清華大學
Tsinghua University

Biweekly Report

姓名：梁玮达

论文题目：UNSUPERVISED AUDIOVISUAL SYNTHESIS VIA EXEMPLAR AUTOENCODERS

日期：2021/8/12

研究背景

- 学习一个低维的表征，忽略说话人的身份信息（然而低维的表征很难捕捉到训练集之外的语音的韵律和周围环境等信息）

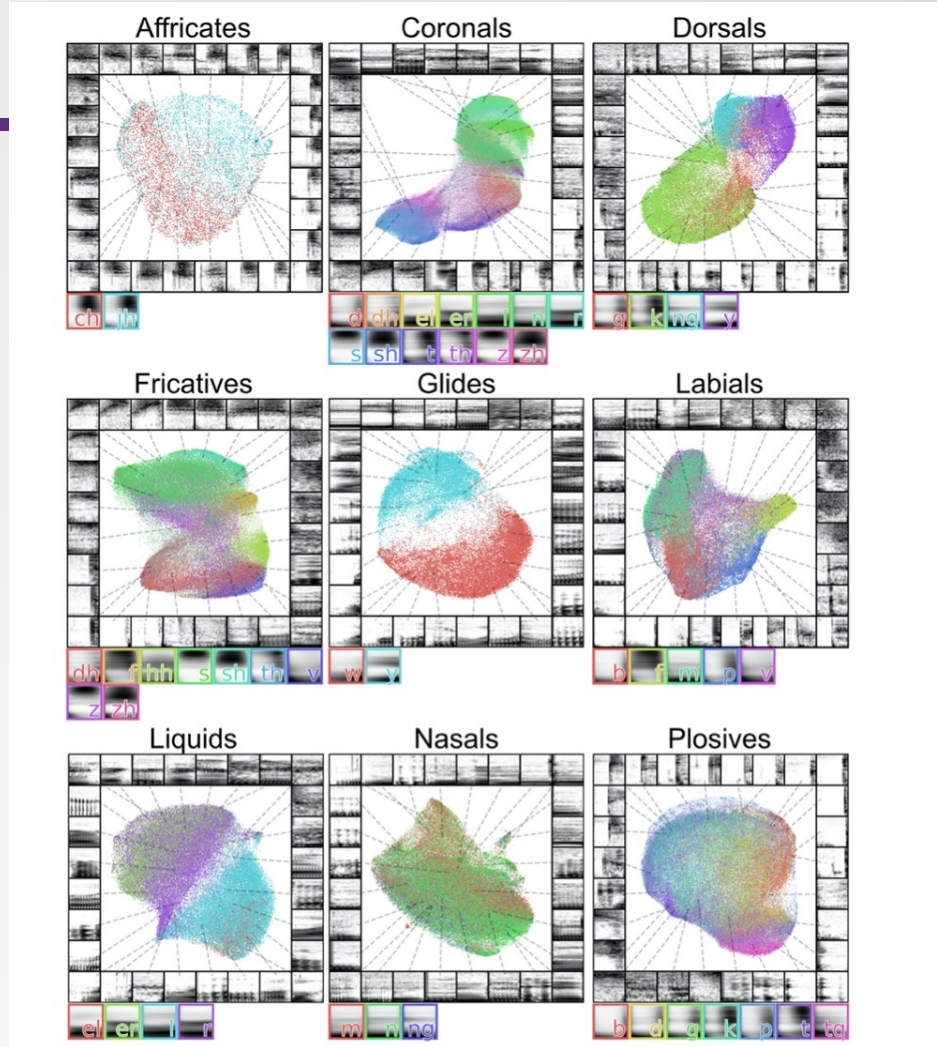
- 学习特定说话人的语音转换（限制了测试阶段的说话人已知）

- 本文将生成模型和特定说话人转换系统结合起来。训练阶段给定一个目标语音和特定的风格，训练一个自编码器；测试阶段，我们验证这个自编码器可以将任何输入语音转变为目标语音。

