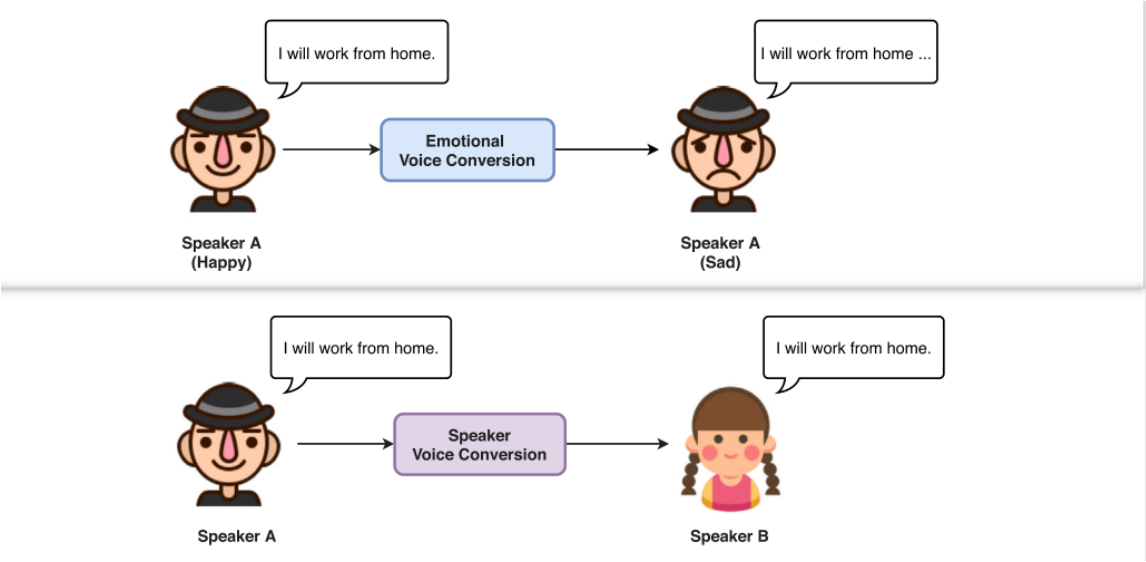
Voice Conversion 一

- 情感语音转换（EVC）是一种旨在将话语的情感状态从一种转换为另一种，同时保留语言信息和说话人身份的技术
  -
    

  - 语音转换的一些方法：
    - 统计方法：Gaussian mixture model (GMM) 高斯混合模型 ； partial least square regression偏最小二乘回归 ； frequency warping 频率扭曲；sparse representation 稀疏表示法
    - 深度学习方法：deep neural network (DNN) 深度神经网络；recurrent neural network (RNN) 递归神经网络；generative adversarial network (GAN) 生成对抗网络 ；sequence-to-sequence model with attention mechanism 具有注意机制的序列到序列模型
    - 非并行训练数据的语音转换技术：
      - domain translation 域翻译； multitask learning 多任务学习；speaker disentanglement 说话者分离；
  - 语音转换与情感语音转换的区别：
    - 语音转换：人们认为，与韵律相关的特征与说话人无关，人们认为，这些特征将从源传递到目标。因此，频谱转换一直是焦点
    - 语音情感转换：情绪本质上是超段和复杂的，涉及频谱和韵律 ，通过逐帧频谱映射来转换情绪是不够
  - 情感语音转换的一些方法：
    - 统计方法：GMM technique GMM技术；sparse representation technique 稀疏表示技术； an incorporated framework of hidden Markov model (HMM) 隐马尔可夫模型（HMM）的综合框架； GMM and fundamental frequency (F0) segment selection method GMM和基频（F0）段选择方法

    - 深度学习方法：DNN 深度神经网络； highway neural network 高速神经网络; deep bi-directional long-short-term memory network (DBLSTM) 深度双向长短期存储网络 ； sequence-to-sequence model 序列对序列模型
    - CycleGAN StarGAN leverage text-to-speech (TTS) automatic speech recognition (ASR)
- 在建模情绪时，通常考虑两个研究问题： （1）如何描述和表示情绪；以及（2）如何建模人类的情绪表达和感知过程
    - 1. 情绪可以用分类或维度表示来表征
        - 六种基本情绪理论（Ekman，1992），其中情绪被分为六个离散的类别，即愤怒、厌恶、恐惧、快乐、悲伤和惊讶
        - 罗素环状模型：由唤醒、效价和支配地位定义。

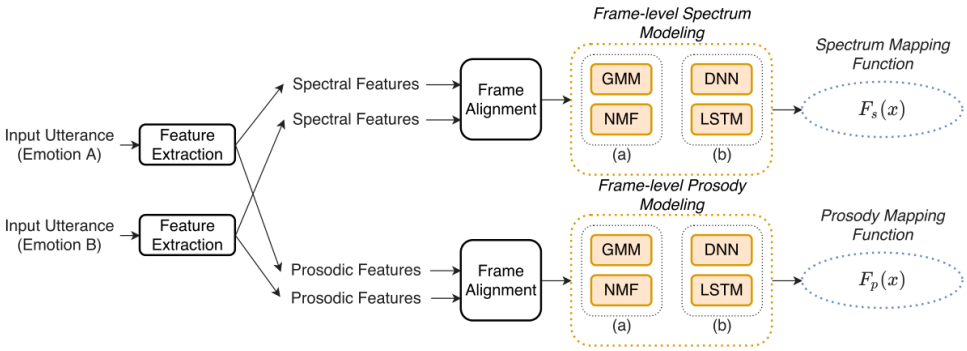
          例如，在 valence-arousal (V-A) 表示中，快乐言语的特征是正价和唤醒值，而悲伤言语的特征是所有负值。 另一方面，生气又可以分为热生气和冷生气
    - 2. 情绪感知和生产：
        - 情绪相关研究中的声学特征：频谱特征、持续时间、F0轮廓和能量包络（spectral features, duration, F0 contour and energy envelope）
- 使用并行数据进行情感语音转换
    - 并行训练数据，即同一说话人的一对相同内容但具有不同情绪的话语
        - 三个步骤：特征提取、帧对齐和特征映射

          动态时间扭曲（DTW）（Müller，2007）和**基于模型的语音识别器对齐**（Ye和Young，2004）或**注意力机制对齐**（Zhang等人，2019b；Tanaka等人，2019）通常用于帧对齐。
    - 特征提取及建模
        - 给定成对的话语，特征提取寻求获得表征语音中情绪的特征。
        - 将得到的成对特征序列对齐以获得帧级对齐

          常用的频谱特征包括梅尔倒谱系数（MCC）、线性预测倒谱系数和线谱频率（LSF）
        - 情感语音转换中，频谱和韵调成分都需要相同程度的关注。通常考虑几个韵律特征，如音高、能量和持续时间
        - F0是一个基本的韵律成分，它描述了从音节到话语的不同持续时间的语调，无论是语言的还是情感的
            - 对F0变体建模的方法：stylization methods multi-level modeling 风格化方法和多层次建模
            - 多级建模方法，连续小波变换（CWT）已被广泛用于建模分层韵律特征：如F0和能量轮廓
    - 特征映射
        - 特征映射旨在捕获源和目标特征之间的关系
        - 统计方法
            - 使用分类和回归树将源语音的基音轮廓分解为层次结构，然后使用GMM和基于回归的聚类方法

        - 基于GMM的情感语音转换框架，以学习频谱和韵律映射
        - 一种基于示例的情感方法，其中并行示例用于编码源语音信号并合成目标语音信号
        - 进一步扩展到统一的基于示例的情感语音转换框架（Ming等人，2015），该框架同时学习频谱特征和基于CWT的F0特征的映射
    - 深度学习方法：
        - DNN, deep belief network , highway neural network , and DBLSTM ； DNN、深度信念网络（DBN）、高速神经网络 和DBLSTM
    - 帧级特征映射并没有明确处理持续时间的映射，但这是韵律的一个重要元素

      1.序列到序列编码器-解码器架构代表了持续时间映射的解决方案。利用注意力机制，神经网络在训练过程中学习特征映射和对齐，并在运行时推断过程中自动预测输出持续时间
      2.编码器-解码器模型（Robinson等人，2019）就是一个例子，其中音调和持续时间被联合建模
    -



