语音转换调研（2021-2022-arXiv）

VC 模型

1. A Comparative Study of Self-supervised Speech Representation Based Voice Conversion：基于自监督Speech Representation 的VC，先识别再合成，基于S3PRL toolkit。
2. End-to-End Voice Conversion with Information Perturbation：采用信息扰动来去除源语音中与说话人相关的信息，端到端。
3. End-to-End Zero-Shot Voice Style Transfer with Location-Variable Convolutions：提出了基于位置变量卷积的语音转换 (LVC-VC)（端到端zero-shot语音转换）。
4. VQMIVC: Vector Quantization and Mutual Information-Based Unsupervised Speech Representation Disentanglement for One-shot Voice Conversion：one-shot VC，采用矢量量化进行内容编码，在训练期间引入互信息 (MI) 作为相关度量。
5. Nvc-net: End-to-end adversarial voice conversion：基于对抗训练的many-to-many VC，能直接生成音频而不需要vocoder（E2E）。
6. Non- parallel any-to-many voice conversion by replacing speaker statistics：在 VC 系统中提出了一个统计替换层来直接修改隐藏状态以获得目标说话者的统计信息。
7. Assem-VC: Realistic Voice Conversion by Assembling Modern Speech Synthesis Techniques：组合各种技巧实现 any to many 的VC，有源码，SOAT模型。
8. Towards end-to-end F0 voice conversion based on Dual-GAN with convolutional wavelet kernels：使用预训练的F0网络，端到端模型。
9. Global rhythm style transfer without text transcriptions：在不依赖任何文本的情况下从语音中分离出全局韵律风格，基于自编码器。
10. Conversion with Disentangled Universal Linguistic Representations：跨语言，把 PPG 替换成 Universal Linguistic Representations。

GAN

1. StarGAN-VC+ASR: StarGAN-based Non-Parallel Voice Conversion Regularized by Automatic Speech Recognition：使用ASR辅助改进StarGAN-VC。
2. Subband-based Generative Adversarial Network for Voice Conversion：SGAN-VC，在unseen的数据上效果 超过了StarGANv2-VC+ASR ，显式的利于每个不同子带的特征，分别转换每个子带。
3. Glow-WaveGAN 2: High-quality Zero-shot Text-to-speech Synthesis and Any-to-any Voice Conversion：Zero-shot TTS + VC，从Glow-WaveGAN改进而来。

ASR

1. Voicy: Zero-shot non-parallel voice conversion in noisy re- verberant environments：噪声环境下的zero-shot语音转换。
2. HiFi-VC: High Quality ASR-Based Voice Conversion：基于ASR特征、音高追踪、波形预测模型。
3. Tvqvc: Transformer based vector quantized variational autoencoder with ctc loss for voice conversion：基于CTC损失和VQ来获取高层次的语言信息。
4. StarGAN-VC+ASR: StarGAN-based Non-Parallel Voice Conversion Regularized by Automatic Speech Recognition：使用ASR（基于 GMM 的 phoneme model）改进了StarGAN-VC。

VAE

1. Robust Disentangled Variational Speech Representation Learning for Zero-shot Voice Conversion：zero-shot，disentangle 的方法，基于VAE
2. Conditional deep hierarchical variational autoencoder for voice conversion：基于深度分层 VAE 的 VC，非自回归解码器
3. Enhancing Zero-Shot Many to Many Voice Conversion with Self-Attention VAE：添加自注意力，zero-shot VC，many to many。
4. Towards Improved Zero-shot Voice Conversion with Conditional DSVAE：zero-shot VC，提出了条件 DSVAE。
5. Contrastively Disentangled Sequential Variational Autoencoder：C-DSVAE，提取和分离潜在空间中的静态（时不变）和动态（时变）因子，提出了一种新的变分下界。

TTS

1. GlowVC: Mel-spectrogram space disentangling model for language-independent text-free voice conversion：基于flow的多语言、多说话人的VC，基于Glow-TTS构建，有两个版本：GlowVC-conditional and GlowVC-explicit。
2. Glow-WaveGAN 2: High-quality Zero-shot Text-to-speech Synthesis and Any-to-any Voice Conversion：Zero-shot TTS + VC，从Glow-WaveGAN改进而来。
3. Transfer learning from speech synthesis to voice conversion with non-parallel training data：TTS-VC，借鉴于TTS，训练TTS模型，把输入从文本变成语音，decoder一样。

噪声

1. Learning Noise-independent Speech Representation for High-quality Voice Conversion for Noisy Target Speakers
2. An Evaluation of Three-Stage Voice Conversion Framework for Noisy and Reverberant Conditions

3. Voicy: Zero-shot non-parallel voice conversion in noisy re- verberant environments：噪声环境下的zero-shot语音转换。

流式 or low latency

1. Streaming non-autoregressive model for any-to-many voice conversion：基于完全非自回归模型，声学模型（基于Transformer）和声码器（HiFi-GAN）都是流式的，PPGs based，any to any。
2. FastS2S-VC: Streaming Non-Autoregressive Sequence-to-Sequence Voice Conversion：非自回归，S2S-VC，速度快了70-100倍，基于 teacher-student learning framework，many to many。
3. AC-VC: Non-parallel Low Latency Phonetic Posteriorgrams Based Voice Conversion：基于PPG，any to many，57.5 ms 的 look-ahead，AC是指Almost Causal。

歌声转换

1. A Hierarchical Speaker Representation Framework for One-shot Singing Voice Conversion：提出SVC 分层说话人表示框架，可以捕捉不同粒度的说话人特征
2. Multi-Singer: Fast Multi-Singer Singing Voice Vocoder With A Large-Scale Corpus：发布数据集、提出基于GAN的多歌声转换声码器，样例在这。

其他

1. Speak Like a Dog: Human to Non-human creature Voice Conversion：从人到狗的语音转换