号频谱。

[0018] 本发明相比现有技术,具有以下有益效果:

[0019] 本发明从传声器阵列和声源的相对空间位置入手,利用复高斯混合模型(CGMM),期望最大化(EM)方法以及最小方差无失真响应(MVDR)波束来增强目标源方向的语音。其中时频掩模的使用能够避免使用不准确的先验知识,例如阵列几何和平面波传播假设,从而提供稳健的导向矢量估计。在此基础上,为了提高在低信噪比、非稳态噪声等多种复杂噪声场景下算法的有效性,采用了对噪声不敏感的图像模态的信息作为补充,用唇部图像生成可靠的VAD判决。在 CGMM分类系统的前端融合 VAD 可有效提高语音时频掩膜的准确性,从而得到更好的音质和语音可懂度,为后续语音识别任务提供更优质的前端输入。

附图说明

[0020] 图1是本发明的结合图像和视频处理的多模语音增强处理流程图。

[0021] 图2是用卷积神经网络方法标记出目标说话人人脸的五官定位,并裁出嘴唇区域的处理结果。

[0022] 图3为嘴唇图像处理部分的2D CNN-RNN神经网络的框架,其中包括二维卷积层组成的编码器,随后经过长短期记忆网络块,接着得到此刻唇动状态VAD的预测。

[0023] 图4为一个声源时的问题框架示意图。

[0024] 图5为模拟含噪语音生成的空间示意图。

具体实施方式

[0025] 下面结合附图和具体实施例,进一步阐明本发明,应理解这些实例仅用于说明本发明而不用于限制本发明的范围,在阅读了本发明之后,本领域技术人员对本发明的各种等价形式的修改均落于本申请所附权利要求所限定的范围。

[0026] 一种服务型机器人语音交互的音视频混合语音前端处理方法,如图1所示,包括以下步骤:

[0027] 步骤1,模型训练:采集训练音视频样本,将训练音视频样本中视频部分按帧分成图像,将训练音视频样本中语音部分按对应帧图像进行标签,得到对应帧的干净语音VAD标签。将图像和对应帧的干净语音VAD标签导入CNN-RNN图像分类网络中,对图像中的唇动状态和应帧的干净语音VAD标签进行训练,得到训练好的CNN-RNN图像分类网络。

[0028] 步骤2,采集目标说话人嘴部动作视频和对应的含噪语音。嘴部动作视频用卷积神经网络方法标记出目标说话人人脸五官定位,并裁出嘴唇区域图像,截取如图2所示。嘴唇区域图像逐帧进行90×110像素的灰度图重塑,并归一化数据格式到16位浮点数,得到嘴唇区域灰度图像,将嘴唇区域灰度图像输入到图像活动语音检测器。

[0029] 步骤3,图像活动语音检测器根据输入的嘴唇区域灰度图像检测到目标说话人正

在说话，则将嘴唇区域灰度图像输入到训练好的CNN-RNN图像分类网络中，得到此帧嘴唇区域灰度图对应的图像语音VAD概率，如图3所示，首先第一列嘴唇区域灰度图序列经过二维卷积层组成的编码器，随后经过长短期记忆网络块，接着得到此刻唇动状态的预测，输出根据图像信息判决此帧为图像语音VAD概率。

[0030] 步骤4，使用清晰视频数据集的音频、多通道噪声数据集，并根据相应的传声器空间位置以及随机声源位置模拟出相应传声器采集到的含噪语音，如图5所示，对含噪语音做短时傅里叶变换得到短时傅里叶频谱，其中信号处理的参数设置见表1。

[0031] 表1 音频算法的实验参数

[0032]

采样频率	16kHZ
帧长	32ms
帧重叠比例	50%
窗函数	Hanning
EM 算法迭代次数	3
最大麦克风数	6
在线处理一批的长度	640ms

[0033] 步骤5，将图像语音VAD概率通过非线性的映射函数得到映射后图像语音概率，映射后图像语音概率与相应帧的所对应音频信号的短时傅里叶频谱进行时域上的加权操作，进行图像VAD和传声器阵列信号的多模融合，得到图像VAD加权后的传声器阵列信号频谱。

[0034] 其中映射函数的定义域和值域都在[0,1]，可以理解为一种额外设计的激活函数，目的是为了让加权操作更加平滑。映射具体函数关系见式(1)，加权方式见式(2)：

