# DISENTANGLING CONTENT AND FINE-GRAINED PROSODY INFORMATION VIA HYBRID ASR BOTTLENECK FEATURES FOR VOICE CONVERSION

*Xintao Zhao[1,†], Feng Liu[2], Changhe Song[1], Zhiyong Wu[1,3,‡], Shiyin Kang[2,‡], Deyi Tuo[2], Helen Meng[1,3]*

[1] Tsinghua-CUHK Joint Research Center for Media Sciences, Technologies and Systems,
Shenzhen International Graduate School, Tsinghua University, Shenzhen, China
[2] Huya Inc, Guangzhou, China
[3] Department of Systems Engineering and Engineering Management,
The Chinese University of Hong Kong, Hong Kong SAR, China

{zxt20, sch19}@mails.tsinghua.edu.cn, {liufeng1, kangshiyin, tuodeyi}@huya.com, {zywu, hmmeng}@se.cuhk.edu.hk

## ABSTRACT

Non-parallel data voice conversion (VC) have achieved considerable breakthroughs recently through introducing bottleneck features (BNFs) extracted by the automatic speech recognition(ASR) model. However, selection of BNFs have a significant impact on VC result. For example, when extracting BNFs from ASR trained with Cross Entropy loss (CE-BNFs) and feeding into neural network to train a VC system, the timbre similarity of converted speech is significantly degraded. If BNFs are extracted from ASR trained using Connectionist Temporal Classification loss (CTC-BNFs), the naturalness of the converted speech may decrease. This phenomenon is caused by the difference of information contained in BNFs. In this paper, we proposed an any-to-one VC method using hybrid bottleneck features extracted from CTC-BNFs and CE-BNFs to complement each other advantages. Gradient reversal layer and instance normalization were used to extract prosody information from CE-BNFs and content information from CTC-BNFs. Auto-regressive decoder and Hifi-GAN vocoder were used to generate high-quality waveform. Experimental results show that our proposed method achieves higher similarity, naturalness, quality than baseline method and reveals the differences between the information contained in CE-BNFs and CTC-BNFs as well as the influence they have on the converted speech.

***Index Terms***— voice conversion, hybrid bottleneck features, cross entropy, connectionist temporal classification, disentangling

## 1. INTRODUCTION

Voice conversion (VC) aims at converting a speech uttered by source speaker as if it was uttered by a target speaker, while maintaining speech content. Conventional VC systems require paired training utterances which are expensive in practice. Non-parallel VC is a more challenging but more practical task in applications. The training utterances are only recorded by the target speaker, which is easier for collection than parallel data. A common way to achieve non-parallel VC is disentangling the linguistic and non-linguistic information carried by the source and target utterances respectively, then combining them and synthesizing the converted utterance. In order to generate waveform with high quality and similarity, it is crucial to keep these information accurate and disentangled as much as possible. Some
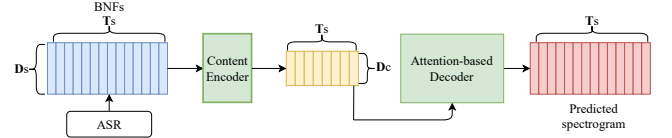
---

† Work performed while interning at Huya Inc.
‡ Corresponding author.



**Fig. 1**. Baseline model architecture

methods tried to disentangle content information with timbre and pitch in an unsupervised manner [1–3]. However, due to the difficulty of the unsupervised learning, the quality and intelligibility of converted waveform inevitably degrades.

Recognition-synthesis based VC is another representative approach of the non-parallel VC and has been shown to achieve noticeable performance improvement [4–7]. Such methods provide more robust linguistic representations based on the knowledge from an automatic speech recognition (ASR) model trained with huge amount of data. Therefore, information contained in the representations is considered to be highly related to the content information and disentangled with speaker-related information to a certain degree [7]. An intermediate embedding is extracted from a certain layer of ASR model, either phonetic posteriorgrams (PPGs) [4] from the last layer regardless of preformed extractor structure, or bottleneck features (BNFs) [8] from the penultimate layer with well-designed dimension, and then fed into a VC system. BNFs based VC framework, as shown in Fig. 1 is proven to be more robust when facing recognition error than PPGs in practice.

