Interspeech 2024
1-5 September 2024, Kos, Greece

USD-AC: Unsupervised Speech Disentanglement for Accent Conversion

Jen-Hung Huang, Wei-Tsung Lee*, Chung-Hsien Wu*

Department of Computer Science and Information Engineering,
National Cheng Kung University, Tainan, Taiwan
tony22316@gmail.com, p76104582@gs.ncku.edu.tw, chwu@csie.ncku.edu.tw

Abstract

This study proposes USD-AC, an innovative **U**nsupervised **S**peech **D**isentanglement **A**ccent **C**onversion that does not require parallel data and text transcription for training, solving challenges such as limited labeled data and generalizability issues. USD-AC, grounded in speech decomposition, aims to separate accent features from linguistic content, enhancing its adaptability across various accent conversion tasks. It utilizes a pre-trained ASR model to extract linguistic content and incorporates accent embedding for accent representation. Adversarial training effectively disentangles accent information from other attributes, boosting conversion performance. USD-AC achieves remarkable outcomes for known speakers and accents and exhibits exceptional generalization to unseen speakers, accents, and content. Through experimental comparison, USD-AC based on unsupervised learning has shown superiority and generalization ability compared to supervised learning methods.
Index Terms: accent conversion, unsupervised learning, voice conversion, speech decomposition

1. Introduction

Even within the realm of the same language, phonetic nuances diverge due to the influence of varied backgrounds and life experiences. This occurrence emerges from the intricate interplay of geographical, societal, and cultural factors upon individuals, giving rise to distinctions in vocabulary, grammar, and usage of language. These distinctions extend even to the realm of phonetic attributes such as phonemes, tonality, emphasis, rhythm, and more, where they manifest in distinctive and vivid characteristics.

However, conspicuous variations in accent impede the convenience of daily communication and compromise the precision of human-machine interactions. Thus, devising methods to mitigate the impact of accents is of paramount importance. These methods encompass auditory training to familiarize oneself with accents or pronunciation exercises to modulate one's accent. The technology of accent conversion [1] (AC) aims to transform spoken sentences from one accent into another while preserving the speaker's voice characteristics and linguistic content. Since the converted speech retains the same timbre features as the original speaker, it serves as a valuable reference for accent learning [2]. Leveraging accent conversion technology enhances the facilitation of acquiring the nuances of diverse accents, and may even find applications in dubbing for audiovisual entertainment [3] or serve as an auxiliary system for interpersonal interactions.

*denotes equal contribution.

In contrast to the mere alteration of a speaker's timbre through voice conversion (VC), the impact of accents extends beyond tonal and rhythmic modifications; it also leads to changes in the pronunciation of phonemes. To address the modification of phonetic disparities between diverse accents, prior accent conversion methodologies have necessitated the use of parallel corpora during synthesis [4, 5, 6]. Alternatively, models have been developed based on the content-specific distinctions among accents to train accent translators [7, 8, 9]. The approach of employing parallel corpora during synthesis requires the pre-collection of corresponding target-accent sentences, which poses limitations on its applicability.

Non-parallel accent conversion, employing speaker-independent automatic speech recognition (SI-ASR), extracts phoneme posterior grammars (PPG) or bottleneck features as content features. These features are then fed into a neural network-based translator to model the linguistic content distinctions between native and non-native speech, thus forming the basis for transformation. To facilitate supervised learning for training the translator, the collection of annotated accent data is requisite. This may encompass transcriptions at either word or phoneme levels, phoneme manipulation details including substitution, deletion, and insertion errors, or parallel datasets. Subsequently, a synthesis architecture akin to text-to-speech (TTS) is employed to convert the linguistic content features into mel spectrograms. Lastly, these mel spectrograms are transformed into waveforms via pre-trained vocoders, such as WORLD [10], WaveNet [11], and HiFi-GAN [12].

Despite being more practical than parallel accent conversion, the training process of supervised non-parallel accent conversion heavily relies on accurate text transcriptions or parallel data, presenting certain challenges. Primarily, obtaining accurate annotations for a substantial amount of speech data entails a significant investment of both time and financial resources. Furthermore, this method encounters limitations when expanding the accent conversion system to accommodate novel accents or languages; acquiring parallel data for each new dataset becomes impractical. Additionally, there exists a risk of encountering training biases, including potential inadequacies in speaker generalization. These issues collectively impede the scalability of accent conversion approaches.