先验知识

- 语言的音位倾向于在频谱空间中很好地聚集

- 具有足够小瓶颈的自编码器可以将目标语音样本（target）之外的源语音（source）映射到训练目标空间中，保留源语音的内容和目标语音的风格



Tim Sainburg et al, bioRxiv, 2020

研究方法

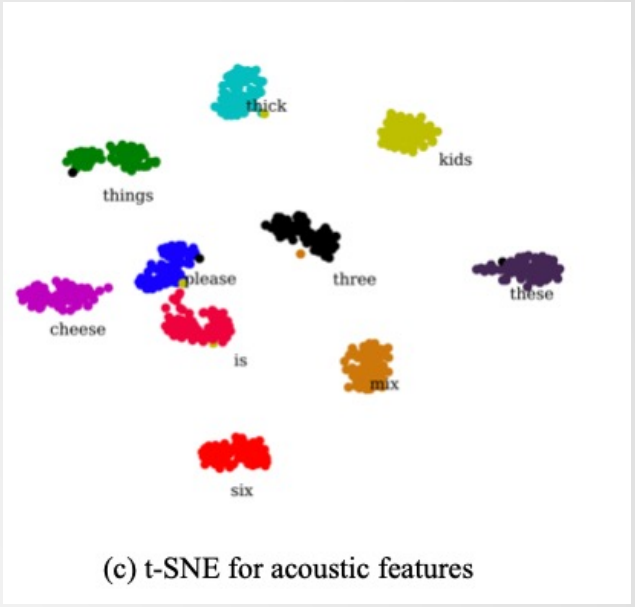
- 自编码器
  - 无监督神经网络
  - 使用瓶颈层重建输入，通常用于降维或特征学习（如去噪自编码器将带噪语音输入映射到训练数据中）

- 本文着重发掘自编码器本身的映射特征，将未知说话人的输入映射到目标说话人的空间中，可以验证此方法适用于非线性自编码器

研究方法

- 样例自编码器(EXEMPLAR AUTOENCODERS)
  - 可以由一个未知说话人的语音作为输入
  - 语音包含两种类型的信息：内容w和风格s
    - x=f(s,w)，f为语音的生成函数
  - 对于不同单词，发声方式不同；然而不同人使用相同种类的发声方式来对同一个词汇发声
  - 给定两种风格s1和s2，有如下结论

$$\text{Error}(f(s_1, w_0), f(s_2, w_0)) \leq \text{Error}(f(s_1, w_0), f(s_2, w)), \forall w \in W, \quad \text{where} \quad s_1, s_2 \in S.$$



(c) t-SNE for acoustic features

Kangle Deng, ICLR, 2021

研究方法

- 样例自编码器可以保存内容的同时转变风格
  - 给定训练样本x，希望学习编码器E和解码器D以最小化重建误差，即

$$\min_{E,D} \sum_i \text{Error}\Big(x_i, D(E(x_i))\Big).$$

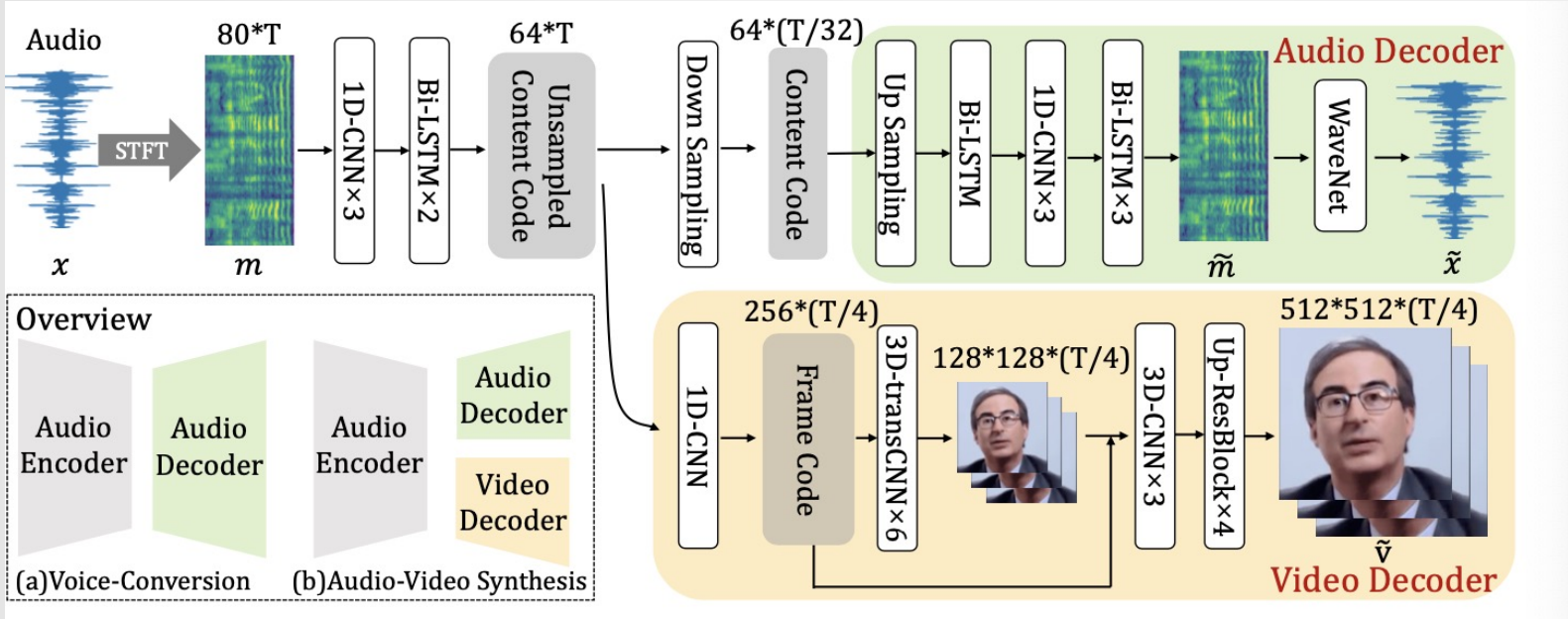
  - 给定一个充分小的瓶颈，线性自编码器将样本点之外的点映射到输入子空间，以最小化重建误差
    - 自编码器捕捉到的是特定数据集的风格信息
    - 瓶颈激活会捕捉到样本特定的内容信息