- 非并行数据的情感语音转换
    - 非平行数据：指的是多情感话语，这些话语在情感中不共享相同的词汇内容
    - 两种方法：auto-encoder (Kingma and Welling,2013), and (2) CycleGAN (Zhu et al., 2017) method
    - 三种场景：情感领域翻译；情感韵律与语言内容的分离；利用 TTS 或 ASR 系统；
    - 1.情感领域翻译
        - CycleGAN （转换两种情绪的有效性）
            - 包含三个损失，即(1) adversarial loss 对抗性损失, (2) cycle-consistency loss 循环一致性损失, (3) identity mapping loss 身份映射损失
            - 对抗性损失衡量转换后的特征分布和目标特征分布之间的可区分程度；
                - 对于正向映射，定义如下：

$$L_{ADV}(G_{X\to Y}, D_Y, X, Y) = \mathbb{E}_{y\sim P(y)}[D_Y(\mathbf{y})] + \mathbb{E}_{x\sim P(x)}[log(1 - D_Y(G_{X\to Y}(\mathbf{x}))] \quad (12)$$

                  转换数据的分布与目标数据的分布越接近，损失就越小

- 循环一致性损失通过循环转换消除了对并行训练数据的需要；保证输入和输出特征之间的语言一致性

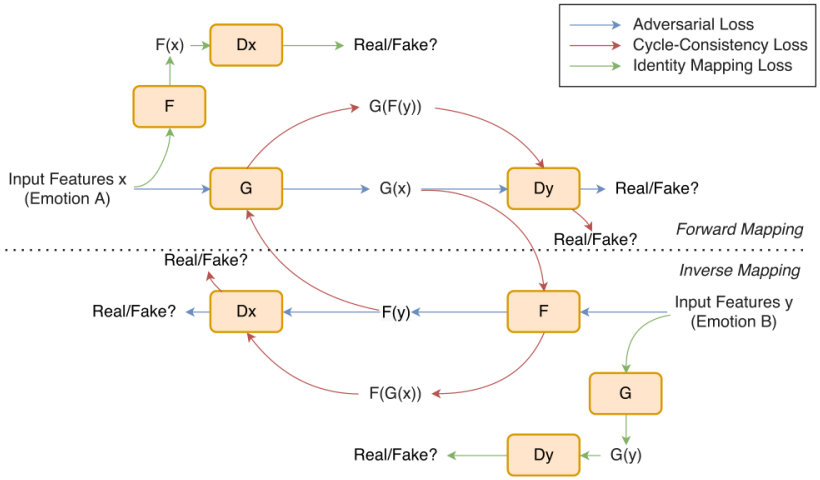
$$L_{CYC}(G_{X\to Y}, G_{Y\to X}) = \mathbb{E}_{\mathbf{x}\sim P(\mathbf{x})}[\|G_{Y\to X}(G_{X\to Y}(\mathbf{x})) - \mathbf{x}\|_1] + \mathbb{E}_{\mathbf{y}\sim P(y)}[\|G_{X\to Y}(G_{Y\to X}(\mathbf{y})) - \mathbf{y}\|_1] \quad (13)$$

L1 范数函数，或最小绝对误差，已知它会产生更清晰的频谱特征
这种损失促使 GX→Y 和 GY →X 通过循环转换找到最佳伪对 (x, y)

- 身份映射损失有助于保留输入和输出之间的语言内容

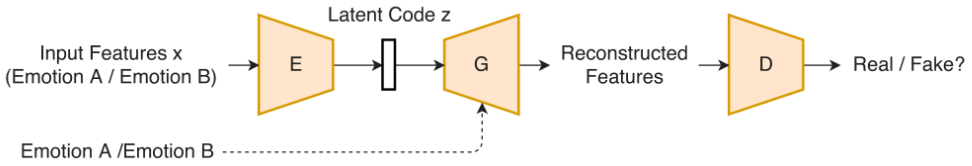
$$L_{ID}(G_{X\to Y}, G_{Y\to X}) = \mathbb{E}_{\mathbf{x}\sim P(x)}[\|G_{Y\to X}(\mathbf{x}) - \mathbf{x}\|] + \mathbb{E}_{\mathbf{y}\sim P(y)}[\|G_{X\to Y}(\mathbf{y}) - \mathbf{y}\|] \quad (14)$$

- StarGAN （多域翻译，多对多说话人语音转换）



- 由一个按类参数化的生成器/判别器对和一个用于验证类正确性的域分类器组成

- 2.情感韵律与语言内容的分离
  - 如果我们能够将情感风格与语言内容分开，我们就可以在不改变语言内容的情况下独立地修改情感风格
  - 自动编码器（Kingma和Welling，2013）是一种常用于语音分离和重构的技术
    - 组成：（1）编码器（2）发生器和（3）鉴别器
    - 编码器逐帧学习生成潜在代码 Z，该潜在代码 Z 应该包含与情感无关的信息，例如语言内容和说话者身份。生成器学习从潜在代码 Z 和情感身份重构输入特征，鉴别器学习区分重建的特征和真实的特征
    -
    -



- 在实践中，我们可以训练情绪分类器，以从语音中获得深层特征表示，即深层情感特征

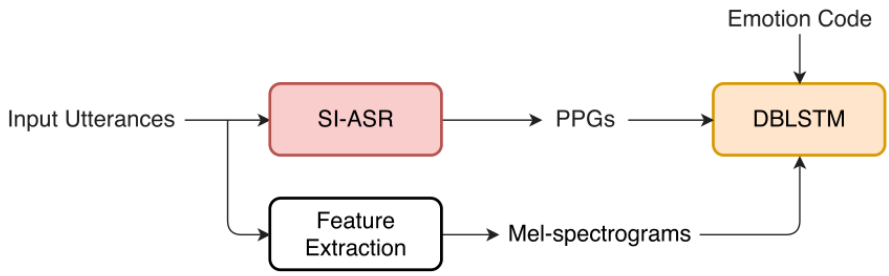
  成功实现： VAE-GAN ； VAW-GAN

- 自动编码器在潜在空间中学习情感不变的潜在代码和情感相关的风格代码

  一对内容编码器和风格编码器被用来从说话内容中分离情感风格

- 3. 利用 TTS 或 ASR 系统
  - 语音转换任务与TTS任务共享许多共同的属性。例如，两者都需要高质量的语音声码器
  - 研究了情绪语音转换和文本到语音之间的多任务学习。在此框架中，训练单个序列到序列模型以优化VC和TTS，其中VC系统受益于TTS在训练期间学习的潜在语音表示。在运行时推断中，给定参考语音，系统可以将输入话语的情感风格从一个转换到另一个
  - 语音后验图（PPG）被用作情感语音转换的辅助输入特征。

    PPG来自自动语音识别系统，该系统被假定为**与说话人无关且与情绪无关**。利用来自PPG的语音信息，所提出的情感语音转换系统对多说话人和多情感语音数据具有更好的泛化能力
  -



利用ASR的基于PPG的情感语音转换框架示例；
使用预训练的说话人独立自动语音识别（SI-ASR）从输入语音中提取PPG。红色框表示预先训练的模型

- 数据集的相关研究
  - lexical variability（词汇变化性）, language variability（语言变化性）, speaker variability（说话人变化性）, confounding factors（混杂因素）, recording environment（录音环境）
  - lexical variability（词汇变化性）
    - 语音合成系统可以受益于具有大量词汇变化的训练数据。然而，这种具有多种情绪话语的数据库并不广泛可用，并且大多数数据库仅包含每种情绪的有限话语。（总结：同一词汇会有多种不同的情感）
  - language variability（语言变化性）
    - 另一个限制是语言表示不平衡。 缺乏多语言情感语音数据库，使得情感语音合成或情感语音转换的相关研究更加困难。（总结：大部分数据集基于英语，其他

语言种类的数据集缺乏）

- speaker variability（说话人变化性）
  - 缺乏说话人的可变性，因此不适合研究说话人独立的情感语音转换（可建立依赖于说话人的情感语音转换，即数据集中已存在的人的情感语音转换）
- confounding factors（混杂因素）
  - 年龄和口音 ； 笑声和叹息
- recording environment（录音环境）
  - 低信噪比（SNR）或低采样率的语音样本可能会损害合成音频的质量

    一些情感语音数据库，其中包含来自各种说话者的大量语音数据。语音数据通常从动态对话、电影或电视节目中收集，并由受试者进行评估
  - 识别系统可能从外部噪声中受益，但是对于合成系统不利
-