To address this issue, we drew inspiration from recent concepts in speech decomposition in the field of speech conversion [13, 14, 15]. This inspiration led to the development of USD-AC, an unsupervised accent conversion method. USD-AC decouples accent sentences into multiple independent attributes, including linguistic content, accent, speaker identity, and other prosodic information. Subsequently, it replaces the source accent with the target accent and re-synthesizes the mel spectrogram. Experimental results have demonstrated that our

10.21437/Interspeech.2024-664

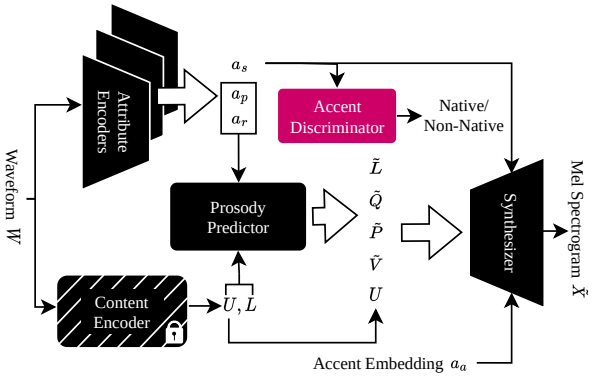

Figure 1: *Model Architecture*

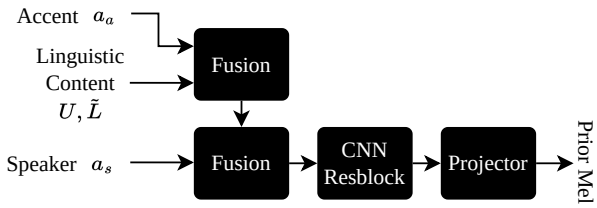

Figure 2: *Filter network with accent embedding.*

approach exhibits superior generalization capabilities compared to traditional supervised accent conversion methods, even for unseen speakers, accents, and content.

2. Methodology

In this study, we extended the architecture based on UUVC [15] to propose the unsupervised accent conversion model USD-AC, centered around accent disentanglement. The architecture is depicted in Figure 1. Since both the pitch attribute a_p and rhythm attribute a_r have specific learning objectives, uncontrollable accent attributes solely reside in the speaker attribute a_s and the linguistic content U. To disentangle the accent attributes from these two components, we implemented two key enhancements:

1) Use a well-trained ASR as a content encoder to reduce the influence of accents on linguistic content, and add accent embedding a_a to represent accent information.

2) The information about the accent in the speaker attribute is removed through the accent adversarial loss, making the accent embedding the only feature that can represent the accent.

2.1. Speech Decomposition

UUVC employs a frozen HuBERT [16] as the content encoder E_{θ_u} to extract unsupervised discrete speech units $U_{rep} = (u_j \in \mathcal{U})_{j=1}^{J}$ from the speech waveform $\mathbf{w}$. And repeated adjacent units are merged to form the linguistic content $U = (u_k \in \mathcal{U})_{k=1}^{K}$ along with their corresponding durations $L = (l_k \in \mathbb{N})_{k=1}^{K}, \sum_k l_k = J$. The architecture of an autoencoder is employed to decompose the speech into rhythm $a_r \in \mathbb{R}^{d_r}$, pitch-energy $a_p \in \mathbb{R}^{d_p}$, and speaker style $a_s \in \mathbb{R}^{d_s}$ by attribute encoders $E_{\theta_i}, i \in \{r, p, s\}$. Subsequently, the prosody predictor H_ϕ, composed of the duration network N_{ϕ_r} and pitch-energy network N_{ϕ_p}, utilizes the decomposed attribute features as input. Ground truth coefficients for each attribute, including duration L, pitch $P = (p_j \in \mathbb{R})_{j=1}^{J}$, energy $Q = (q_j \in \mathbb{R})_{j=1}^{J}$, and voicing $V = (v_j \in \{0, 1\})_{j=1}^{J}$, serve as training targets. Notably, P and V are predicted by the pitch estimator, such as YAAPT [17] and CREPE [18].

