基于内容-说话人分解的声码器

王东 李蓝天

背景

传统声码器一般基于源-滤波器（Source-Filter）模型，将声音信号 x(t)分解成声门激励 e(t)和声道调制 h(t)两部分，写成卷积形式如下：

x(t) = x(t) *e(t)

这一基础分解方式是很多语音信号处理技术的基础，如语音编码中的参数编码器，语音合成中的声码器。

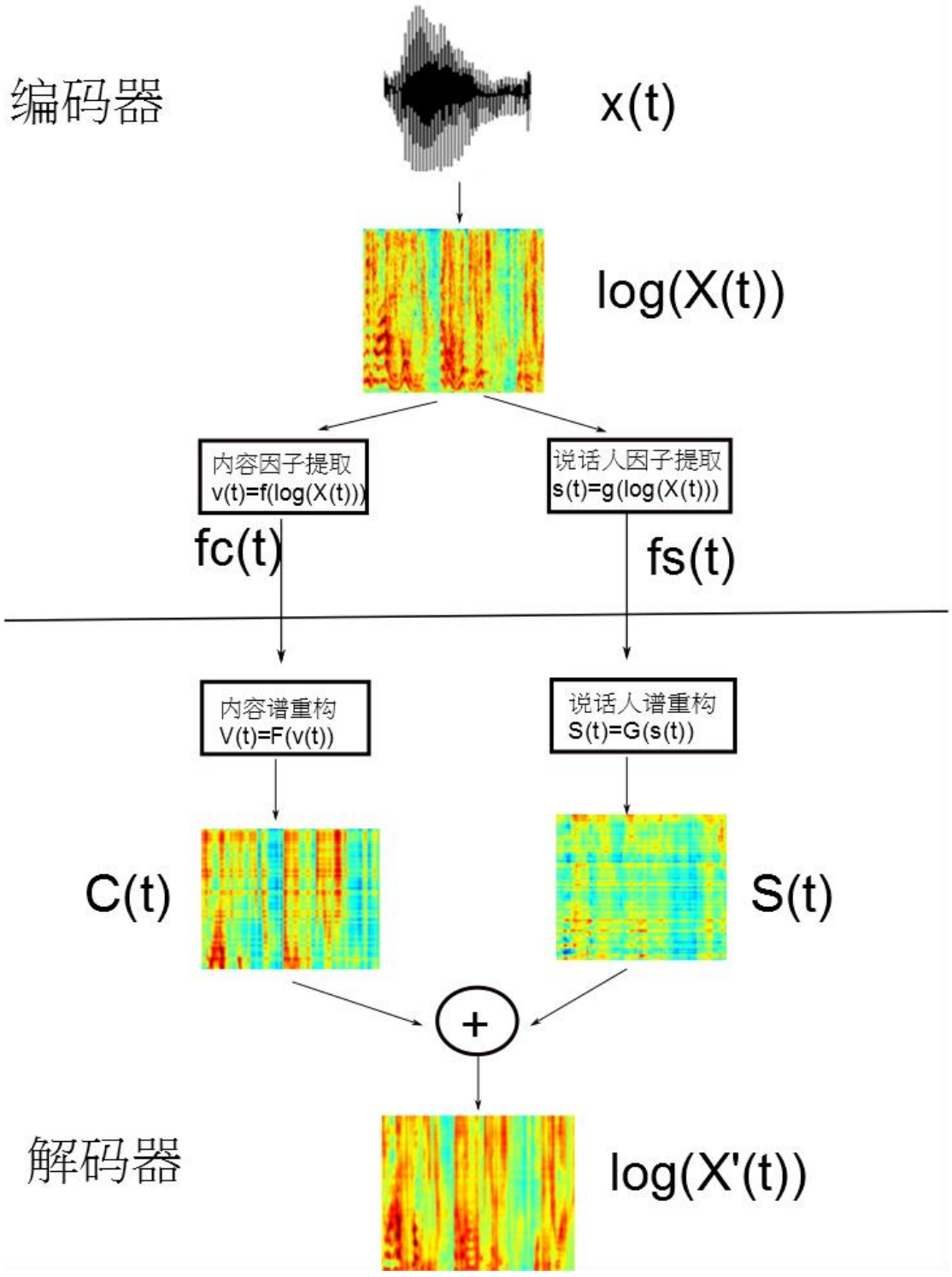
本发明提出一种基于内容-说话人的语音信号分解模型。和源-滤波器模型将信号分解成激励和调制的卷积不同，我们提出将声音分解成发音内容 c(t)与发音人特征 s(t)的卷积，即:

x(t) = c(t) * s(t)

这一内容-发音人分解可以用来设计一种全新的声码器。

发明内容

本发明提出一种将声音信号进行内容-说话人分解（Content-Speaker 分解，CS 分解）的方法，并依此设计一个有效的声码器。该系统结构如下图所示，其中包括一个编码器和一个解码器。编码器将声音分解成一个内容因子 fc(t)和一个说话人因子 fs(t);解码器通过这两个因子重构输入语音。


编码器
x(t)
log(X(t))
内容因子提取
v(t)=f(log(X(t)))
说话人因子提取
s(t)=g(log(X(t)))
fc(t)
fs(t)
内容谱重构
V(t)=F(v(t))
说话人谱重构
S(t)=G(s(t))
C(t)
S(t)
+
解码器
log(X'(t))

系统训练

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 fc(t)和说话人因子 fs(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：

   fc(t) = f(log(X(t)))　　（1）

   fs(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 fc 和 fs 尽可能代表说话内容和说话人信息。对于内容因子 fc，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 fs，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 fc(t)和 fs(t)，训练重构模型 F 和 G，将内容因子重构成内容谱 C(t)，将说话人因子重构成声纹谱 S(t)：

   C(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：

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

系统运行

上述模型训练完成以后即可得到一个基于 CS 分解的声码器，其中 f 和 g 为编码器，F 和 G 为解码器。

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

本发明提供了一种新的语音信号分解方法，与传统源-滤波器分解相比，分解方式与任务具有更强的相关性。基于该方法设计的语音声码器在众多应用场景中有重要应用价值，例如可以作为高效的语音编码工具，编码器输出的因子 fs(t)和 fc(t)可用于精简语音编码，适合低带宽网络传输。