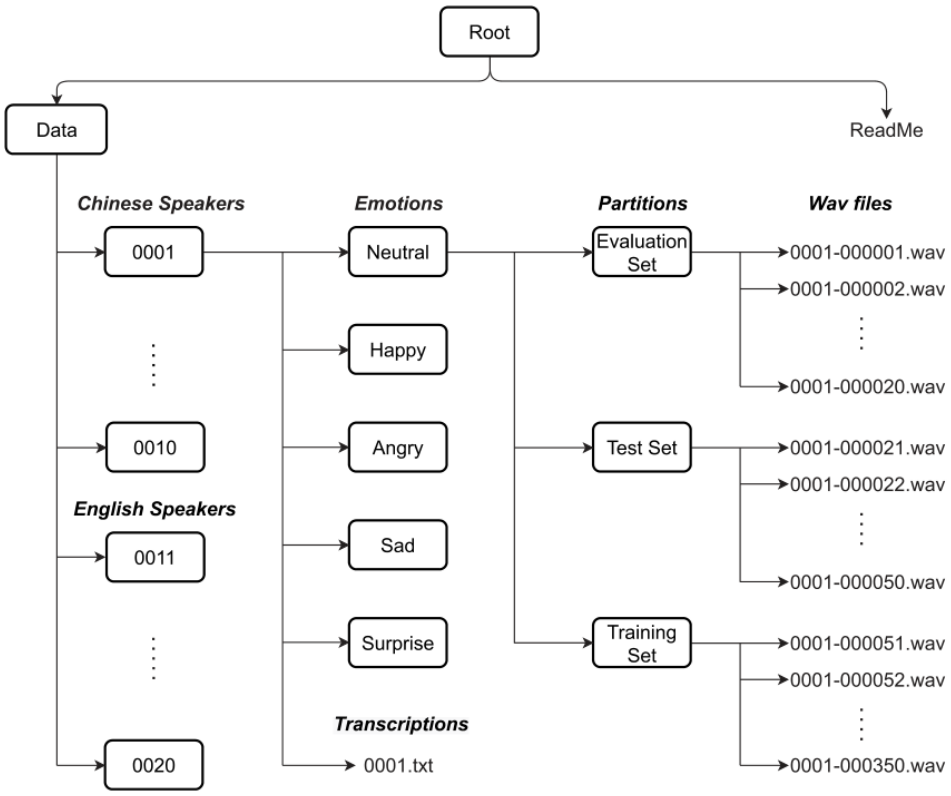
Table 1
An overview of existing emotional speech databases.

Database	Language	Emotions	Size	Modalities	Content type	Remarks
RAVDESS (Livingstone and Russo, 2018)	EN	Neutral, happy, angry, sad, calm, fear, disgust	24 actors (12 male and 12 female) x 2 sentences x 8 emotions x 2 intensities (normal, strong)	Audio/Visual	Acted	Limited lexical variability
CREMA-D (Cao et al., 2014)	EN	Neutral, happy, angry, sad, fear, disgust	91 actors (48 male and 43 female) x 12 sentences x 6 emotions, 1 sentence is expressed with 3 intensities (low, medium, high), others are not specified	Audio/Visual	Acted	Limited lexical variability
EmoDB (Burkhardt et al., 2005)	GE	Neutral, happy, angry, disgust, bored	10 speakers (5 male and 5 female) x 10 sentences x 7 emotions	Audio/Visual	Acted	Limited lexical variability
MSP-IMPROV (Busso et al., 2016)	EN	Neutral, happy, angry, sad	12 actors, 9 h of data	Audio/Visual	Acted/Improvised	Contains external noise, and overlapping speech
IEMOCAP (Busso et al., 2008)	EN	Neutral, happy, angry, sad, frustrate	10 speakers (5 male and 5 female), 12 h of data	Audio/Visual	Acted/Improvised	Contains external noise, and overlapping speech
DES (Engberg et al., 1997)	DA	Neutral, happy, angry, sad, surprise	4 speakers (2 male and 2 female), 10 min of speech	Audio	Acted	Limited lexical variability
CHEAVD (Li et al., 2017)	CN	Neutral, happy, angry, sad, surprise, fear	219 speakers, two hours emotional segments from films and TV shows	Audio/Visual	Improvised	Contains external noise
CASIA (Zhang et al., 2008)	CN	Neutral, happy, angry, sad, surprise, fear	4 speakers (2 male and 2 female) x 500 utterances x 6 emotions	Audio/Visual	Acted	Limited speaker variability Not free to use
AmuS (El Haddad et al., 2017)	EN, FR	Neutral, amuse	4 speakers (3 male and 1 female), about 3 h of data	Audio	Acted	Only single emotion
EmoV-DB (Adigwe et al., 2018)	EN, FR	Neutral, amuse, angry, sleepy, disgust	4 English speakers (2 male and 2 female), 1 French speaker (male), about 5 h of data	Audio	Acted	Contains other non-verbal expressions (laughter, sigh) Incomplete public version Limited speaker variability
SAVEE (Jackson and Haq, 2014)	EN	Neutral, happy, sad, angry, surprise, disgust, fear	4 speakers (all male), 480 utterances in total	Audio/Visual	Acted	Limited speaker variability (all male) Limited lexical variability
VESUS (Sager et al., 2019)	EN	Neutral, happy, angry, sad, fear	10 actors (5 male and 5 female) x 250 distinct phrases	Audio	Acted	Phrases, not complete sentences
JL-Corpus (James et al., 2018)	EN	Neutral, happy, angry, sad, excite	4 speakers (2 male and 2 female) x 5 emotions x 15 sentences x 2 repetitions; 4 speakers (2 male and 2 female) x 6 variants (questions/interrogative, apologetic, encouraging/reassuring, concerned, assertive, anxiety) x 10 sentences x 2 repetitions	Audio	Acted	Limited speaker variability Limited lexical variability
eNTERFACE'05 (Martin et al., 2006)	EN	Angry, disgust, fear, happy, sad, surprise	42 speakers (34 male and 8 female) from 14 nationalities, 1116 video clips in total	Audio/Visual	Improvised	With different accents Limited lexical variability
TESS (Pichora-Fuller and Dupuis, 2020)	EN	Neutral, angry, sad, fear, disgust, surprise, pleased	2 speakers (female) x 200 words x 7 emotions (in the carrier phrase "Say the word _")	Audio	Acted	Limited speaker variability Limited lexical variability
DEMoS (Parada-Cabaleiro et al., 2020)	IT	Neutral, angry, happy, sad, fear, disgust, surprise, guilty	9365 emotional and 332 neutral samples produced by 68 native speakers (23 females, 45 males)	Audio	Improvised	Imbalanced setting for each emotion
EMOVO (Costantini et al., 2014)	IT	Neutral, disgust, fear, anger, joy surprise, sad	6 actors (3 male and 3 female) x 14 sentences x 7 emotions	Audio	Acted	Limited speaker variability Limited lexical variability
VAM Corpus (Grimm et al., 2008)	GE	Valence, activation, dominance	12 h of audio–visual recordings of a German TV talk show	Audio/Visual	Improvised	Contain other non-verbal expressions and external noise
JTES (Takeishi et al., 2016)	JP	Neutral, sad, joy, angry	23 h and 31 min of 50 spoken sentences of 5 emotions acted by 100 speakers (50	Audio	Acted	Not publicly available

- ESD 数据集

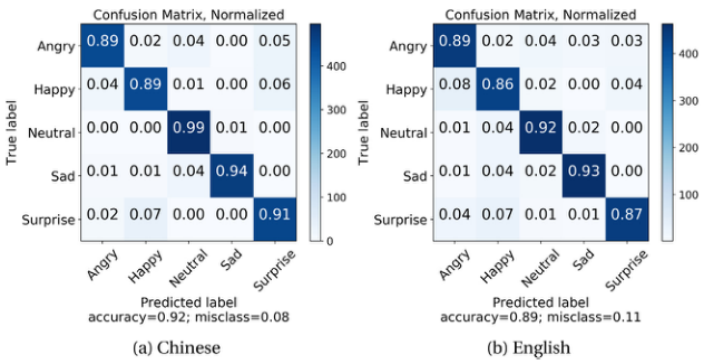
- 分两种类型；acted and improvised（扮演和即兴表演）

  acted：广泛使用在合成系统；缺点：情绪刻板化；
  improvised: 广泛使用在识别系统； 缺点：难以诱导强烈且分化良好的情绪
- 数据库存在问题解决：
  - 1.充足的词汇变化；2.中文英文结合跨语言；3.男女各10个比例均衡；4.年龄限制，标准发言；5.记录设备环境好
- 设计：



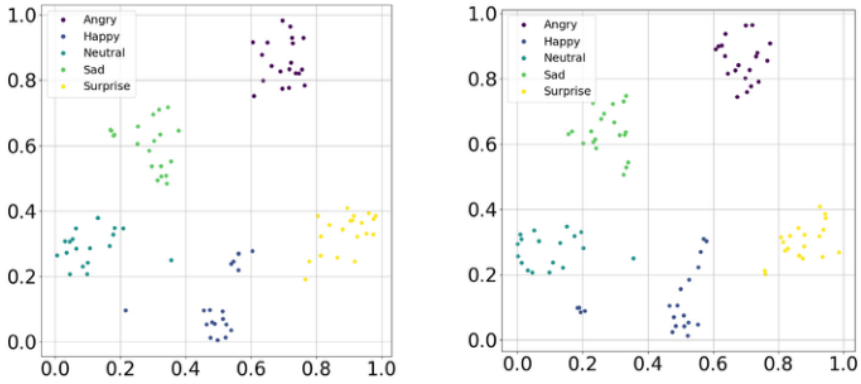
- ESD上的情感语音转换
  - 为了验证ESD数据库中情感表达的质量，开发了一个最先进的说话人独立语音情感识别系统