Prior approaches in recognition-synthesis based VC extract BNFs from ASR model trained with cross entropy loss (CE loss) on the time-aligned training data (CE-ASR) and illustrate noticeable performance [9, 10]. However, we find that the BNFs extracted by CE-ASR (CE-BNFs) contain extra prosody information and timbre information that may leak into the voice conversion task [5]. Though the converted waveform could get fine-grained prosody variation that is similar to the source utterance and has high naturalness, the similarity to target speaker and speech quality degrade significantly due to the preserved redundant source timbre information in CE-BNFs. A PPGs-based approach based on the ASR model trained with connectionist temporal classification loss (CTC-ASR) [11] is proposed recently [6]. By only preserving the non-blank-token frames of the PPGs based on the peaky behavior of the CTC models, a pure
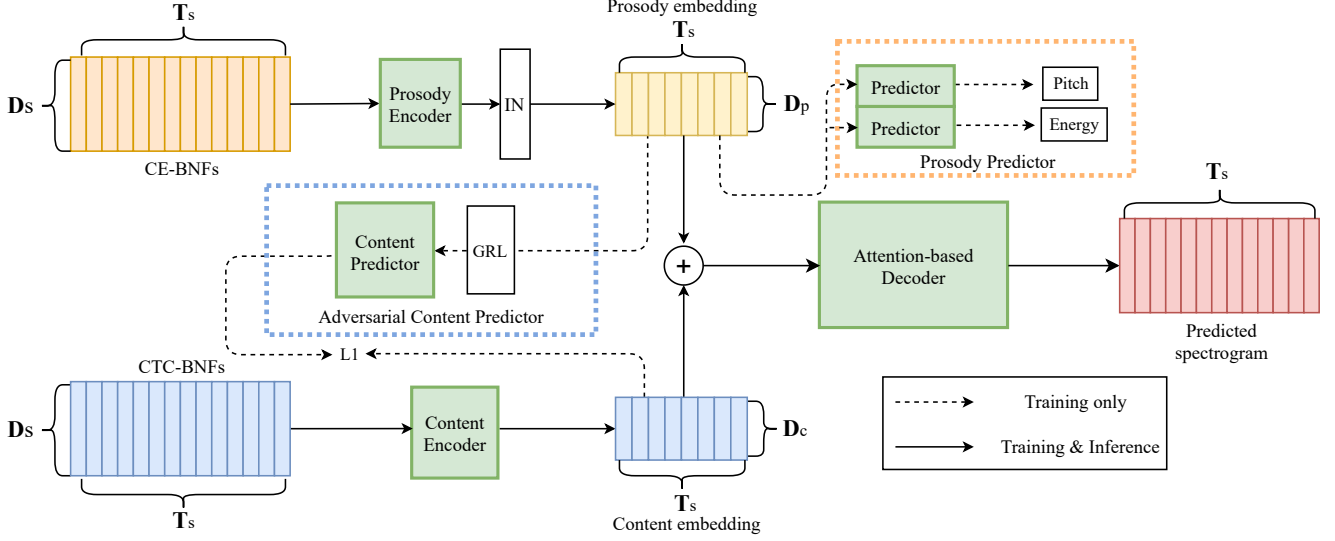
**Fig. 2**. Proposed model architecture.

linguistic representation is extracted. Considering that for utterances with same content but spoken by different speakers, the diversity of converted speeches may degrade if only non-blank-token frames are fed into VC system. Inspired by the influence of PPGs with different training losses on VC mentioned in [6], we studied the impact of BNFs with different ASR training losses to VC and find that the BNFs extracted from CTC-ASR (CTC-BNFs) preserve less non-linguistic information indicating better disentangling characteristics. Therefore, the converted waveform has limited naturalness but higher target speaker similarity and quality due to the elimination of prosody and timbre information from source speaker. It is easy to find that CE-BNFs and CTC-BNFs can complement each other.