[0035] $$t\left(x(\tau)\right)=\frac{exp\left(10x(\tau)\right)-exp\left(-10x(\tau)\right)}{exp\left(10x(\tau)\right)+exp\left(-10x(\tau)\right)} \quad (1)$$

[0036] $$y_{f,t} \leftarrow y_{f,t} * t\left(x(\tau)\right) \quad (2)$$

[0037] 其中，$t\left(x(\tau)\right)$表示映射后图像语音概率，$x(\tau)$是图像语音VAD概率，即CNN-RNN图像分类网络预测结果，τ是图像的时间帧索引，$y_{f,t}$表示短时傅里叶频谱，f表示频域，t表示时刻。

[0038] 步骤6，将得到的图像VAD加权后的传声器阵列信号频谱输入基于复数高斯混合模型CGMM的时频掩模估计器，然后用最大似然法估计CGMM 参数，得到时频掩膜序列。然后对于所有频域点数，依次在线递归更新空间相关矩阵，含噪语音和噪声的协方差矩阵，以及聚类的混合权重。最后更新所有源的期望协方差矩阵并作时间平滑，分离它们的特征向量作为对应源导向矢量的估计，用MVDR波束的空间最优权矢量滤波器得到目标方向增强的语音信号。

[0039] 一、问题框架

[0040] $k \in \{1, \ldots, K\}$是源索引，K表示源信号个数，$m \in \{1, \ldots, M\}$ 是传声器索引，M表示传声器个数。在时域中，第 m 个传声器的语音信号$y_m(t)$可以写为：

[0041] $$y_m(t)=\sum_k\sum_\tau h_m^{(k)}(\tau)s^{(k)}(t-\tau)+n_m(t) \quad (3)$$

[0042] 其中，$y_m(t)$表示第 m 个传声器的语音信号，$s^{(k)}(t)$表示第 k 个源信号的噪声信号，$n_m(t)$表示第 m 个传声器采集到的噪声信号，$h_m^{(k)}(\tau)$表示对应于第k 个源和第 m 个传声器之间的脉冲响应，如图4所示，τ是图像的时间帧索引，t表示时刻。

[0043] 第 m 个传声器的语音信号$y_m(t)$通过应用短时傅立叶变换(shorttime Fourier transform, STFT)，公式 (3)可以在频域中表示为：

[0044] $$y_m(f,t)=\sum_k h_m^{(k)}(f)s^{(k)}(f,t)+n_m(f,t) \quad (4)$$

[0045] 其中，$y_m(f,t)$为$y_m(t)$的频域表示，$h_m^{(k)}(f)$为$h_m^{(k)}(\tau)$的频域表示，$s^{(k)}(f,t)$为$s^{(k)}(t)$的频域表示，$n_m(f,t)$为$n_m(t)$的频域表示。

[0046] 这里我们假设脉冲响应的长度远小于 STFT 窗口的长度，因此，脉冲响应和源信号在时域中的卷积表示为时不变频率响应和时变源信号在频域中的乘积，引入矢量符号，式(4)可以改写为：

[0047] $$\mathbf{y}(f,t)=\sum_k \mathbf{r}^{(k)}(f)s^{(k)}(f,t)+\mathbf{n}(f,t) \quad (5)$$

[0048] 其中：

[0049] $$\mathbf{y}(f,t)=\left[y_1(f,t),y_2(f,t),\ldots,y_M(f,t)\right]^T$$

[0050] $$\mathbf{r}^{(k)}(f)=\left[h_1^{(k)}(f),h_2^{(k)}(f),\ldots,h_M^{(k)}(f)\right]^T \quad (6)$$

[0051] $$\mathbf{n}(f,t)=\left[n_1(f,t),n_2(f,t),\ldots,n_M(f,t)\right]^T$$

[0052] 其中，$\mathbf{y}(f,t)$表示被噪声混合的观测信号，$\mathbf{r}^{(k)}(f)$表示第 k 个信号源和各个传声器之间的频率响应， $\mathbf{r}^{(k)}(f)$ 是导向矢量，$s^{(k)}(f,t)$表示源信号的短时傅立叶变换，$\mathbf{n}(f,t)$表示噪声信号的短时傅立叶变换，T 表示非共轭转置。