    SER模型包括一个LSTM层，然后是一个具有256个节点的ReLUactivated fully connected（FC）层。在保持概率为0.5的LSTM层上应用Dropout。最后，将得到的256个特征向量馈送到softmax分类器，该分类器是具有5个输出的FC层
    - 混淆矩阵：中文和英文系统的总体SER准确率分别为92.0%和89.0%



(a) Chinese (b) English

    - 深层情绪特征是SER系统的中间激活，表征情绪状态。



(a) Chinese (b) English

从测试集中为每种情绪随机选择20个话语，并在二维平面中使用分布随机邻居嵌入（t-SNE）算法可视化其深层情感特征

结果表明，从ESD中提取的深层情感特征可以作为言语情感状态的一个极好的描述。

- 为了提供参考基准，我们使用最先进的技术在ESD数据库上进行情感语音转换实验。

(1) CycleGAN EVC：通过CycleGAN网络本身学习成对情感域翻译

(2) VAWGAN-EVC：依靠情感代码来控制转换

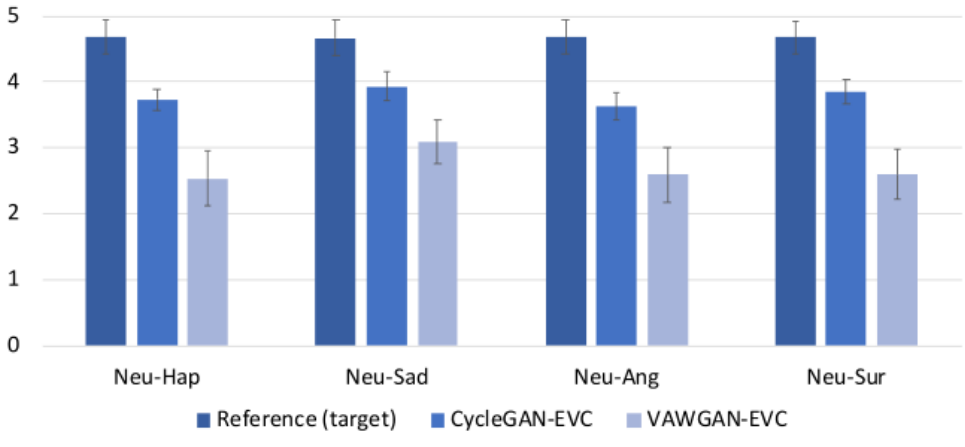
- 客观评价
  - 从频谱中导出24维Mel倒谱系数（MCEP），计算转换后的和目标MCEP之间的Mel倒频失真（MCD）值

Table 5
A comparison of the MCD [dB] values of Zero effort, CycleGAN-EVC (Zhou et al., 2020b), and VAWGAN-EVC (Zhou et al., 2021b) for four emotion conversion pairs. The lower value of MCD indicates the better conversion performance.

MCD [dB]	Neutral-to-Angry	Neutral-to-Happy	Neutral-to-Sad	Neutral-to-Surprise
Zero effort	6.47	6.64	6.22	6.49
CycleGAN-EVC	4.57	4.46	4.32	4.68
VAWGAN-EVC	5.89	5.73	5.31	5.67

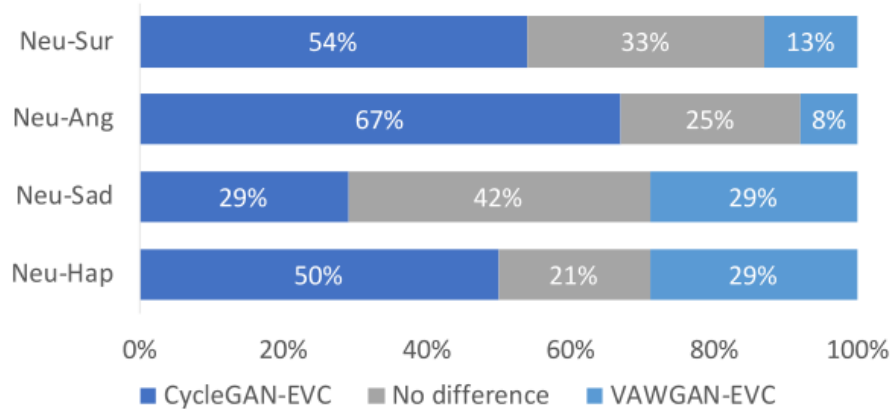
1. Zero effort表示我们直接比较源和目标而不进行任何转换的情况。
2. MCD值均低于Zero effort，结果良好
3.对于所有转换对，CycleGAN EVC始终优于VAWGAN-EVC

- 主观评价
  - （1）主观意见得分 MOS 评估总体语音质量 ： 包括两个方面：（1）如何保存语言信息和说话人身份 （2）转换后的语音的自然度。



对于所有情感转换对，CycleGAN EVC 优于 VAWGAN-EVC

- （2）XAB偏好测试 XAB preference test 评估情感相似性



CycleGAN EVC在Neu Sur、 Neu Ang和Neu Hap方面优于VAWGAN-EVC，并且在Neu Sad方面取得了与VAWGAN-EVC相当的结果

- （3）对比
  - 相同的训练数据量下：CycleGAN EVC： 72小时 VAWGAN-EVC 7小时
  - VAWGAN-EVC 允许多对多情绪的转换，
  - 比CycleGAN EVC更可控，可以扩展为独立于说话人的情绪语音转换

以上内容整理于 幕布文档

二

- 语音转换：
  - 典型的语音转换管道包括语音分析、映射和重构模块，如图1所示，称为分析-映射-重构管道。语音分析器将源说话人的语音信号分解为代表超段信息和段信息的特征，映射模块将其转换为目标说话人的特征，最后由重构模块重新合成时域语音信号

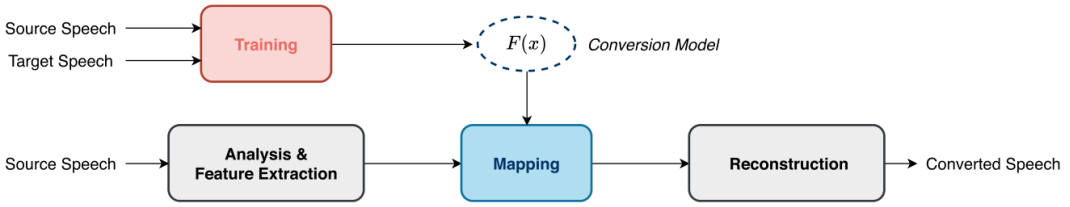


Fig. 1. Typical flow of a voice conversion system. The pink box represents the training of the mapping function, while the blue box applies the mapping function at run-time, in a 3-step pipeline process $\mathcal{Y} = (\mathsf{R} \circ \mathsf{F} \circ \mathsf{A})(\mathcal{X})$.

  - 技术可以以不同的方式分类，例如，基于训练数据的使用 - 并行与非并行，统计建模技术的类型 - 参数与非参数，优化范围 - 帧级别与话语级别， 以及转换的工作流程 - 直接映射与跨语言。
  - 深度学习对语音转换的贡献：
    - 允许映射模块从大量的语音数据中学习，从而极大地提高了语音质量和与目标说话人的相似度
    - 深度学习对语音编码技术产生了深远的影响

      神经声码器是一种神经网络，它学会从声学特征重建音频波形，神经声码器第一次变得可训练和数据驱动

      优秀的声码器： WaveNet ； WaveRNN； WaveGlow和FloWaveNet
    - 深度学习脱离了传统的分析-映射-重构的流程

      神经声码器是可训练的，它可以与映射模块联合训练，甚至与分析模块联合训练，成为端到端解决方案
- 语音转换的典型流程
  - 语音转换系统仅修改与说话人相关的语音特征，如共振峰、基频（F0）、语调、强度和持续时间，同时保留与说话者无关的语音内容
  - 将源和目标语音特征表示为x和y

$$\mathbf{y} = \mathsf{F}(\mathbf{x})$$

F 称为逐帧映射函数

  - 一、语音分析与重构

- 语音分析的目的：是将语音信号分解成某种形式的中间表示，以便有效地操纵或修改语音的声学特性。
- 假设语音信号是根据底层物理模型（如源滤波器模型）生成的，并将语音信号帧表示为一组模型参数，将源说话者的中间表示表示为x，语音分析可以用一个函数来描述：

$$\mathbf{x} = \mathsf{A}(\mathcal{X})$$

- 语音重构可以被看作是语音分析的逆函数，它对修改后的参数进行操作并生成可听语音信号。它与语音分析一起工作
  - 将修改的中间表示表示为y，那目标说话者的重构语音信号即为

$$\mathcal{Y} = \mathsf{R}(\mathbf{y})$$

- 语音转换可以通过三个功能的组合来描述：

$$\begin{aligned}\mathcal{Y} &= (\mathsf{R} \circ \mathsf{F} \circ \mathsf{A})(\mathcal{X}) \\ &= \mathsf{C}(\mathcal{X})\end{aligned}$$