$$H_\phi(U, a_r, a_p) = \begin{cases} \tilde{L} \leftarrow N_{\phi_r}(U, a_r) \\ \tilde{P}, \tilde{V}, \tilde{Q} \leftarrow N_{\phi_p}(U, \tilde{L}, a_p) \end{cases} \tag{1}$$

$$\begin{aligned} \tilde{X} &\leftarrow G_\phi(U, \tilde{L}, \tilde{P}, \tilde{V}, \tilde{Q}, a_s, a_a) \\ &= S_{\phi_s}(\tilde{P}, \tilde{V}, a_s) + F_{\phi_f}(U, \tilde{L}, a_s, a_a) \oplus M_{\phi_m}(\tilde{Q}) \end{aligned} \tag{2}$$

After predicting various speech coefficients, the synthesis of the corresponding mel spectrogram X for $\mathbf{w}$ is achieved through the synthesizer G_ϕ. This synthesizer comprises three modules: the source network S_{ϕ_s}, responsible for pitch harmonics correspondence; the Filter network F_{ϕ_f} [19], responsible for spectral envelope correspondence; and the energy network M_{ϕ_m}, which estimates volume level. Unlike methods applied to VC, USD-AC, as shown in Figure 2, introduces a learnable accent embedding $a_a \in \mathbb{R}^{d_a}$ as input to F_{ϕ_f}, thereby explicitly correlating the synthesis process with the accent.

We employ the same loss function $\mathcal{L}_{\text{UUVC}}$ as UUVC to learn the decoupling of fundamental speech attributes and the synthesis of mel spectrograms. In the equation, $\mathcal{L}_P$ and $\mathcal{L}_Q$ use the bin weights of the gaussian-blurred P and Q as training targets for $\tilde{P}$ and $\tilde{Q}$, and, similar to $\mathcal{L}_V$, the loss is calculated using binary cross entropy. $\mathcal{L}_L$ computes the mean square error between the predicted duration $\tilde{L}$ and the ground truth L. Furthermore, following the implementation of UUVC[1], pitch and energy bin weights are separately weighted sum with their respective codebooks in S_{ϕ_s} and M_{ϕ_m}, yielding corresponding features $\mathbf{E}^p$ and $\mathbf{E}^q \in \mathbb{R}^{J \times d}$. $\mathcal{L}_p$ and $\mathcal{L}_q$ represent the mean square error between the estimated bin weight embedding and the true bin weight embedding. Lastly, $\mathcal{L}_{\text{mel}}$ is composed of L1 loss and adversarial loss $\min_{\theta,\phi} \max_{\psi_{\text{mel}}} D_{\psi_{\text{mel}}}(\tilde{X}||X)$ by mel spectrogram discriminator $D_{\psi_{\text{mel}}}$.

$$\mathcal{L}_{\text{UUVC}} = \mathcal{L}_P + \mathcal{L}_Q + \mathcal{L}_V + \mathcal{L}_L + \mathcal{L}_p + \mathcal{L}_q + \mathcal{L}_{\text{mel}} \tag{3}$$

2.2. Decouple Accent from Speaker Style

Although we introduced the accent attribute a_a into the synthesis process, we cannot ensure that the information related to the accent is effectively separated from other attributes. Therefore, as a first step, we replaced HuBERT with a pre-trained ASR as the content encoder E_{θ_u}. This was done to reduce the influence of accent information on linguistic content. Additionally, we introduced an accent discriminator D_{ψ_a}, implemented as an MLP, on the speaker attribute a_s to remove accent features from the speaker style. The goal of the accent discriminator D_{ψ_a} is to maximize the predicted probability of the native accent $\bar{a}_s$ while minimizing the predicted probability of the foreign accent $\hat{a}_s$. On the other hand, the speaker encoder E_{θ_s} aims to maximize the discrimination probability of $\hat{a}_s$ by D_{ψ_a}. With the assistance of pre-trained ASR and the accent adversarial loss $\mathcal{L}_{\text{accent}}$, USD-AC decouples the accent information into a learnable accent embedding a_a.

$$\mathcal{L}_{\text{accent}} = \min_{\theta_s} \max_{\psi_a} D_{\psi_a}(\bar{a}_s||\hat{a}_s) \tag{4}$$