[0053] 源分离 (或语音增强) 问题的目标是凭借被噪声混合的观测信号$\mathbf{y}(f,t)$估计每个目标源信号$s_{f,t}^{(k)}$。

[0054] 二、结合图像信息的CGMM-MVDR在线方法：

[0055] 初始化协方差矩阵$\mathcal{R}_{f,0}^{(v)} \leftarrow 0$，掩膜和$\Lambda_{f,0}^{(v)} \leftarrow 0$，聚类的混合权重$\alpha_f^{(v)}=\frac{1}{K+1}$，

取前1000ms作为空间相关矩阵$\mathbf{R}_{f,0}^{(v)}$的粗略估计。$v\in\{k+n,n,k\}$分别表示含噪语音、噪声、干净语音。

[0056] 首先通过基于复数高斯混合模型CGMM的时频掩模估计器进行CGMM的EM方法掩膜估计，在掩膜估计期望步骤(E step)中后验概率用以下式子计算：

[0057] $$\lambda_{f,t}^{(v)}\leftarrow\frac{\alpha_f^{(v)}p\left(y_{f,t}|v,\Theta'\right)}{\sum_v\alpha_f^{(v)}p\left(y_{f,t}|v,\Theta'\right)} \tag{7}$$

[0058] 其中，$\lambda_{f,t}^{(v)}$表示v类的掩膜，$\alpha_f^{(v)}$表示v类的混合权重，p表示条件概率，v表示含噪语音、噪声、干净语音中的任意一类，Θ表示一系列CGMM参数。

[0059] 步骤5得到的图像VAD加权后的传声器阵列信号频谱得到混合权重为$\alpha_f^{(v)}$的复高斯混合模型，如下所示：

[0060] $$\mathrm{y}_{f,t}\sim\sum_v\alpha_f^{(v)}\mathcal{N}_c\left(0,\phi_{f,t}^{(v)}\mathrm{R}_f^{(v)}\right) \tag{8}$$

[0061] 其中，$\mathcal{N}_c$表示复数高斯混合分布，$\phi_{f,t}^{(v)}$表示时频点的信号方差，$\mathbf{R}_f^{(v)}$表示v类的空间相关矩阵。

[0062] 具有均值μ和协方差矩阵Σ的多元复高斯分布为：

[0063] $$\mathcal{N}_c\left(\mathrm{x}|\mu,\sigma\right)=\frac{1}{|\pi\Sigma|}exp\left(-\left(\mathrm{x}-\mu\right)^H\Sigma^{-1}\left(\mathrm{x}-\mu\right)\right) \tag{9}$$

[0064] 其中，$\mathcal{N}_c\left(\mathrm{x}|\mu,\sigma\right)$表示随机变量为X均值为$\mu$方差为$\sigma$的复数高斯混合分布，$\mathrm{x}$表示随机变量，$\mu$表示均值，$\sigma$表示方差，$H$表示共轭转置。

[0065] 在掩膜估计最大化步骤(M step)中，CGMM 参数用以下式子更新：

[0066] $$\varphi_{f,t}^{(v)}\leftarrow\frac{1}{\mathrm{M}}\mathrm{tr}(\mathbf{y}_{f,t}\mathbf{y}_{f,t}^{\mathrm{H}}\mathbf{R}_f^{(v)-1}) \tag{10}$$

[0067] 其中，$\varphi_{\mathrm{f,t}}^{(\mathrm{v})}$表示$v$类时频点的信号方差，$\mathrm{M}$表示空间相关矩阵的维度，$\mathrm{tr}$表示取矩阵的迹，$\mathbf{y}_{f,t}$表示含噪语音的观测信号的时频点，$\mathbf{R}_f^{(v)-1}$表示空间相关矩阵取逆。