In this paper, to retain the good disentangling characteristics of CTC-BNFs, and meanwhile preserve the fine-grained prosody information contained in CE-BNFs, we proposed an any-to-one VC system using hybrid BNFs extracted by two ASR systems with CTC and CE loss respectively. The proposed approach attempted to extract prosody information from CE-BNFs, and content information from CTC-BNFs. Compared to mel-spectrogram, CE-BNFs has more stable data distribution on diversified input waveform and is more suitable for extracting prosody embedding in our any-to-one VC task. On the other hand, the proposed model tried to extract content embedding from CTC-BNFs rather than CTC-PPGs because BNFs are more robust to recognition error. The contributions of our work include:

1) We proposed a VC method by virtue of hybrid BNFs from CTC-BNFs and CE-BNFs. This method leverages the complementary information of the two BNFs and achieved high similarity, naturalness and quality of the converted speech.

2) We demonstrated the information differences contained in CTC-BNFs and CE-BNFs through appropriate experiments.

3) An adversarial-prediction structure was proposed to disentangle content and prosody information.

4) Experimental results revealed that BNFs are more robust to ASR recognition errors than PPGs for VC.

## 2. METHODOLOGY

The proposed any-to-one VC system, shown in Fig.2, consists of five components: (i) A prosody encoder followed by instance normalization (IN) [12] to encode CE-BNFs into prosody embedding. (ii) A content encoder to encode CTC-BNFs into content embedding. (iii) An adversarial content predictor with gradient reversal layer (GRL) [13] to reduce the information overlapping between prosody embedding and content embedding. (iv) A prosody predictor consists of two sub-predictors to predict pitch and energy so as to force the prosody encoder work correctly. (v) An auto-regressive decoder with location-sensitive attention [14] for high-quality spectrogram generation. Although (iii) and (iv) increase the complexity at training stage, the increase in the number of parameters is acceptable as we do not need to load their parameters at inference stage.

### 2.1. ASR system

The very recent conformer [15] based ASR acoustic model achieved state-of-the-art recognition performance due to the combination of Transformer and CNN models. The conformer models are good at capturing content-based global interactions as well as local features. We trained two ASR models with conformer encoder but different loss functions: CE loss and CTC loss. We extract 256-dims BNFs from the last hidden layer of conformer encoder.

### 2.2. Encoder-Decoder model

As shown in Fig.2, the proposed conversion model is an encoder-decoder architecture, following the setting of Tacotron2 [16]. The input CTC-BNFs were passed through a content encoder firstly. The content encoder consists of a stack of 2 fully-connected layers and a set of 1D convolutional layers, followed by a multi-layer highway network [17] and a bidirectional GRU layer. The outputs of content encoder were then concatenated with prosody embedding extracted by prosody encoder. The concatenated content and prosody embeddings were further fed into the auto-regressive decoder to generate high-quality spectrogram.

## 2.3. Prosody information extraction

Prosody encoder consists of a Conv1d layer and a GroupNorm layer followed by a BLSTM structure. The prosody embedding has a significant narrow dimension compared to CE-BNFs to restrict the information passed by. To further eliminate the timbre information, an IN module was introduced here. The output of prosody encoder was fed into adversarial content predictor, prosody predictor and decoder at the same time. Each sub-predictor in prosody predictor consists of 3 fully-connected layers followed by GeLU activation [18] and layer normalization [19]. The prosody loss is defined as:

$$F_{prosody} = IN(E_{prosody}(BNF_{ce})) \tag{1}$$

$$\hat{E}, \hat{P} = P_{prosody}(F_{prosody}) \tag{2}$$

$$L_{energy} = \|E - \hat{E}\|_1 \tag{3}$$

$$L_{pitch} = \|P - \hat{P}\|_1 \tag{4}$$

$$L_{prosody} = \lambda_p * L_{pitch} + \lambda_e * L_{energy} \tag{5}$$

where $BNF_{ce}$ is the input CE-BNFs, $F_{prosody}$ is the prosody embedding, $E_{prosody}$ and $P_{prosody}$ correspond to prosody encoder and predictor. $P$, $\hat{P}$ are the ground truth and predicted pitch, $E$, $\hat{E}$ are the ground truth and predicted energy respectively. $L_{prosody}$ is the total prosody loss, where we simply set $\lambda_p$ and $\lambda_e$ to 1.