The final loss function of USD-AC is a combination of $\mathcal{L}_{\text{UUVC}}$ and $\mathcal{L}_{\text{accent}}$, mixed according to the decoupling weight λ

[1]https://github.com/b04901014/UUVC

ranging between 0 and 1. The $\mathcal{L}_{\text{accent}}$ term utilizes mean square error, as proposed by Mao et al. [20], to calculate the adversarial loss. We conducted further ablation studies to test the impact of different λ values on the system's performance.

$$\mathcal{L}_{\text{USD-AC}} = \mathcal{L}_{\text{UUVC}} + \lambda \mathcal{L}_{\text{accent}} \quad (5)$$

3. Experiments

Our architecture draws inspiration from the model configuration of UUVC and utilizes fine-tuned wav2vec 2.0[2] [21] as the pre-trained ASR, which predicts 32 English characters and special tokens. To ensure that ASR's textual units, which comprise 32 classes, do not compromise the quality of synthesis due to a lack of rich phonetic information similar to HuBERT's 200 discrete units, we increased the layers of both the source network and filter network from 16 to 24. This enhancement is aimed at improving the quality of the synthesized mel spectrogram. We further used a pre-trained HiFi-GAN to convert the mel spectrogram into 22kHz speech. To reduce training time, we initialized three attribute encoders and the prosody predictor with the pre-trained models from UUVC, which were pre-trained on the VCTK dataset [22]. For optimization, we employed the Adam optimizer with $\beta 1 = 0.5$ and $\beta 2 = 0.9$, and no weight decay. λ (the weight of $\mathcal{L}_{\text{accent}}$) is set to 0.25. The learning rate was set to 1e-4, the total training steps amounted to 150k steps, and there was no warm-up phase. We used a batch size of 32, following the batch calculation method described in UUVC. In practice, the total duration of all speech samples in a small batch did not exceed 32 seconds.

3.1. Datasets

The goal of this study is to convert foreign-accented English to native English. Previous papers have utilized databases such as CMU-ARCTIC [23] and L2-ARCTIC [24]. To augment the number of native accent samples and speaker diversity, we have included the VCTK dataset [22], commonly used in voice conversion. To reduce complexity and stabilize training, accents other than Indian and Israeli in CMU-ARCTIC and VCTK are categorized as native accents. Finally, the three datasets mentioned above have been classified into eight accents.

To balance the number of non-native speakers, we utilized the Israeli accent (with only one speaker in CMU-ARCTIC, rxr) for the unseen accent test, and one male and one female speaker with the Indian accent (ksp and slp in CMU-ARCTIC) were used for unseen speaker testing and excluded from the training data. All remaining speakers were used for the training set, with 5% of each speaker's speech randomly selected as validation data. Additionally, for every non-native accent in L2-ARCTIC, we randomly selected 15 speakers from the Speech Accent Archive (SAA) [25] as external test data.

Lastly, for subjective evaluation, speakers were randomly selected from six non-native accents excluding Israeli accents, with one male and one female speaker from each, totaling 12 speakers. Each speaker was randomly assigned two sentences from the test dataset for accent conversion, resulting in a total of 24 synthesized speech samples.

3.2. Evaluation Setup

This study conducts four objective evaluations to measure the performance of the proposed method: ASR performance [26], MOSNet [27] for naturalness, accent embedding similarity (AES), and speaker embedding similarity (SES) [15]. Experimental results are evaluated on speeches synthesized by HiFi-GAN using ground-truth Mel spectrograms to mitigate quality differences.

The pre-trained ASR model for Character Error Rate (CER) calculation is HuBERT, fine-tuned on the LibriSpeech dataset to achieve 1.9% WER. MOSNet is employed as an objective measure of naturalness, producing scores from 1 to 5, with higher scores indicating greater similarity to human-generated speech. Accent and speaker similarity are assessed using trained verification models, comparing cosine similarity between features of converted and native accents (AES) or original L2 speaker's speeches (SES). These verification models are adversarially trained to decouple accent and speaker characteristics, enhancing evaluation accuracy. These models's accuracy exceeds 90% for accent classification and achieves less than 1% equal error rate for speaker verification.

