一种高精度声纹谱提取方法

王东 李蓝天

背景

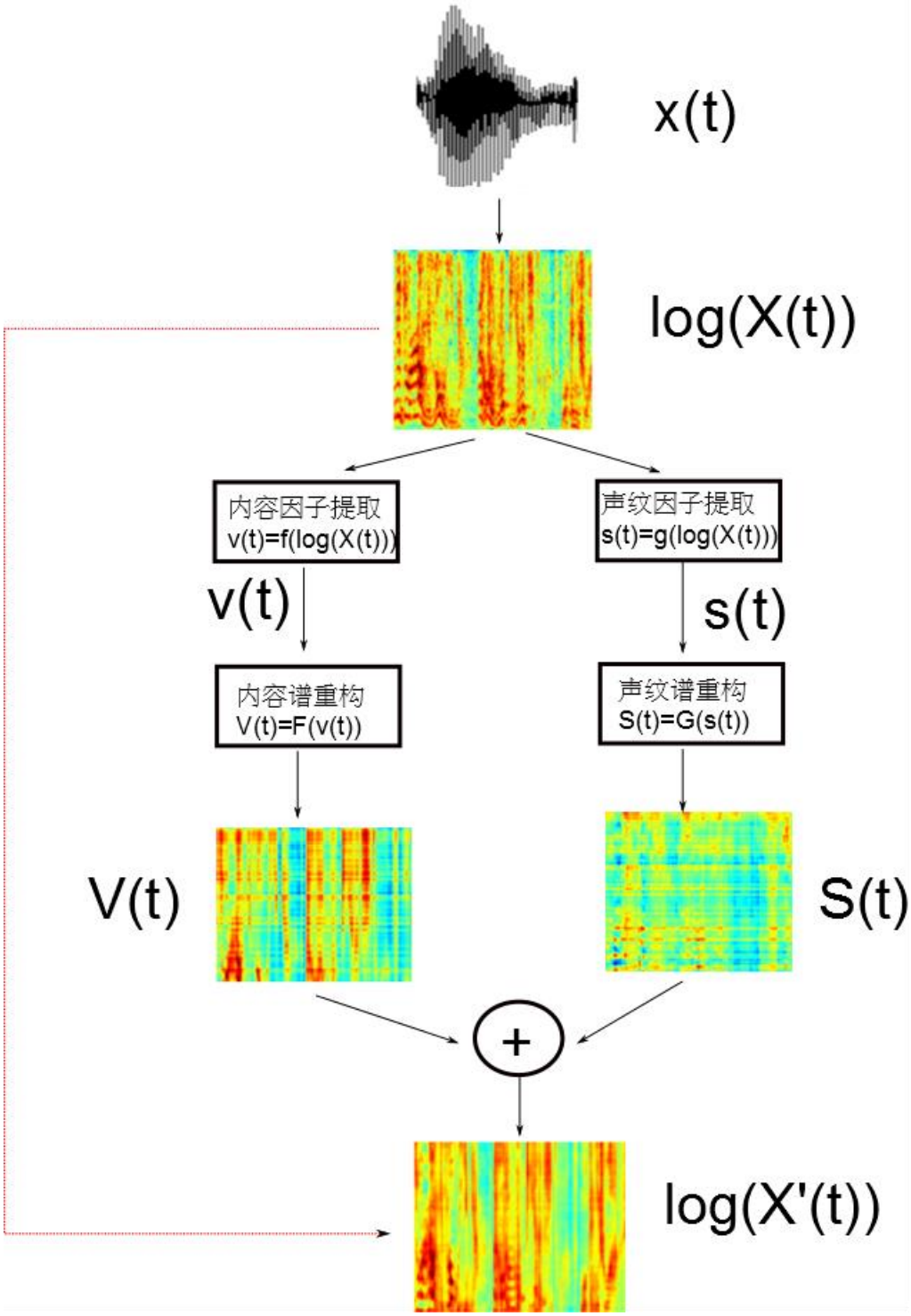
声纹是判断声音信号中包含的说话人特征。在司法实践中，声纹是对目标进行验证的有效工具之一，在司法鉴定领域具有重要意义。传统声纹比对方法一般通过频谱来实现，即先将声音转换成频谱，再由鉴定专家通过观察两段声音的频谱来判断这两段声音是否来源于同一个说话人。然而，频谱中不仅包含说话人信息，也包含说话内容信息，通常说话内容的变动更为显著，这意味着鉴定专家通过频谱看到的大多数信息是和发音变异相关的，与说话人其实没有太大关系。这种信息的混杂使得声纹比对非常困难，验证误差较大。在司法鉴定中，这种高误差率导致声纹验证无法列为重要的呈堂证供。

我们希望通过一种特殊处理过程，在语音频谱中滤除说话人内容信息，只保留说话人信息，即得到声纹谱，再基于此实现精准声纹验证方法。

发明内容

本发明提出一种因子分析与合成方法滤除语音频谱中的说话内容信息，留下说话人信息，从而生成说话人谱。该系统分为训练和生成两个过程，其中训练过程训练一个模型 M，该模型将频谱 X(t)转换成声纹谱 S(t)；生成过程基于模型 M 对待分析语生成相应的声纹谱。