## 2.4. Adversarial learning for information disentanglement

An adversarial content predictor network inspired by [3] was designed to reduce the overlap of information between prosody embedding and content embedding. The network consists of a GRL and a prediction head layer [20] as content predictor. The output of the adversarial content predictor network has the same size as the content embedding extracted by the content encoder. The gradient of the adversarial network is reversed by GRL before backward propagated to the prosody encoder. The adversarial loss is defined as:

$$F_{content} = E_{content}(BNF_{ctc}) \tag{6}$$

$$\hat{F}_{content} = P_{content}(GRL(F_{prosody})) \tag{7}$$

$$L_{adv} = \left\|\hat{F}_{content} - F_{content}\right\|_1 \tag{8}$$

where $BNF_{ctc}$ is the input CTC-BNFs of content encoder $E_{content}$, $F_{content}$ and $\hat{F}_{content}$ correspond to the content embedding generated by content encoder and the output of adversarial content predictor. $P_{content}$ and $GRL$ means content predictor and gradient reversal layer, respectively. The optimization of $L_{adv}$ forces prosody embedding to contain as little overlapped content information as possible because of the reversed gradient.

## 2.5. Voice conversion loss

We computed the reconstruction loss between the input and reconstructed mel spectrograms. To speed up the training process, we generated a diagonal matrix and calculate the alignment loss between the matrix and alignment of decoder as $L_{align}$. The reconstruction loss and the final objective loss functions are given as follows:

$$L_{rec} = \|S - \hat{S}\|_1 \tag{9}$$

$$Loss = \lambda_{adv} * L_{adv} + \lambda_{rec} * L_{rec} + \lambda_a * L_{align} + L_{prosody} \tag{10}$$

where $S$ and $\hat{S}$ are the mel spectrograms of the ground truth and reconstructed speeches respectively, and $\lambda$s are the loss weights for all the losses.

Finally, Hifi-GAN [21] was used as the vocoder to get high fidelity waveform and fast decoding speed.
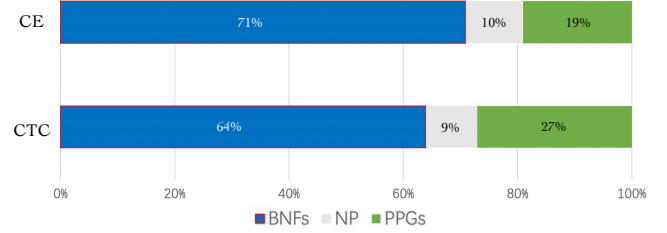


**Fig. 3**. Results of the ABX-test for **Naturalness**. BNFs and PPGs correspond to the input of the baseline VC model. CE and CTC indicate the loss which is used in training ASR system. NP means no preference

# 3. EXPERIMENT

## 3.1. Training setup

The voice conversion experiments were performed on a homegrown Mandarin corpus uttered by a female speaker. The original corpus duration is 2 hours in total. The training data is augmented by changing the speaking rate from 0.9 to 1.5 times to enhance the prosody diversity. ASR systems were trained with nearly 3000 hours Mandarin corpus uttered in live streaming scene, which may contain noisy and accented speech. We extract 256-dim BNFs and 331-dim PPGs from ASR model separately in our experiments.