In addition to subjective evaluations, we also invited 5 students to conduct subjective evaluations on the naturalness of speech, the degree of accent, and the similarity to the speaker for the systems. For naturalness evaluation, we utilize the Mean Opinion Score (MOS) ranging from 1 to 5, with reference speech samples drawn from the Voice Conversion Challenge 2018 [28] to calibrate listener ratings as per guidelines in [29]. Accent evaluation employs a nine-point Likert scale (1-9) to rate foreign accent presence, with higher scores indicating a stronger foreign accent. For speaker similarity, ABX testing is employed, where listeners discern which system's converted speech resembles the original L2 input speech in timbre.

3.3. Comparing Baseline

To facilitate a comparison with past non-reference supervised accent conversion techniques, in cases where open-source implementations were unavailable, we employed a baseline architecture inspired by [8] and made the following adjustments: 1) Utilized the same ASR model as our proposed method to extract bottleneck features, replacing the ASR model used in the original paper. 2) Adopted the ECAPA-TDNN[3] [30] as the speaker encoder to extract speaker features. ECAPA-TDNN was trained on Voxceleb 1 [31] and Voxceleb 2 [32] datasets, achieving a 0.8% equal error rate on the cleaned Voxceleb 1 test set. To ensure training stability and generation quality for the autoregressive model, scheduled sampling [33] training techniques were applied. VCTK was excluded from the data used for training the USD-AC to create parallel data for training the baseline model. The 'bdl' US speaker from CMU-ARCTIC was employed to extract target bottleneck features.

4. Results

In our experiments, we validated the performance of USD-AC and the supervised learning baseline in converting the speech from a foreign to a native accent. This assessment extended to unseen speakers, accents, and speech content. Furthermore, to demonstrate the performance of the proposed accent disentanglement, we compared different weights for the accent adversarial loss and different linguistic encoders through ablation experiments on the validation set.

To confirm that there is a statistical performance difference between our proposed system and the baseline, we use the

[2]https://huggingface.co/facebook/wav2vec2-base-960h

[3]https://huggingface.co/speechbrain/spkrec-ecapa-voxceleb

Table 1: *Results of unseen accent.*

Methods	CER (%) ↓	AES ↑	SES ↑	MOSNet ↑
Baseline	9.88	0.592	0.581	2.536
USD-AC	**3.27**†	**0.846**†	**0.803**†	**2.822**†

Table 2: *Results of unseen speakers.*

Methods	CER (%) ↓	AES ↑	SES ↑	MOSNet ↑
Baseline	**3.12**†	0.592	0.671	2.565
USD-AC	6.07	**0.793**†	**0.780**†	**2.838**†

method proposed by [34] to calculate the distribution of the performance difference, and mark the evaluation results with † to indicate that the 95% confidence interval does not include the value 0.0.

4.1. Comparison with Baseline

Tables 1 and 2 demonstrate that our approach outperforms the supervised learning baseline in terms of conversion performance and speech quality for unseen speakers and accents. Particularly notable is the baseline's tendency to exhibit conversion failures and consequently increased CER for unseen accents, thereby confirming the limitations of existing supervised accent conversion methods due to the quantity of available labeled data, which affects their generalization capability. In table 3, to assess performance in accent conversion for unseen speech content, we retrained USD-AC from scratch with the same training data as the baseline and compared it against the baseline using the SAA as test data. Results indicate that due to insufficient training data, the baseline based on the Tacotron 2 [35] architecture struggles to model unseen speech content, resulting in the majority of evaluation metrics falling below the USD-AC without VCTK. Moreover, USD-AC's performance significantly improves with the addition of VCTK, validating our approach's capability to improve accent conversion and speaker preservation through the use of additional unannotated data.

4.2. Human Evaluation Results

The subjective evaluation is shown in Table 4. It can be seen that the score of the proposed method is almost 4 in terms of naturalness, which is very close to the ground truth. The degree of non-native accent in the speech was significantly lower compared to baseline and true L2 scores. Although in the speaker retention test, the score is lower than the baseline, in the statistical test, the 95% confidence interval of the distribution of the performance difference is between (-1.917, 0.583), so we cannot reject that the two systems have The null hypothesis of identical speaker retention efficacy. In summary, subjective evaluations show that the proposed model can produce more natural speech without non-native accents and also exhibits close speaker similarity to a baseline that leverages a large corpus to train the speaker encoder.