[0068] 在每个 EM 迭代步骤里被最大化的 Q 函数为：

[0069] $$\begin{aligned}Q(\Theta|\Theta')&=\mathbb{E}\left[log\,p\left(\mathrm{y}|\Theta,\mathrm{v}\right)\right]_{\mathrm{v}}\\&=\sum_{f,t}\sum_v\lambda_{f,t}^{(v)}log\,\alpha_f^{(v)}\mathcal{N}_c\left(\mathrm{y}_{f,t}|0,\phi_{f,t}^{(v)}\mathrm{R}_f^{(v)}\right)\\&=\sum_{f,t}\sum_v\lambda_{f,t}^{(v)}\left\{log\,\alpha_f^{(v)}-M\,log\,\phi_{f,t}^{(v)}-log\,det\,\mathrm{R}_f^{(v)}\right.\\&\left.-\frac{1}{\phi_{f,t}^{(v)}}tr\left(\mathrm{y}_{f,t}\mathrm{y}_{f,t}^H\mathrm{R}_f^{(v)-1}\right)\right\}\end{aligned} \tag{11}$$

[0070] 直至EM方法迭代达到指定次数。

[0071] EM方法迭代指定次数后，第B_l批处的空间相关矩阵由下式递归估计：

[0072]
$$\mathrm{R}_{f,l}^{(v)} \leftarrow \frac{\Lambda_{f,l-1}^{(v)}}{\Lambda_{f,l-1}^{(v)}+\sum_{t\in B_l}\lambda_{f,t}^{(v)}}\mathrm{R}_{f,-1}^{(v)} + \frac{1}{\Lambda_{f,l-1}^{(v)}+\sum_{t\in B_l}\lambda_{f,t}^{(v)}}\sum_{t\in B_l}\lambda_{f,t}^{(v)}\frac{1}{\phi_{f,t}^{(v)}}\mathrm{y}_{f,t}\mathrm{y}_{f,t}^{H} \tag{12}$$

[0073] 含噪语音和噪声的协方差矩阵被在线递归更新为：

[0074]
$$\mathcal{R}_{f,l}^{(v)} \leftarrow \frac{\Lambda_{f,l-1}^{(v)}}{\Lambda_{f,l-1}^{(v)}+\sum_{t\in B_l}\lambda_{f,t}^{(v)}}\mathcal{R}_{f,l-1}^{(v)} + \sum_{t\in B_l}\lambda_{f,t}^{(v)}\sum_{t\in B_l}\lambda_{f,t}^{(v)}\mathrm{y}_{f,t}\mathrm{y}_{f,t}^{H} \tag{13}$$

[0075] 递归更新混合权重：

[0076]
$$\alpha_{f,l}^{(v)} \leftarrow \frac{1}{T}\sum_{t}\lambda_{f,t}^{(v)} \tag{14}$$

[0077] 以上步骤对于所有频率点都更新完，随后进行导向矢量的估计。

[0078] 通过导向向量估计器进行导向向量估计：

[0079] 先计算含噪语音$k+n$和噪声n的协方差矩阵估计：

[0080]
$$\mathcal{R}_{f,l}^{(v)} = \frac{1}{\sum_t \lambda_{f,t}^{(v)}}\sum_t \lambda_{f,t}^{(v)}\mathbf{y}_{f,t}\mathbf{y}_{f,t}^{H} \tag{15}$$

[0081] 得到k-th语音信号协方差矩阵估计：

[0082]
$$\mathcal{R}_{f,l}^{(k)} = \mathcal{R}_{f,l}^{(k+n)} - \mathcal{R}_{f,l}^{(n)} \tag{16}$$

[0083] 然后对$\mathcal{R}_{f,l}^{(k)}$执行特征向量分解，提取最大特征值相关联的特征向量作为导向向量$\mathbf{r}_{f,l}^{(k)}$的估计。

[0084] 最后进行MVDR波束形成，得到增强语音。

[0085] MVDR波束的k-th源的滤波器系数：

[0086]
$$\mathbf{w}_{f,l}^{(k)} = \frac{\mathcal{R}_{f,l}^{(n)-1}\mathbf{r}_{f,l}^{(k)}}{\mathbf{r}_{f,l}^{(k)H}\mathcal{R}_{f,l}^{(n)-1}\mathbf{r}_{f,l}^{(k)}} \tag{17}$$

[0087] 得到增强的k-th源信号估计：

[0088]
$$\widehat{s}_{f,t}^{(k)} = \mathbf{w}_{f,l}^{(k)H}\mathbf{y}_{f,t} \tag{18}$$