We evaluated the speaker information contained in BNFs extracted from different ASR systems via a speaker classification network. We introduced a ResNet-32 [22] based neural network here. It would randomly select 2s bottleneck frames from each utterance as input and predict which speaker these frames belong to. Nearly 300 hours of Mandarin corpus from AISHELL-1 [23], AISHELL-3 [24], THCHS [25] and VC corpus were used in training. The corpus contains 381 female speakers and 252 male speakers. 20 sentences were randomly chosen from each speaker as the validation set.
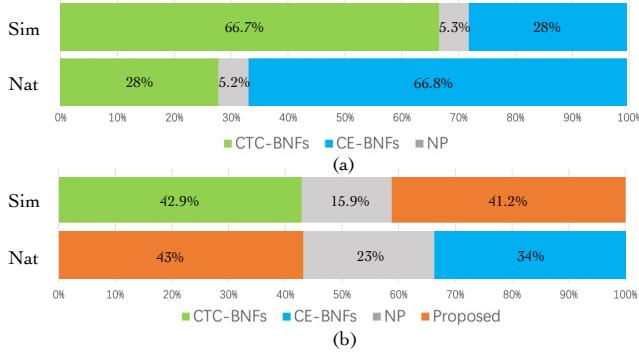
All the audios were down-sampled to 16KHz. Mel spectrograms were computed through a short time Fourier transform (STFT) using a 50ms frame size, 10ms frame hop and a Hann window function. Fbank features were used in ASR system with 25ms frame size and 10ms frame hop. Our ASR systems were based on the implement of WeNet, an E2E speech recognition toolkit[1]. A pretrained HiFi-GAN [21] vocoder was used to synthesize the audios from the mel spectrogram. Baseline and proposed systems were evaluated under the same settings, otherwise mentioned. Demo page is available[2].

## 3.2. Subjective evaluation

In order to show the impact of different linguistic features on the VC system, we conducted an ABX test to compare the naturalness or the timbre similarity of the converted speech. "A" and "B" were converted speech samples obtained with different methods and "X" was a real speech sample obtained from a source speaker in naturalness evaluation which has correct and natural prosody, or a real speech sample from the target speaker in similarity evaluation. Each listener was presented 20 utterances for each ABX pair. The result in Fig.3 shows that BNFs are more robust than PPGs in terms of recognition errors, which lead to significant improvement of naturalness no matter which loss function are used in training the ASR system.

---

[1]https://github.com/wenet-e2e/wenet
[2]https://thuhcsi.github.io/icassp2022-hybrid-bottleneck-vc/

**Fig. 4**. Result of the ABX test for **Naturalness** and **Similarity**. Where **\*-BNFs** means baseline model with different BNFs as input.

As the results in Fig.4 (a) reveals, CE-BNFs get lower similarity but higher naturalness, whereas CTC-BNFs get the adverse result. Fig.4 (b) shows the proposed method combined the merits of the two types of BNFs and generated speech with high similarity and naturalness.

**Table 1**. Speaker classification accuracy and Cosine similarity calculated by CTC-BNFs and CE-BNFs. Cosine similarity stands for the average cosine similarity of all neighbouring BNF vectors.

|  | CTC-BNFs | CE-BNFs |
|---|---|---|
| Cosine similarity | 0.791 | **0.769** |
| Speaker classification accuracy | **32.2%** | 68.89% |

### 3.3. Objective evaluation

Table 1 shows the information difference between CTC-BNFs and CE-BNFs. *Speaker cls acc* indicates the accuracy of our speaker classification task using different BNFs as input. It is shown that CE-BNFs have significantly high accuracy compared to CTC-BNFs, this indicates CE-BNFs preserve much more timbre information. We then calculated average *cosine similarity* between neighboring BNF vectors. Provided that the neighbouring BNF vectors mostly belong to the same phoneme, lower cosine similarity means more prosodic variation and richer prosody information contained in BNFs.