模型训练过程


x(t)
log(X(t))
内容因子提取
v(t)=f(log(X(t)))
声纹因子提取
s(t)=g(log(X(t)))
v(t)
s(t)
内容谱重构
V(t)=F(v(t))
声纹谱重构
S(t)=G(s(t))
V(t)
S(t)
+
log(X'(t))

系统训练如上图所示。具体流程如下：

1. 给定一段语音信号，通过傅立叶变换将其转换成 log 域频谱 log(X(t))
2. 基于该频谱，对每一帧语音信号提取两类因子：内容因子 v(t)和说话人因子 s(t)，前者和说话内容相关，后者和说话人特征相关。提取模型可表示成两个函数 f 和 g，公式化如下：

   v(t) = f(log(X(t)))　　（1）

   s(t) = g(log(X(t)))　　（2）

3. 对提取模型 f 和 g 分别进行训练，使得生成的因子 v 和 s 尽可能代表说话内容和说话人信息。对于内容因子 v，我们通过训练 f 使其对音素的区分能力最大化；对于说话人因子 s，我们通过训练 g，使其对说话人的区分能力最大化。这一区分能力最大化准则可以有多种，一般可基于 Fisher 准则或交叉熵。
4. 基于因子 v(t)和 s(t)，训练重构模型 F 和 G，将内容因子重构成内容谱 V(t)，将说话人因子重构成声纹谱 S(t)：

   V(t)=F(v(t))　　（3）

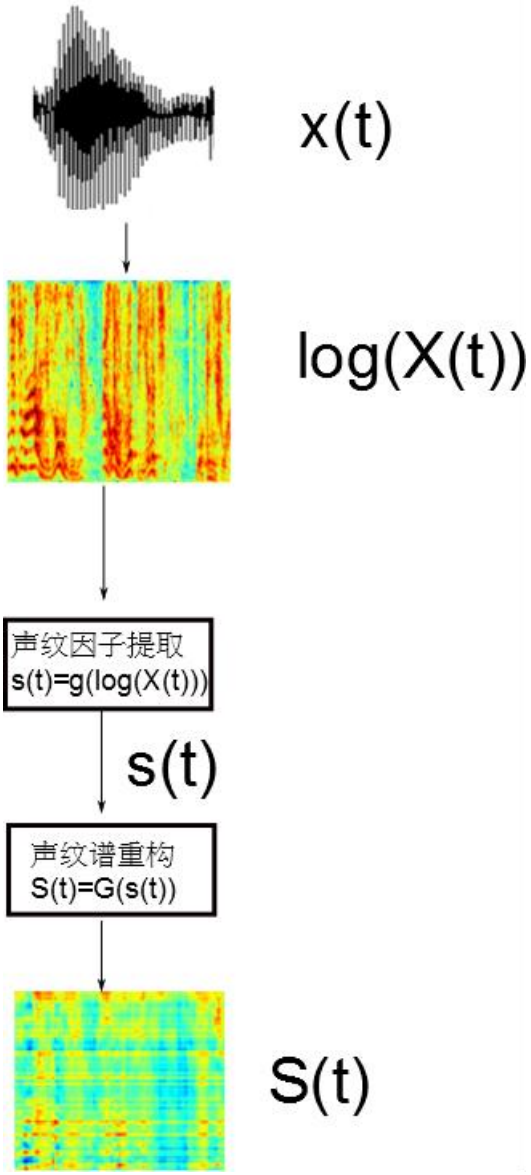
   S(t)=G(s(t))　　（4）

   这两者相加后即得重构的 log 域频谱，在训练时的目标是使重构后的 log 频谱与原 log 频谱尽可以相近，公式化为：

   $$L(F,G) = \sum_t D(\mathrm{V(t)} + \mathrm{S(t)}, \log(\mathrm{X(t)}))$$　　（5）

   其中 D 为距离函数。注意上式是重构模型 F 和 G 的函数，通过优化上式即可得到优化的 F 和 G。

声纹谱生成



系统训练完成以后，只保留声纹因子提取模型和声纹谱重构模型，即可实现由一个输入语音x(t)到声纹谱的生成过程。

实现实例

本发明提出的声纹谱生成方法具有通用性，其中的模型、距离度量、训练准则均可灵活定义。一种典型的实现实例如下：

1. 内容因子模型 f,声纹因子 g，内容重构模型 F，声纹重构模型 G 都基于深度神经网络实现。
2. 内容因子模型 f 采用音素作为区分学习的对象，用模型预测结果和实际音素标记的交叉熵作为训练的目标函数。
3. 声纹因子模型 g 采用说话人作为区分学习对象，用模型预结果和实际说话人权记的交叉熵作为训练的目标函数。
4. 公式(5)中的距离度量 D 采用平方误差。

发明优势

本方法掉了发音内容信息，仅保留声纹谱，可清晰观察到说话人特性，帮助鉴定专家提高验证精度。