[0089] $\widehat{s}_{f,t}^{(k)}$表示增强的k-th源信号估计。

[0090] 由于是在线算法，故以上操作都只针对某一批次的每个时间点$t \in B_l$，结束这一批次以后，需要更新掩膜和：

[0091] $$\Lambda_{f,l}^{(v)} \leftarrow \Lambda_{f,l-1}^{(v)} + \sum_{t \in B_l} \lambda_{f,t}^{(v)} \tag{19}$$

[0092] 然后进行下次批次的更新，直到音频结束。

[0093] 三、.数据集与评价指标

[0094] 噪声来自DEMAND多通道噪声库，纯净目标源来自 TIMIT 库。共模拟数据 120(干净音频)*12(噪声种类)=1440(组)。对于在线处理，每个音频的前 1000ms，约31帧作为训练数据以估计可靠的初始空间相关矩阵。由于 TIMIT 库的音频说话开始时间皆小于1000ms，这样做是可行的。

[0095] 评价指标包括经常被用来衡量语音分离效果的尺度不变的信号失真比(SI-SDR)，其定义为

[0096] $$\begin{cases} \mathbf{x}_{target} := \dfrac{\langle \hat{\mathbf{x}}, \mathbf{x} \rangle \mathbf{x}}{\|\mathbf{x}\|_2^2}, \\ \mathbf{e}_{noise} := \hat{\mathbf{x}} - \mathbf{x}_{target}, \\ SI-SDR := 10 log_{10} \dfrac{\|\mathbf{x}_{target}\|_2^2}{\|\mathbf{e}_{noise}\|_2^2} \end{cases} \tag{20}$$

[0097] 其中，$\mathbf{x}$和$\hat{\mathbf{x}}$分别是干净语音和估计的目标语音，它们被零均值归一化以保证尺度不变性。$\mathbf{x}_{target}$表示干净语音在干净语音和估计语音相关系数的归一化的方向的投影，:=表示编程语言里的赋值语句的符号，$\mathbf{e}_{noise}$表示估计的噪声信号。

[0098] 除了SI-SDR之外，评价指标还有语音质量客观评价指标PESQ。

[0099] 四、实验结果

[0100] 对比是否结合图像信息的CGMM-MVDR在线算法，对不同信噪比混合语音处理前后的效果用处理前后指标的差值表示，数值越大代表改善越大，测试结果如表2所示：

[0101] 表2 测试结果

[0102]

混合处理	SNR	ΔSI-SDR	ΔPESQ
否	-10dB	6.07dB	1.54
	-5dB	6.27dB	1.50
	0dB	6.40dB	1.24
	5dB	6.14dB	1.07
是	-10dB	7.13dB	1.48
	-5dB	6.77dB	1.34
	0dB	6.45dB	1.16
	5dB	5.78dB	0.95

[0103] 标准CGMM-MVDR算法不含图像的多模处理，为混合处理为否的部分。它在含噪语音为0dB左右的时候SI-SDR改善最多，而PESQ则是含噪语音信噪比越低处理前后改善越多。因为含噪语音信噪比越低，初始分数越低。

[0104] 多模混合处理方案在极低信噪比SNR=-10dB时，相对于标准方案，SI-SDR还能再提高1.06dB，Babble类人声噪音此提高幅度更甚。由于多模融合时粗暴的幅度加权，PESQ效果略逊色。但是实际在使用时，由于多模检测为不说话的时间段上本来就不需要语音识别，所以PESQ的逊色只会影响听感，而不影响后续语音识别。反而准确的图像VAD判决会为后续的语音识别任务强调重点识别的地方，在目标说话人闭嘴时忽略其他类似的人声噪声。

[0105] 以上所述仅是本发明的优选实施方式，应当指出：对于本技术领域的普通技术人员来说，在不脱离本发明原理的前提下，还可以做出若干改进和润饰，这些改进和润饰也应视为本发明的保护范围。