Table 2 shows the Mel-cepstral distortion (MCD), root-mean-square error (RMSE) in F0 and energy. MCD was used to measure how close the converted is to the target speech. M2VoC parallel data[3] were used for MCD calculation. We adopted Pysptk tools to extract the MCCs and dynamic time warping (DTW) to align the target and converted speech parameters. Lower value indicates higher similarity. It can be observed that CE-BNFs-VC system gets higher MCD than CTC-BNFs-VC system because the redundant timbre information contained in CE-BNFs limited similarity and quality. Energy and F0 RMSEs were used to measure the naturalness of the converted waveform. Both F0 and energy sequences were performed min-max normalization before calculation. We calculated RMSE between converted and source waveform because source waveform was uttered by real speakers with fine-grained prosody and high naturalness. It is found that CE-BNFs-VC system achieves lower energy and F0 RMSEs because the extra fine-grained prosody information preserved contributes to high naturalness. Overall, the proposed

method outperforms two baseline methods in all metrics.

**Table 2**. MCD, Energy RMSE and F0 RMSE calculated between the waveforms generated by different systems. Energy and F0 RMSEs were calculated between **source** and **converted** waveform while MCD was calculated between **target** and **converted** waveform.

|  | MCD | Energy RMSE | F0 RMSE |
|---|---|---|---|
| CTC-BNFs-VC system | 7.494 | 0.306 | 0.395 |
| CE-BNFs-VC system | 7.760 | 0.298 | 0.365 |
| Proposed-VC system | **7.453** | **0.286** | **0.358** |

### 3.4. Ablation study

To understand how much each component contributes to the model, mean opinion scores (MOS) tests were conducted to evaluate the speech quality, prosody naturalness and timbre similarity of the converted waveform. Each listener was presented 25 utterances for each experiment, and was asked to give a 5-point scale score with 1 point interval. Table 3 shows that all components have significant impact on our proposed model. Model without IN leads to significantly performance degradation in similarity and quality because IN could eliminate timbre information of source speaker. The model without prosody predictor leads to degradation in naturalness. Adversarial learning also plays an important role in disentanglement. If we replace CE-BNFs by mel-spectrogram like [26], due to the variation of source waveform, the prosody extraction will fail in our experiment.

**Table 3**. The MOS with 95% confidence intervals shows the impact of different modules in network on naturalness, similarity and quality. **Mel-Prosody encoder** row is the result gained by feeding mel-spectrogram into prosody encoder to extract prosody embedding.

|  | Naturalness | Similarity | Quality |
|---|---|---|---|
| Proposed | 3.65(-0.00) | 3.75(-0.00) | 3.96(-0.00) |
| w/o IN | 3.16(-0.49) | 2.89(**-0.86**) | 2.90(**-1.06**) |
| w/o Prosody predictor | 3.02(**-0.65**) | 3.19(-0.56) | 3.23(-0.73) |
| w/o Adv predictor | 3.20(-0.45) | 3.36(-0.39) | 3.58(-0.38) |
| Mel-Prosody encoder | 2.17(-1.45) | 2.31(-1.44) | 2.76(-1.20) |

## 4. CONCLUSION

In this paper, we proposed an any-to-one voice conversion model using hybrid BNFs. The proposed method retains the good disentangling characteristics of CTC-BNFs while preserving the fine-grained prosody information contained in CE-BNFs, leading to superior performance than the baseline models in naturalness and similarity. The experiments demonstrate that CE-BNFs contains much more timbre and prosody information than CTC-BNFs. CE-BNFs is a better choice when attempts are made to extract prosody information from various source waveform than the mel-spectrogram. Other than that, our ABX test also proves that BNFs features are more robust choice than PPGs in terms of recognition error. In the future, we will focus on any-to-many VC task and try to extract prosody information and content information separately from two different waveforms to further control the prosody of any content.

---

[3]http://challenge.ai.iqiyi.com/detail?raceId=5fb2688224954e0b48431fe0

# 5. REFERENCES

[1] Chin-Cheng Hsu, Hsin-Te Hwang, Yi-Chiao Wu, Yu Tsao, and Hsin-Min Wang, "Voice conversion from unaligned corpora using variational autoencoding wasserstein generative adversarial networks," *arXiv preprint arXiv:1704.00849*, 2017.