实验数据介绍

- 训练数据：44.1kHz、单通道
  - CelebAudio（Obama 47:34 ）
  - AISHELL-3 Speech Data
    - 男 4:37
    - 女 28:07
- 测试数据：5～10s
  - AISHELL-3 Speech Data
  - Whisper数据集
  - 论文提供的英文语音

实验模型

- 短时傅里叶映射到梅尔倒频谱，使用自编码器重建频谱，并使用wavenet输出语音信号。



$$\text{Error}_{Audio}(x, \tilde{x}) = \mathbb{E}\|m - \tilde{m}\|_1 + L_{WaveNet}(x, \tilde{x}),$$

清華大學 Tsinghua University

实验结果

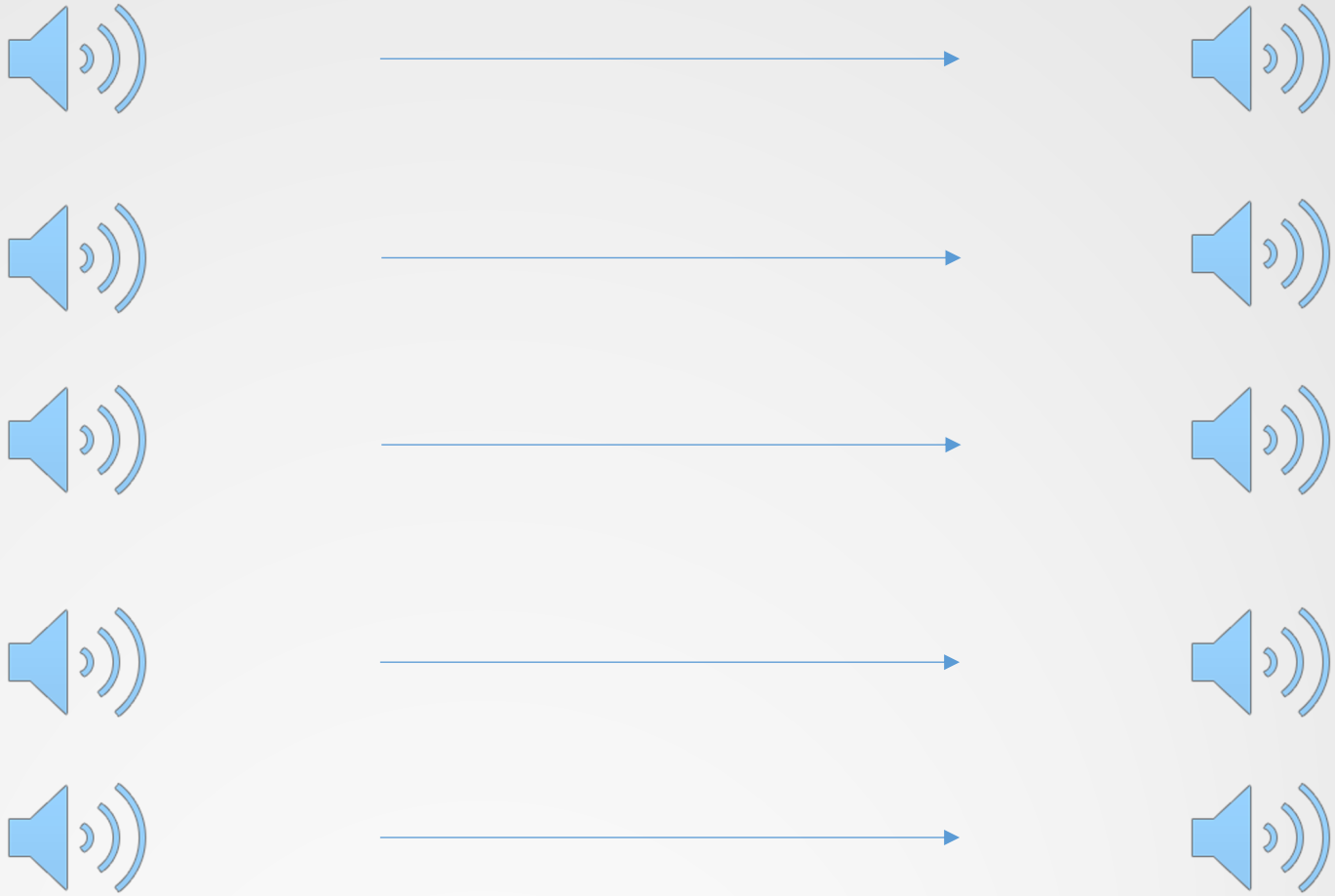
VCTK (Veaux et al., 2016)	Zero-Shot	Extra-Data	SCA (%) ↑	MCD ↓
StarGAN-VC (Kaneko et al., 2019b)	✗	✓	69.5	582.1
VQ-VAE (van den Oord et al., 2017)	✗	✓	69.9	663.4
Chou et al. (Chou et al., 2018)	✗	✓	98.9	**406.2**
Blow (Serrà et al., 2019)	✗	✓	87.4	444.3
Chou et al. (Chou et al., 2019)	✓	✓	57.0	491.1
Auto-VC (Qian et al., 2019)	✓	✓	98.5	408.8
Ours	✓	✗	**99.6**	420.3