Table 3: *Objective evaluation results on the SAA test set.*

Methods	CER (%) ↓	AES ↑	SES ↑	MOSNet ↑
Baseline	79.54	0.567	**0.644**†	2.168
ɥ↦-VCTK	**21.38**†	**0.633**†	0.603	**2.488**†
USD-AC	20.17	0.655	0.747	2.909

Table 4: *Results of Human Evalutions.*

	L1	L2	Baseline	USD-AC
MOS ↑	4.867	4.158	2.483	**3.917**†
Accentedness ↓	1.450	5.467	3.833	**3.067**†
Speaker Similarity (%)↑	-	-	56.67	43.33

Table 5: *Comparison of different linguistic context embedding.*

E_{θ_u}	$\lvert\mathcal{U}\rvert$	$\mathcal{L}_{\text{accent}}$	CER (%) ↓	AES ↑	SES ↑
HuBERT	200	-	15.47	0.629	0.707
HuBERT	200	$\surd$	14.52	0.679	0.663
ASR	32	$\surd$	**9.55**	**0.716**	**0.696**

4.3. Ablation Study

Table 5 confirms that training a well-trained ASR system can mitigate the impact of accents on linguistic content without requiring fine-tuning. Conversely, the discrete speech units extracted by HuBERT undergo changes based on pronunciation differences, leading to incomplete decoupling of linguistic content and accent and thus reducing the performance of accent conversion. In table 6, the impact of different accent adversarial loss weights on accent information disentanglement is compared. While larger weights lead to better disentanglement, they also tend to reduce speaker preservation.

5. Conclusions

This study proposes an unsupervised accent conversion system that does not require parallel data and text transcriptions for training. It demonstrates the potential of unsupervised learning in accent conversion through experiments. The primary contribution lies in the use of speech decomposition techniques to disentangle accents from other speech information. By adjusting the weight of the accent adversarial loss, the speaker embeddings are perturbed, achieving a balance between speaker retention and accent conversion performance. Objective evaluation results demonstrate that the proposed framework maintains consistent conversion performance on completely unseen test sets and can benefit from additional unlabeled data for training. This leads to superior performance across all evaluation metrics when compared to the baseline trained on parallel data. We believe that by leveraging a large amount of unlabeled speech data, this framework can be applied to speech modification tasks in the future, such as those related to articulatory disorders and other low-resource challenges.

Table 6: *Comparison of different weights of accent adversarial loss.*

λ	CER (%) ↓	AES ↑	SES ↑
1	9.55	**0.716**	0.696
0.5	**9.40**	0.690	0.738
0.25	9.51	0.691	0.783
0	9.45	0.603	**0.862**

6. References

[1] D. Felps, H. Bortfeld, and R. Gutierrez-Osuna, "Foreign accent conversion in computer assisted pronunciation training," *Speech communication*, vol. 51, no. 10, pp. 920–932, 2009.

[2] S. Ding, C. Liberatore, S. Sonsaat, I. Lučić, A. Silpachai, G. Zhao, E. Chukharev-Hudilainen, J. Levis, and R. Gutierrez-Osuna, "Golden speaker builder–An interactive tool for pronunciation training," *Speech Communication*, vol. 115, pp. 51–66, 2019.

[3] O. Turk and L. M. Arslan, "Subband based voice conversion," in *Seventh International Conference on Spoken Language Processing*, 2002.

[4] G. Zhao and R. Gutierrez-Osuna, "Using phonetic posteriorgram based frame pairing for segmental accent conversion," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 10, pp. 1649–1660, 2019.

[5] G. Zhao, S. Sonsaat, J. Levis, E. Chukharev-Hudilainen, and R. Gutierrez-Osuna, "Accent conversion using phonetic posteriorgrams," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5314–5318.

[6] S. Ding, G. Zhao, and R. Gutierrez-Osuna, "Accentron: Foreign accent conversion to arbitrary non-native speakers using zero-shot learning," *Computer Speech and Language*, vol. 72, p. 101302, 2021.

[7] G. Zhao, S. Ding, and R. Gutierrez-Osuna, "Converting foreign accent speech without a reference," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2367–2381, 2021.