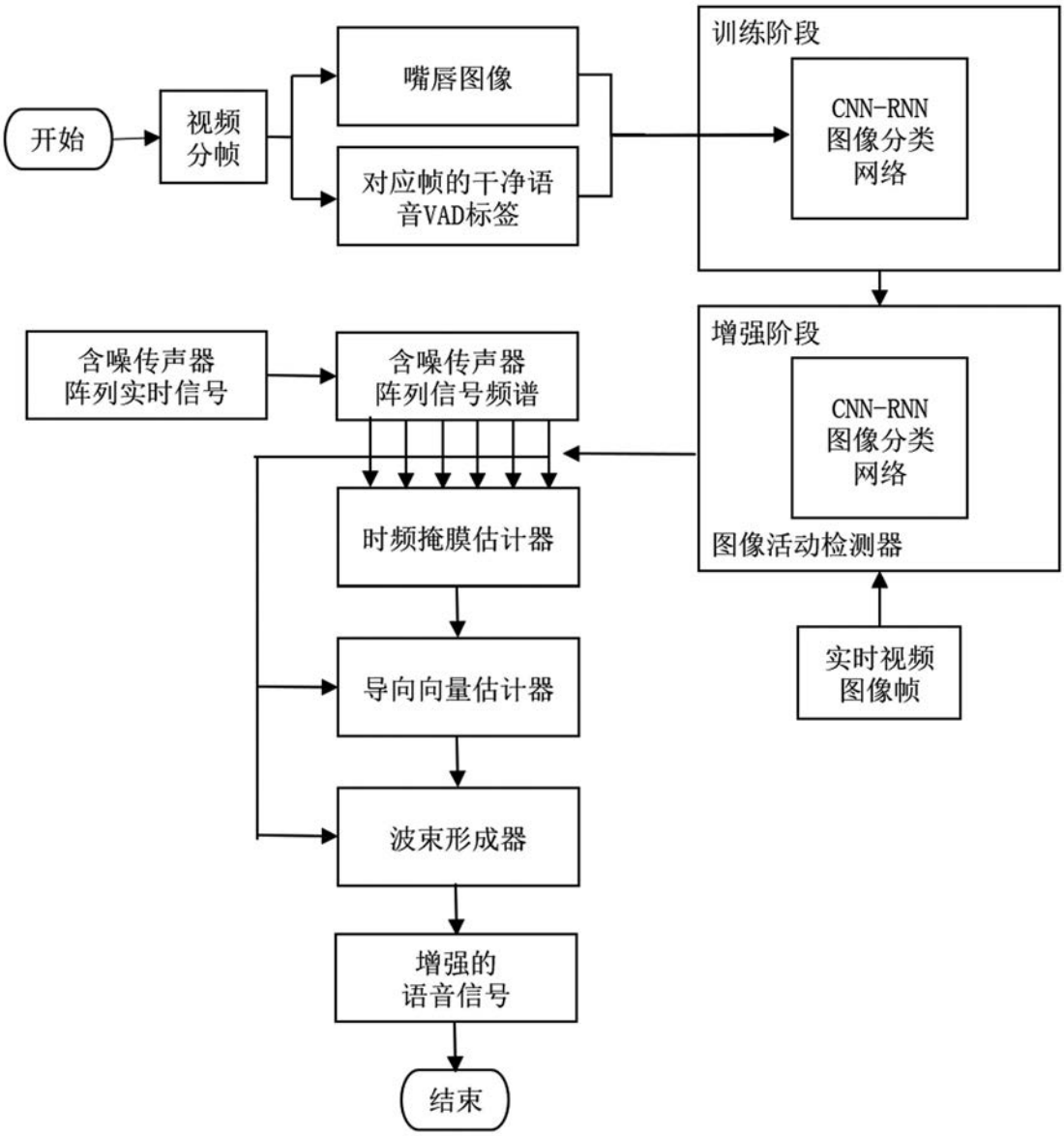

开始
视频分帧
嘴唇图像
对应帧的干净语音VAD标签
训练阶段
CNN-RNN图像分类网络
增强阶段
CNN-RNN图像分类网络
图像活动检测器
含噪传声器阵列实时信号
含噪传声器阵列信号频谱
时频掩膜估计器
导向向量估计器
波束形成器
增强的语音信号
结束
实时视频图像帧