[2] Kaizhi Qian, Yang Zhang, Shiyu Chang, Mark Hasegawa-Johnson, and David Cox, "Unsupervised speech decomposition via triple information bottleneck," in *International Conference on Machine Learning*. PMLR, 2020, pp. 7836–7846.

[3] Jie Wang, Jingbei Li, Xintao Zhao, Zhiyong Wu, and Helen Meng, "Adversarially learning disentangled speech representations for robust multi-factor voice conversion," *arXiv preprint arXiv:2102.00184*, 2021.

[4] Lifa Sun, Kun Li, Hao Wang, Shiyin Kang, and Helen Meng, "Phonetic posteriorgrams for many-to-one voice conversion without parallel data training," in *2016 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2016, pp. 1–6.

[5] Shaojin Ding, Guanlong Zhao, and Ricardo Gutierrez-Osuna, "Improving the speaker identity of non-parallel many-to-many voice conversion with adversarial speaker recognition.," in *INTERSPEECH*, 2020, pp. 776–780.

[6] Ying Zhang, Hao Che, and Xiaorui Wang, "Non-parallel sequence-to-sequence voice conversion for arbitrary speakers," in *2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. IEEE, 2021, pp. 1–5.

[7] Songxiang Liu, Yuewen Cao, Disong Wang, Xixin Wu, Xunying Liu, and Helen Meng, "Any-to-many voice conversion with location-relative sequence-to-sequence modeling," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1717–1728, 2021.

[8] Zhichao Wang, Wenshuo Ge, Xiong Wang, Shan Yang, Wendong Gan, Haitao Chen, Hai Li, Lei Xie, and Xiulin Li, "Accent and speaker disentanglement in many-to-many voice conversion," in *2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. IEEE, 2021, pp. 1–5.

[9] Jing-Xuan Zhang, Zhen-Hua Ling, and Li-Rong Dai, "Non-parallel sequence-to-sequence voice conversion with disentangled linguistic and speaker representations," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 540–552, 2019.

[10] Jing-Xuan Zhang, Zhen-Hua Ling, Li-Juan Liu, Yuan Jiang, and Li-Rong Dai, "Sequence-to-sequence acoustic modeling for voice conversion," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 3, pp. 631–644, 2019.

[11] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 369–376.

[12] Ju-chieh Chou, Cheng-chieh Yeh, and Hung-yi Lee, "One-shot voice conversion by separating speaker and content representations with instance normalization," *arXiv preprint arXiv:1904.05742*, 2019.

[13] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky, "Domain-adversarial training of neural networks," *The journal of machine learning research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[14] Jan Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio, "Attention-based models for speech recognition," *arXiv preprint arXiv:1506.07503*, 2015.

[15] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al., "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.

[16] Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al., "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4779–4783.

[17] Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber, "Highway networks," *arXiv preprint arXiv:1505.00387*, 2015.

[18] Dan Hendrycks and Kevin Gimpel, "Gaussian error linear units (gelus)," *arXiv preprint arXiv:1606.08415*, 2016.

[19] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

[20] Andy T Liu, Shu-wen Yang, Po-Han Chi, Po-chun Hsu, and Hung-yi Lee, "Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6419–6423.

[21] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *arXiv preprint arXiv:2010.05646*, 2020.

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[23] Hui Bu, Jiayu Du, Xingyu Na, and Hao Zheng Bengu Wu, "Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline," in *Oriental COCOSDA 2017*, 2017, p. Submitted.

[24] Yao Shi, Hui Bu, Xin Xu, Shaoji Zhang, and Ming Li, "Aishell-3: A multi-speaker mandarin tts corpus and the baselines," 2015.

[25] Dong Wang and Xuewei Zhang, "THCHS-30 : A free chinese speech corpus," *CoRR*, vol. abs/1512.01882, 2015.

[26] Sri Karlapati, Alexis Moinet, Arnaud Joly, Viacheslav Klimkov, Daniel Sáez-Trigueros, and Thomas Drugman, "Copycat: Many-to-many fine-grained prosody transfer for neural text-to-speech," *arXiv preprint arXiv:2004.14617*, 2020.