- 一些分析与重构的技术介绍：
  - 1.基于信号的表示
    - 音调同步重叠和相加（PSOLA）
      - 原理：将语音信号分解为重叠的语音段，每个语音段代表语音信号的连续基音周期之一。通过重叠和添加这些具有不同音调周期的语音段，可以重建不同语调的语音信号
      - 优点：直接对时域语音信号进行操作，因此分析和重建不会引入显著的伪影
      - 缺点：清音信号（未发音的 unvoiced speech signal）不是周期性的，时域信号的处理不是直接的
    - 谐波加噪声模型（HNM）
      - 工作假设是语音信号可以表示为谐波分量加上由所谓的最大浊音频率界定的噪声分量。谐波分量被建模为直到最大浊音频率的谐波正弦之和，而噪声分量被建模为由时变自回归滤波器滤波的高斯噪声。由于HNM分解由一些可控参数表示，因此可以轻松修改语音
  - 2.基于模型的表示
    - 原理：假定输入信号可以由参数随时间变化的模型数学表示
    - 源滤波器模型
      - 原理：1.将语音信号表示为源激励的结果，该激励由由喉上声道形状确定的滤波器函数调制

- 2.声码器是一种简短的语音编码器，将语音编码为缓慢变化的控制参数，如线性预测编码和梅尔对数谱近似，这些参数描述了滤波器，并在接收端将语音信号与源信息重新合成
- 3.在语音转换中，通过修改可控参数将来自源说话者的语音信号转换为模拟目标说话者

- STRAIGHT 或“使用加权频谱自适应插值的语音转换和表示
  - 将语音信号分解为： 1) 在时间和频率上没有周期性的平滑频谱图； 2) 使用定点算法估计的基频 (F0) 轮廓； 3) 时间-频率周期图，捕捉噪声的频谱形状及其时间包络

- 3. WaveNet 语音编码器
  - 原理：数据驱动的解决方案，需要大量的训练数据
    - 波形X = x1，x2，…，xN的联合概率可以分解为条件概率的乘积

$$p(\mathcal{X}) = \prod_{n=1}^{N} p(x_n | x_1, x_2, \ldots, x_{n-1})$$

    - WaveNet由许多残差块构成，每个残差块由2×1扩张因果卷积、门控激活函数和1×1卷积组成。通过附加的辅助特征h，WaveNet还可以模拟条件分布p（x|h）

$$p(\mathcal{X}|h) = \prod_{n=1}^{N} p(x_n | x_1, x_2, \ldots, x_{n-1}, h)$$

    - 大多数WaveNet声码器仅覆盖语音重建功能,以语音的一些中间表示作为输入辅助特征，并生成语音波形作为输出。
  - 优点：在声音质量方面明显优于传统参数声码器。它不仅可以学习输入特征和输出波形之间的关系，还可以学习输入特性之间的相互作用
  - 缺点：WaveNet使用自回归（AR）方法对波形采样点的分布进行建模，会导致较高的计算成本

- 4. 神经网络编码器的最新进展
  - 说话人独立WaveRNN的神经声码器 (speaker independent WaveRNN-based neural vocoder)

    它可以从域内和域外频谱图生成类似人类的声音
  - WaveGlow　基于流的网络

    WaveGlow得益于Glow和WaveNet的最佳性能，从而提供快速、高效和高质量的音频合成，而无需自动回归
    使用具有单个成本函数的单个网络来实现，即最大化训练数据的可能性，这使得训练过程简单而稳定
  - Parallel WaveGAN

> 一种无蒸馏快速波形生成方法，通过联合优化多分辨率谱图和对抗性损失函数来训练非自回归WaveNet

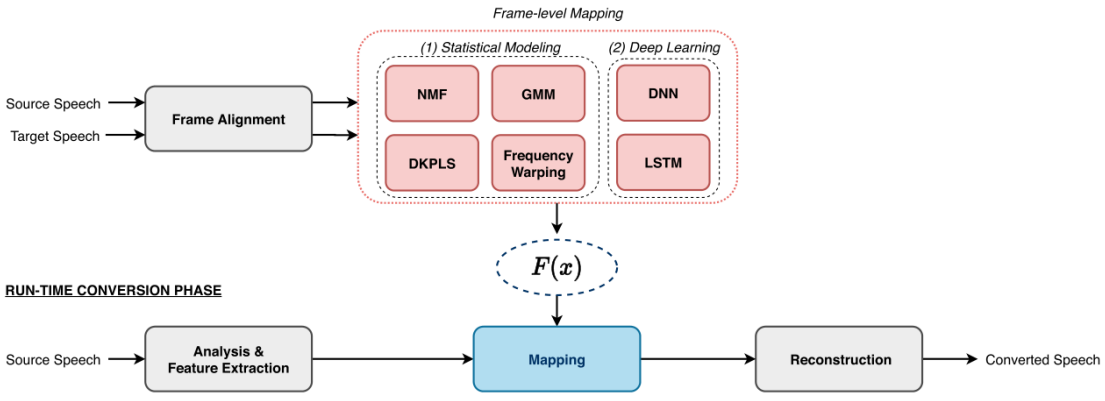
- MelGAN

> 展示基于GAN的方法在语音合成、音乐领域翻译和无条件音乐合成中用于高质量梅尔谱图反演的有效性

- 二、特征提取
  - 为什么进行特征提取：

> 通过语音分析，我们导出了通常包含频谱和韵律分量的语音编码参数，以表示输入语音，可能**不是语音身份转换的最佳参数**
> 语音编码参数被进一步转换为语音特征，称之为特征提取，以**更有效地修改语音转换中的声学特性**

    - 对于频谱分量，特征提取旨在从高维原始光谱中导出低维表示。 一般来说，频谱特征能够很好地代表说话人的个性。 该特征不仅能很好地拟合谱包络，而且能够转换回谱包络
      - 幅度谱可以被扭曲成Mel或Bark频率标度，这对语音转换有感知意义
      - 常用的语音特征包括梅尔倒谱系数（MCEP）、线性预测倒谱系数和线谱频率（LSF）。通常，语音帧由特征向量表示。
      - 短时分析是语音分析的最实用方法，但忽略了上下文。多帧、动态特征和语音段是特征映射中的有效特征
    - 对于韵律分量，特征提取可用于将韵律信号（如基频（F0）、非周期性（AP）和能量轮廓）分解为说话人相关和独立参数
      - 可以继承与说话者无关的韵律模式，同时在特征映射期间转换与说话人相关的韵律。
  - 特征映射
    - 特征映射执行从源说话人到目标说话人的语音特征修改
    - 频谱映射寻求改变声音音色，而韵律转换寻求修改韵律特征，如基频、语调和持续时间
- 并行训练数据语音转换的统计建模
  - 在训练阶段，给定来自源说话者x和目标说话者y的并行训练数据，执行帧对准以对准源语音向量和目标语音向量，以获得配对语音特征向量z = {x，y}



  - Gaussian Mixture Models (高斯混合模型) 参数化技术
    - 原理：

        - 使用高斯混合模型表示来自源和目标说话者的两组频谱包络之间的关系
        - 高斯混合模型是一个连续的参数函数，经过训练可以对光谱映射进行建模
    - 优点：提高了语音质量
    - 缺点：逐帧转换过程导致的具有不适当动态特性的频谱移动，以及转换后频谱的过度平滑
    - 解决：研究了联合密度高斯混合模型（JD-GMM）
        - 以使用最大似然估计联合建模光谱特征序列及其方差，这增加了光谱特征的全局方差
        - JD-GMM 方法包括两个阶段：离线训练和运行时转换阶段。
            - 在训练阶段，采用高斯混合模型（GMM）对成对特征向量序列z = {x, y}的联合概率密度p(z)建模，表示源语音x和目标语音y的联合分布 :为了估计 JD-GMM 的模型参数，使用期望最大化 (EM) 算法 来最大化训练数据的似然性。

$$p(\mathbf{z}) = \sum_{k=1}^{K} w_k^{(z)} \mathcal{N}\left(\mathbf{z}|\mu_k^z, \Sigma_k^{(z)}\right)$$

$$\mu_k^z = \begin{bmatrix} \mu_k^x \\ \mu_k^y \end{bmatrix}, \Sigma_k^{(z)} = \begin{bmatrix} \Sigma_k^{(xx)} & \Sigma_k^{(xy)} \\ \Sigma_k^{(yx)} & \Sigma_k^{(yy)} \end{bmatrix}$$