图1

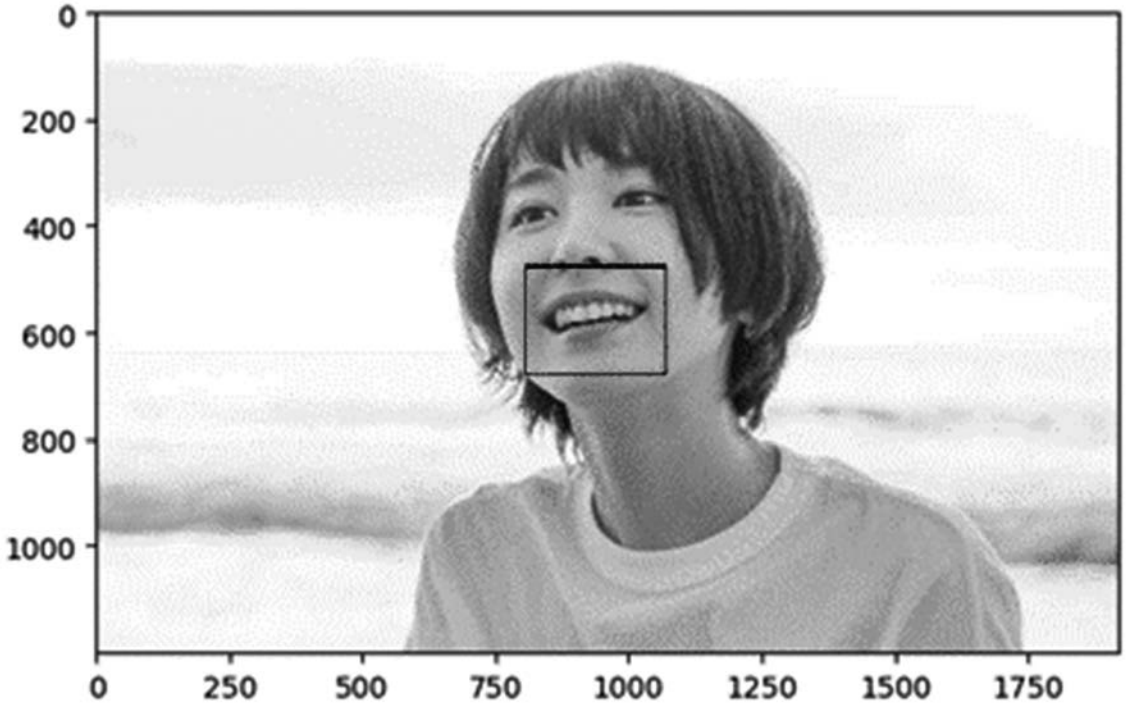

0
200
400
600
800
1000
0
250
500
750
1000
1250
1500
1750

图2

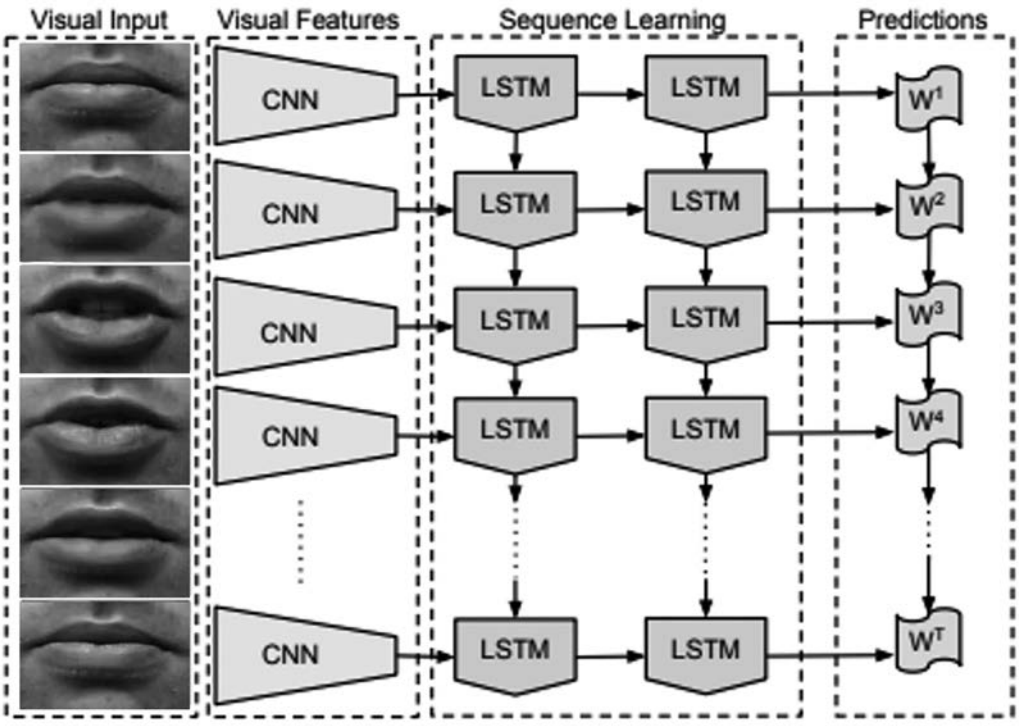

Visual Input
Visual Features
Sequence Learning
Predictions
CNN
LSTM
W1
W2
W3
W4
WT

图3

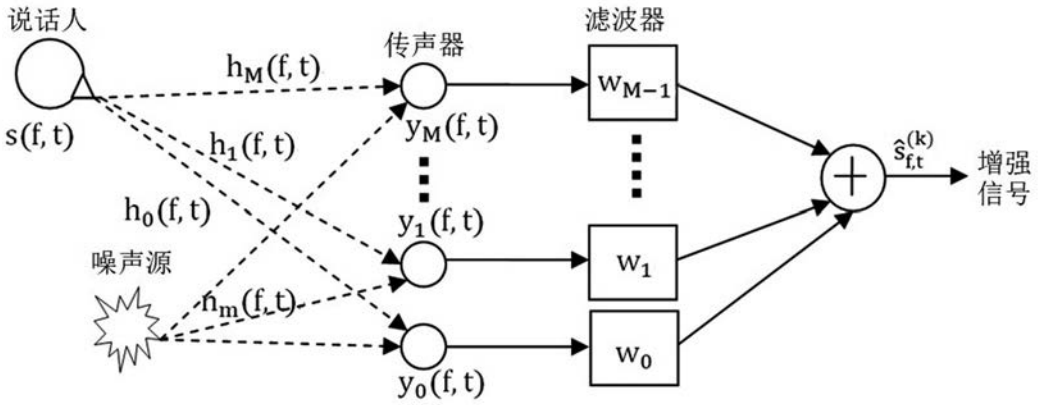