实验结果—男（英）

男转男、跨语言

男转男、同语言

男转女、跨语言



实验结果—男（汉）

男转男、同语言

男转男、跨语言

男转女、同语言

实验结果—女（汉）

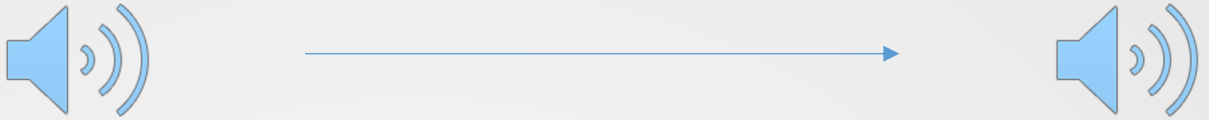
女转女、同语言

女转男、跨语言

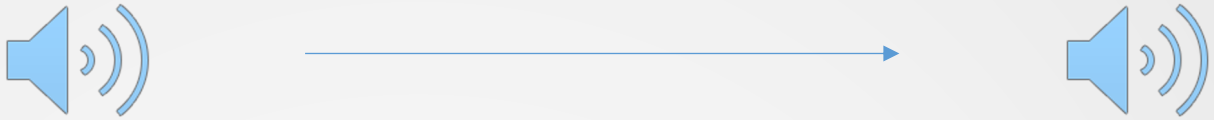
女转男、同语言

实验结果—whisper

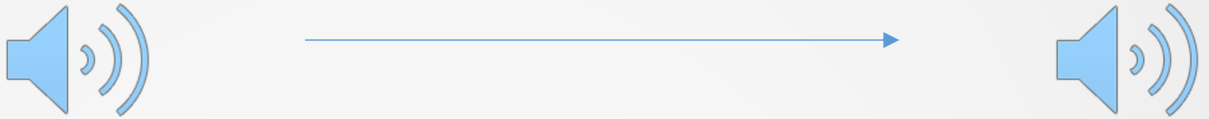
whisper->男、同语言



whisper->男、跨语言



whisper->女、同语言



总结与思考

- 本文主要借助自编码器和瓶颈激活分别可以保留语音的风格信息和内容信息的特点，利用很少的训练数据，实现了特定目标风格的语音转换

- 由于网络可以很好地保留训练语音的风格信息，而忽略掉测试语音的风格信息，因此，可以考虑将原文对语言风格的转换进行扩展，扩展到whisper的语音转换，这也是接下来的实验目标

谢谢聆听
请大家批评指正！