[8] W. Quamer, A. Das, J. M. Levis, E. Chukharev-Hudilainen, and R. Gutierrez-Osuna, "Zero-shot foreign accent conversion without a native reference," in *Interspeech*, 2022.

[9] T.-N. Nguyen, N.-Q. Pham, and A. H. Waibel, "Accent conversion using pre-trained model and synthesized data from voice conversion," in *Interspeech*, 2022.

[10] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE TRANSACTIONS on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.

[11] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A generative model for raw audio," in *Proc. 9th ISCA Workshop on Speech Synthesis Workshop (SSW 9)*, 2016, p. 125.

[12] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in Neural Information Processing Systems*, vol. 33, pp. 17 022–17 033, 2020.

[13] C. H. Chan, K. Qian, Y. Zhang, and M. Hasegawa-Johnson, "Speechsplit2. 0: Unsupervised speech disentanglement for voice conversion without tuning autoencoder bottlenecks," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6332–6336.

[14] K. Qian, Y. Zhang, S. Chang, M. Hasegawa-Johnson, and D. Cox, "Unsupervised speech decomposition via triple information bottleneck," in *International Conference on Machine Learning*. PMLR, 2020, pp. 7836–7846.

[15] L.-W. Chen, S. Watanabe, and A. Rudnicky, "A unified one-shot prosody and speaker conversion system with self-supervised discrete speech units," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[16] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[17] K. Kasi and S. A. Zahorian, "Yet another algorithm for pitch tracking," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, 2002, pp. I–361–I–364.

[18] J. W. Kim, J. Salamon, P. Q. Li, and J. P. Bello, "Crepe: A convolutional representation for pitch estimation," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 161–165, 2018.

[19] H.-S. Choi, J. Lee, W. Kim, J. Lee, H. Heo, and K. Lee, "Neural analysis and synthesis: Reconstructing speech from self-supervised representations," *Advances in Neural Information Processing Systems*, vol. 34, pp. 16 251–16 265, 2021.

[20] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. P. Smolley, "Least squares generative adversarial networks," *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2813–2821, 2016.

[21] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[22] J. Yamagishi, C. Veaux, and K. MacDonald, "CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92)," 2019.

[23] J. Kominek and A. W. Black, "The CMU arctic speech databases," in *Speech Synthesis Workshop*, 2004.

[24] G. Zhao, S. Sonsaat, A. Silpachai, I. Lucic, E. Chukharev-Hudilainen, J. M. Levis, and R. Gutierrez-Osuna, "L2-ARCTIC: A non-native english speech corpus," in *Interspeech*, 2018.

[25] Weinberger and Steven, "Speech Accent Archive," 2015.

[26] T. Hayashi, R. Yamamoto, K. Inoue, T. Yoshimura, S. Watanabe, T. Toda, K. Takeda, Y. Zhang, and X. Tan, "Espnet-tts: Unified, reproducible, and integratable open source end-to-end text-to-speech toolkit," in *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2020, pp. 7654–7658.

[27] C.-C. Lo, S.-W. Fu, W.-C. Huang, X. Wang, J. Yamagishi, Y. Tsao, and H.-M. Wang, "MOSNet: Deep learning-based objective assessment for voice conversion," in *Proc. Interspeech 2019*, 2019, pp. 1541–1545.

[28] J. Lorenzo-Trueba, J. Yamagishi, T. Toda, D. Saito, F. Villavicencio, T. H. Kinnunen, and Z. Ling, "The voice conversion challenge 2018: Database and results," 2018.

[29] P. C. Loizou, "Speech quality assessment," in *Multimedia analysis, processing and communications*. Springer, 2011, pp. 623–654.

[30] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification," in *Proc. Interspeech 2020*, 2020, pp. 3830–3834.

[31] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: A large-scale speaker identification dataset," in *Interspeech*, 2017.

[32] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep speaker recognition," in *Interspeech*, 2018.

[33] S. Bengio, O. Vinyals, N. Jaitly, and N. Shazeer, "Scheduled sampling for sequence prediction with recurrent neural networks," *Advances in neural information processing systems*, vol. 28, 2015.

[34] M. Keller, S. Bengio, and S. Wong, "Benchmarking non-parametric statistical tests," *Advances in neural information processing systems*, vol. 18, 2005.

[35] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. 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.