说话人
s(f,t)
传声器
$h_M(f,t)$
$h_1(f,t)$
$h_0(f,t)$
噪声源
$h_m(f,t)$
$y_M(f,t)$
$y_1(f,t)$
$y_0(f,t)$
滤波器
w_{M-1}
w_1
w_0
$\hat{s}_{f,t}^{(k)}$
增强信号

图4

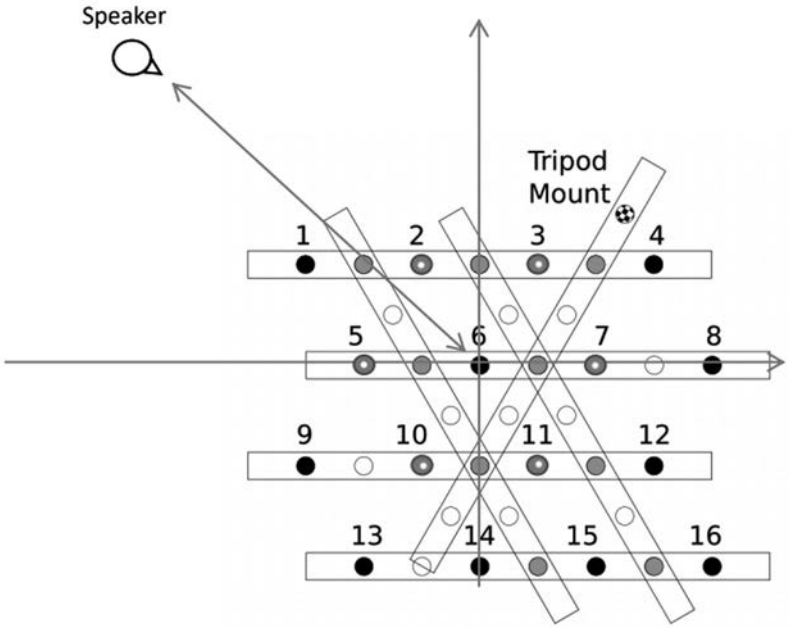

Speaker
Tripod Mount
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16

图5