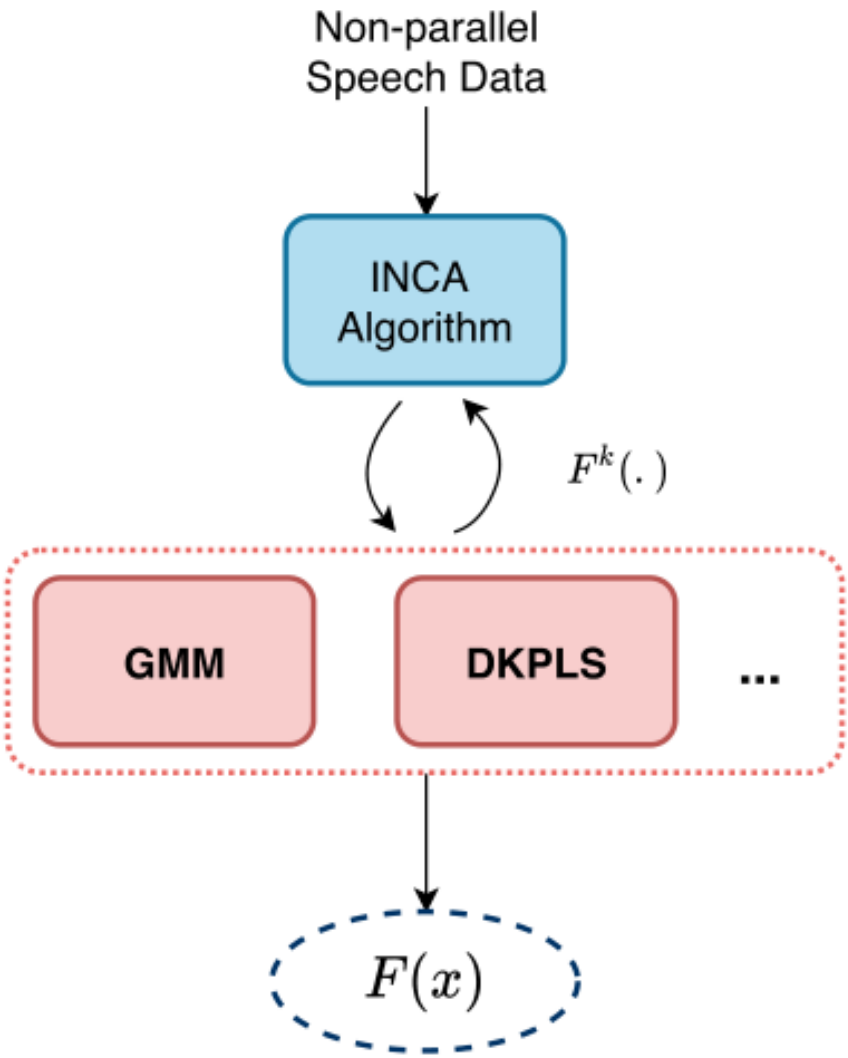
其中K是高斯分量的个数，Wk是每个高斯分量的权重，μ和Σ是第k个高斯分量N的均值向量和协方差矩阵，

            - 在运行时转换阶段，JD-GMM 模型参数用于估计转换函数。
- Dynamic Kernel Partial Least Squares (动态核偏最小二乘) 参数化技术
    - 原理：
        - 考虑语音特征之间的时间相关性
        - 基于源特征的核变换以允许非线性建模，并连接相邻帧以对动态进行建模
        - 非线性变换利用了 GMM 方法所没有的数据的全局属性
    - 优点：在语音质量方面优于 GMM 方法 [113]。 这种方法简单高效，不需要大量调整
- Frequency Warping （频率扭曲） 参数化技术
    - 上述两种技术的缺点：
        - 会出现过度平滑，因为它们使用最小均方误差或最大似然函数作为优化标准。因此，该系统产生表示统计平均值的声学特征，并且无法捕获时间和频谱动态的期望细节。
        - 参数化技术通常采用低维特征，例如梅尔倒谱系数 (MCEP) 或线谱频率 (LSF)，以避免维数混乱

        - 低维特征注定会丢失光谱细节，因为它们具有低分辨率。统计平均和低分辨率特征都会导致输出语音的消音效果
    - 频率变形技术通过频率变形功能直接将高分辨率源频谱转换为目标人的频谱
    - 最新文献中，频率扭曲函数要么由单个参数实现，基于 VTLN 的方；，要么表示为分段线性函数，已经成为主流方案
    - 分段线性扭曲函数的目标：通过最小化频谱距离或最大化转换后的频谱与目标频谱之间的相关性来对齐源频谱与目标光谱之间的一组频率
- Non-Negative Matrix Factorization （非负矩阵分解） 非参数方法
    - 将一个矩阵分解为两个矩阵，一个字典和一个激活矩阵，其性质是所有三个矩阵都没有负元素
    - 优点：基于NMF的技术在训练数据非常有限的语音转换中表现出了有效性
    - 缺点：基于NMF的方法中，目标谱图被构造为样本的线性组合。因此，也可能出现过平滑问题
        - （1）Sparse Representation 稀疏表示
            - 对激活矩阵应用稀疏约束，称为基于样本的稀疏表示
        - (2) Phonetic Sparse Representation 语音稀疏表示
            - 建立在语音子词典和运行时词典选择的基础上
            - 基于样本的稀疏表示语音转换中，多个语音子词典始终优于单个词典
        - （3）Group Sparse Representation 组稀疏表示
            - 提出组稀疏表示：以在统一的数学框架下制定基于样本的稀疏表示和语音稀疏表示
            - 通过组稀疏性正则化，只有与输入特征相关的语音子字典可能在运行时推断时被激活
- 使用非平行训练数据进行语音转换的统计建模
    - INCA Algorithm INCA算法
        - 是最近邻搜索步骤和转换步骤对齐方法的迭代组合，通过在目标声学空间中找到每个源向量的最近邻居来学习映射函数。
        - 主要思想是在上一次最近邻对齐后得到的中间语音x，可以作为下一次迭代时的源语音

$$\mathbf{x}_s^{k+1} = \mathsf{F}^k(\mathbf{x}_s^k)$$

        - 在训练期间，重复优化过程，直到当前中间语音x，足够接近目标语音y



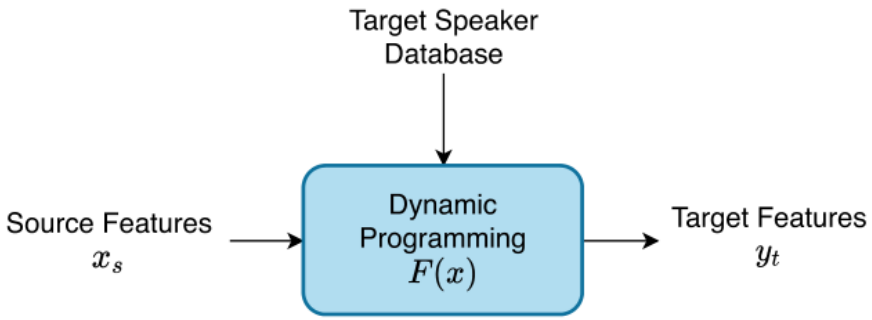
逐帧映射函数的训练是最近邻搜索步骤（INCA对齐）和转换步骤（参数映射函数）之间的迭代过程

- INCA首先使用GMM方法实现，用于语音转换以估计线性映射函数
- INCA通过DKPLS方法实现, 使用INCA对齐算法从源和目标数据集中找到相应的帧，这允许DKPLS回归找到对齐数据集之间的非线性映射

- Unit Selection Algorithm 单元选择算法
  - 单元选择算法广泛应用于自然语音的生成，可以产生高的说话人相似度和语音质量，因为合成的波形是直接从目标说话人单元库中选取
  - 有人建议利用单元选择合成系统从非并行数据中生成训练句子的并行版本，但它需要一个庞大的语音数据库来开发语音单元库
  - 原理：给定来自源说话者的 M 个语音特征向量 X = {x1, x2, ..., xM } 的话语，应用动态规划来找到来自目标说话者的特征向量序列 yi ，其最小化成本函数，

$$\mathbf{Y} = \arg\min_{y} \left(\alpha \sum_{i=1}^{M} d_1(\mathbf{x}_i, \mathbf{y}_i) + (1-\alpha) \sum_{i=2}^{M} d_2(\mathbf{y}_i, \mathbf{y}_{i-1}) \right) \tag{11}$$

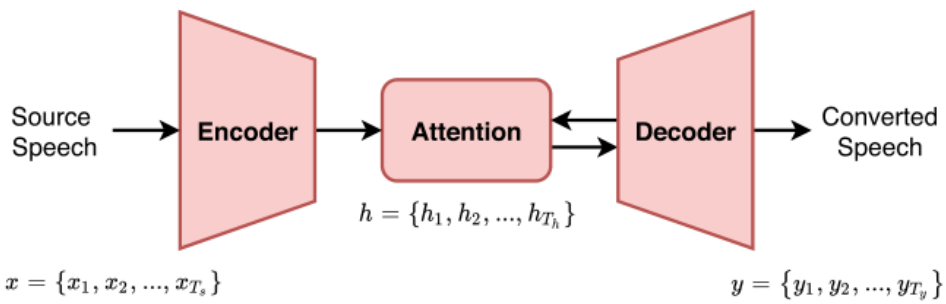
- d1（·）表示源和目标特征向量之间的声学距离，利用声学距离，我们确保从目标说话者中检索到的语音特征接近源的语音特征
- d2（·）是两个目标特征向量间的级联成本。；利用级联成本，我们鼓励从目标说话人数据库中的连续语音帧在多帧片段中一起检索



单元选择算法的运行时推断，该算法不使用参数对映射函数建模，而是直接从目标说话人数据库中搜索输出特征序列，并在话语级别优化输出

- Speaker Modeling Algorithm（说话人建模算法）
  - 文本无关说话人特征的技术很容易用于非并行训练数据，其中说话人可以通过一组参数（如GMM或i向量）建模
    - 使用基于 GMM 的技术预先对参考说话人之间的关系进行建模，并将该关系应用于新说话人
    - 提出了一种执行两种映射的特征语音方法，一种是从源说话人映射到从参考说话人训练的特征语音（或平均语音），另一种是从特征语音映射到目标说话人
  - 启发：在说话人验证中，联合因子分析方法将说话人从其他语音内容中分离出来，以进行有效的说话人验证
    - 从非平行先验数据中估计语音成分和因子负荷，在训练过程中，我们只估计说话者身份因子的低维集合和绑定协方差矩阵
  - 在i-vector说话人空间中执行语音转换，其中i-vector用于将说话人与语言内容分离
    - 原理：无论说话人或语音内容如何，都可以以无监督的方式提取i向量
    - 研究了一种将输入语音的声学特征向i向量空间中的目标语音移动的方法
      - 学习一个函数，将源话语的i向量映射到目标话语的i矢量。通过映射函数，我们能够将源语音逐帧转换为目标语音
- 语音转换的深度学习
  - 一、帧对齐并行数据的深度学习
    - DNN Mapping Function （DNN映射函数）
      - 优点：允许源和目标特征之间的非线性映射，并且对要建模的特征的维度几乎没有限制，其他声学特征（如基频和能量轮廓）的转换也可以类似地进行
        - 1.提出了一种DNN，用于将低维频谱表示（如梅尔倒谱系数（MCEP））从源映射到目标说话人

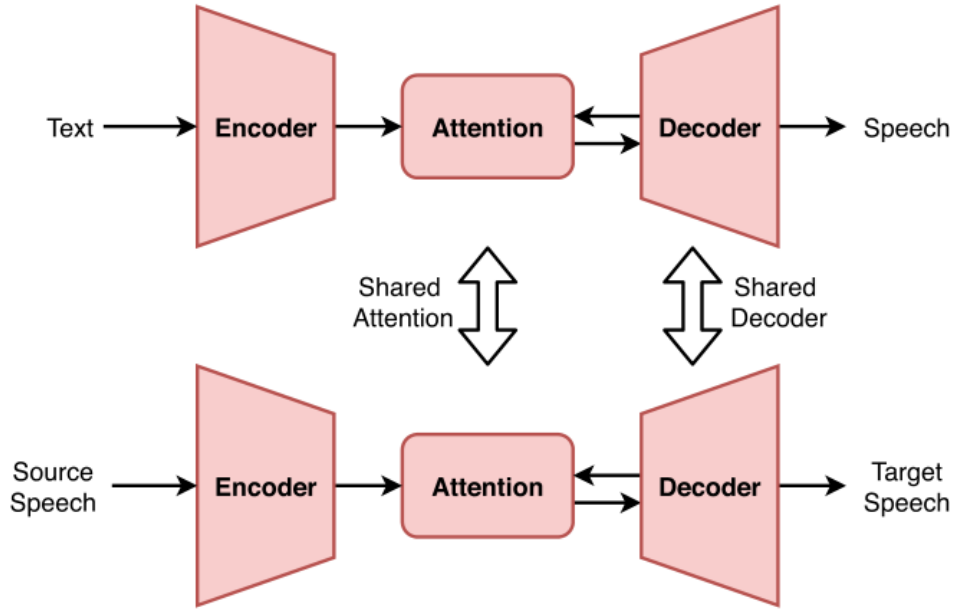
        - 2.使用深度信念网（DBN）从源和目标倒谱系数中提取潜在特征，并使用具有一个隐藏层的神经网络来执行潜在特征之间的转换。
        - 3.研究来自多个说话者的深度自动编码器来进一步推进，以导出语音频谱特征的紧凑表示
    - LSTM Mapping Function （LSTM映射函数）
        - 原因：模拟语音转换中语音帧之间的时间相关性
        - 1.探索循环时间受限玻尔兹曼机 (RTRBM) 的使用，这是一种循环神经网络
            - 优点：长短期记忆 (LSTM) 在序列到序列建模方面的成功激发了 LSTM 在语音转换中的研究，从而提高了语音输出的自然性和连续性
        - 2.通过堆叠BLSTM网络架构的多个隐藏层，提出了一种深度双向LSTM网络（DBLSTM），即使不使用动态特征，它也被证明优于 DNN 语音转换
            - 缺点：
                - 1. 需要来自源说话人和目标说话人的大型语音语料库进行训练，限制了实际应用的范围
                - 2. DNN和LSTM技术在训练数据准备过程中依赖于外部帧对齐器
                - 3. 遵循3步流水线的典型流程，在转换过程中不会改变语音持续时间
- 二、并行数据的带有注意力机制的编解码器
    - 语音转换的研究问题围绕对齐和映射展开
        - 训练期间，更精确的对齐有助于构建更好的映射函数，在运行时推断中，帧级映射范例不会在转换期间改变语音的持续时间
    - 利用注意力机制，神经网络在训练过程中同时学习特征映射和对齐。在运行时推断时，网络根据所学内容自动决定输出持续时间。不需要帧对齐
    - 序列到序列转换网络（SCENT）和AttS2S VC、
        - 1.源语音x = {x1，x2，…，xTs }，编码器网络首先将输入的特征序列转换为隐藏表示，h = {h1, h2, ..., hTh}，在较低的帧速率下 Th < Ts，适合解码器处理。
        - 2.每个解码器时间步，注意力模块通过注意力概率聚合编码器输出并产生上下文向量
        - 3.解码器使用上下文向量逐帧预测输出声学特征
        - 4.设计了一个后置滤波网络来增强转换后的声学特征的准确性，以生成转换后的语音 y = {y1, y2, ..., yTy}



> 在训练期间，注意机制学习源序列和目标序列之间的**映射动态。**
> 在运行时推理时，解码器和注意机制交互以同时执行**映射和对齐**

- 具有注意力的编码器-解码器结构标志着与帧级映射范式的背离
  - 注意力不逐帧执行映射，而是允许解码器关注多个语音帧，并在解码过程中使用组合来预测输出帧
  - 通过注意机制，转换后的语音 Ty 的持续时间通常不同于源语音 Ts 的持续时间，以反映源和目标之间说话风格的差异
  - 代表了一种同时处理频谱和韵律转换的方法

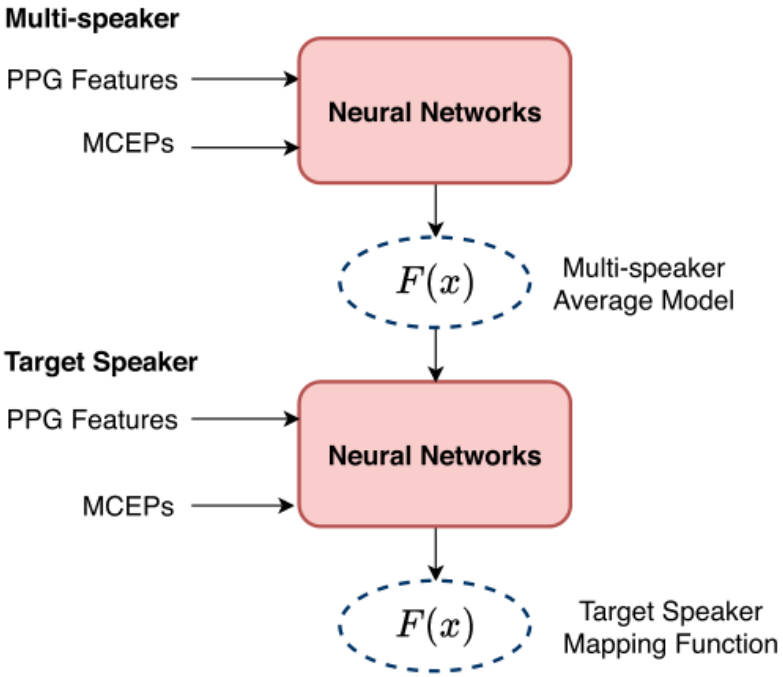
- 三、Beyond Parallel Data of Paired Speakers（除成对说话人之外的并行数据）
  - 背景原因：深度学习已经实现了许多不需要并行数据的语音转换场景
    - Non-parallel data of paired speakers (成对说话人的非并行数据)
    - Leveraging TTS systems （利用文本语音合成系统）
    - Leveraging ASR systems （利用自动语音识别系统）
    - Disentangling speaker from linguistic content （将说话者从语音内容中分离）
  - Non-parallel data of paired speakers (成对说话人的非并行数据)
    - 在语音转换中，将一种语音转换为另一种语音，同时保留语言和韵律内容
      - CycleGAN基于对抗学习的概念，即训练生成模型，以在两个神经网络（称为生成器（G）和鉴别器（D））之间的最小-最大博弈中找到解决方案
  - Leveraging TTS systems （利用文本语音合成系统）
    - 背景原因
      - 在一个大型语音数据库上训练TTS系统，该数据库在给定语言内容的情况下提供高质量的语音重构机制；
      - 其次，TTS系统配备了语音转换所需的质量关注机制
    - 利用 TTS 知识的策略建立在共享注意力知识和/或共享解码器架构的思想之上。
    - 提出了一种语音转换网络的迁移学习技术，可以从 TTS 注意机制派生的语音上下文向量中学习，并与 TTS 系统共享解码器。



该体系结构具有具有双输入和双注意机制的多源序列到序列模型
通过仅将文本作为输入，系统执行语音合成
该系统还可以单独使用语音，或同时使用文本和语音（表示为混合TTS和VC）作为语音转换的输入。

- 提出了一种语音转换系统，称为Cottron，建立在多说话人的Tacotron TTS架构之上
  - 在运行时推理中，预训练的 TTS 系统用于推导源语音的与说话人无关的语言特征。 此过程以输入语音的转录为指导，因此，在运行时推理时需要源语音的文本转录
  - 该系统使用 TTS 编码器来提取与说话人无关的语言特征，或解开说话人的身份。
  - 然后，解码器以注意力对齐的独立于说话人的语言特征为输入，以目标说话人身份为条件，生成目标说话人的语音。
- Leveraging ASR systems （利用自动语音识别系统）
  - 背景原因：大多数ASR系统已经用大量语料库进行了训练。他们已经用不同的方式很好地描述了语音系统
  - 使用ASR模型产生的上下文后验概率序列和序列对序列学习来生成目标语音特征序列
    - 类似于编码器-解码器结构，只是使用语音识别器作为编码器，使用语音合成器作为解码器
  - 另一项研究是通过 ASR 系统引导序列到序列语音转换模型，该系统增加了具有瓶颈特征的输入。
    - 一种端到端的语音到语音序列传感器 Parrotron
      - 学习将具有多种口音和缺陷的任何说话者的语音频谱图转换为单个预定义目标说话者的声音。

        - Parrotron 通过使用辅助 ASR 解码器来预测输出语音的转录本，以编码器潜在表示为条件来实现这一点。
        - Parrotron的多任务训练优化了解码器以生成目标语音，同时限制了潜在表示仅保留语言信息。ASR解码器旨在从语音中分离说话者的身份。
    - 另一种看待语音转换的方式是，语音由两个部分组成，说话者相关部分和说话者无关部分。
        - 平均建模技术：其中构建映射函数以将语音后验图 (PPG) 转换为声学特征。
        - PPG 功能源自 ASR 系统，可以将其视为与说话者无关。
        - 从多说话人、非平行语音数据中训练映射函数。 不需要为每个目标说话人训练一个完整的转换模型。
        - 平均模型可以适应具有少量目标语音的目标。



将 PPG 特征映射到 MCEP 特征以进行语音转换的平均建模方法的训练阶段

- Disentangling speaker from linguistic content （将说话者从语音内容中分离）
    - 语音可以被认为是说话者语音身份和语言内容的组合。能够将说话人从语言内容中分离出来，我们就可以独立于语言内容来改变说话人身份
        - 自动编码器技术
            - 自动编码器学习将其输入重现为输出。 不需要并行训练数据
            - 编码器学习用潜在代码表示输入，解码器学习从潜在代码重建原始输入。
            - 潜在代码可以看作是一个信息瓶颈，一方面，它可以传递必要的信息，例如与说话者无关的语言内容，为了完美重建，另一方面，迫使一些信息被丢弃，例如说话人、噪声和信道信息
    - 基于 VAE 的语音转换框架

        - 解码器通过调节编码器提取的潜在代码和单独的说话人代码来重构话语，这可能是一个 one-hot 向量， i-vector 或者 d-vector
        - 通过在说话人身份上显式调节解码器，编码器被迫从多说话人数据库中捕获潜在代码中与说话人无关的信息
        - 缺点：倾向于生成过度平滑的语音
            - 解决：生成对抗网络 (GAN)
            - 将 GAN 概念纳入 VAE，V AE-GAN 被研究用于非平行训练数据的语音转换 和跨语言语音转换 。 结果表明，VAE-GAN产生的语音比标准 V AE 方法听起来更自然
- 语音转换评估
    - 客观评价：用于频谱的梅尔倒谱失真（MCD），用于韵律的PCC（皮尔逊相关系数）和RMSE（均方根误差）
        - 频谱转换：梅尔倒谱失真（MCD）通常用于测量两个光谱特征之间的差异
            - 它在转换后的和目标Mel倒谱系数或MCEP之间计算，ˆy和y

$$MCD[dB] = \frac{10}{\ln 10}\sqrt{2\sum_{i=1}^{24}(m_{k,i}^{t} - m_{k,i}^{c})^2}$$

            K代表第几帧， i 代表转换的和目标MCEP中的第i个系数
            MCD越低表示性能越好
        - 韵律转换：语音韵律的特点是语音持续时间、能量轮廓和音高轮廓
            - PCC和RMSE用来 测量两个语音话语之间的韵律轮廓或能量轮廓的线性相关性。

$$\rho(F0^{c}, F0^{t}) = \frac{cov(F0^{c}, F0^{t})}{\sigma_{F0^{c}}\sigma_{F0^{t}}}$$

            σF0c和σF0分别是转换后的F0序列（F0c）和目标F0序列的标准差（F0）
            较高的PCC值表示更好的F0转换性能
            - 转换后的F0和对应的目标F0之间的RMSE定义为

$$RMSE = \sqrt{\frac{1}{K}\sum_{k=1}^{K}(F0_{k}^{c} - F0_{k}^{t})^2}$$

            其中F0ck和F0tk分别表示转换后的和目标F0特征
            K是F0序列的长度，或帧的总数
            较低的RMSE值代表更好的F0转换性能
    - 主观评价：平均意见得分（MOS）、偏好测试和最佳最差评分
        - MOS： 1， 2, 3, 4, 5 评分
        - 偏好测试： AB / ABX
        - Best-Worst Scaling (BWS)

        - 原因：人们对介于两者之间的任何事物的偏好可能是模糊和不准确的
        - 方法：每次只向听众呈现几个随机选择的选项
- 使用深度学习方法进行评估
    - 语音质量感知评估 (PESQ) 是一项 ITU-T 推荐标准，被广泛用作行业标准。 它提供客观的语音质量评估，预测人类感知的语音质量
    - 缺点：PESQ公式要求存在参考语音
    - 解决：
        - Quality-Net 一种端到端模型，用于预测PESQ评级，这是人类评级的代理
        - 一种基于CNN的自然度预测因子，用于预测人类MOS评分
        - MOSNet 基于深度神经网络 可以学习预测人类MOS评级

            MOSNet标志着最近在自动感知质量评估方面取得的进展，该评估是免费和开源的

以上内容整理于 幕布文档