Conference Name #1234

Assignment 1

Anshul Thakur (B21CS085)
Samay Meher (B22AI048)

Task

Voice conversion is the process of modifying a source speaker's voice to resemble that of a target speaker while preserving the linguistic content of the original speech, and in the zero-shot or one-shot context, the challenge is compounded by the fact that only a minimal amount or even a single sample from the target speaker is available. This task is critically important in various real-world applications including personalized voice assistants, where users expect the system to adapt to their unique vocal characteristics, and in media and entertainment for tasks such as dubbing and voice acting, where achieving a natural and convincing replication of a target voice can significantly enhance the viewer's experience. Moreover, voice conversion technology has significant potential in accessibility solutions by providing customized speech synthesis for individuals with speech impairments, and in security and privacy applications where voice anonymization is required. The ability to perform effective voice conversion with scarce data not only pushes the boundaries of current machine learning methodologies—particularly in the realms of deep learning and representation learning—but also opens up new possibilities for cross-domain applications such as multilingual speech synthesis and adaptive human-computer interaction. Voice conversion can also help in creating better audio/speech deepfake detection frameworks as it can help in adversarial training.

link for repository: https://github.com/dr-ghost/Speech-Understanding

1. ControlVC: Zero-Shot Voice Conversion with Time-Varying Controls on Pitch and Speed

1.1. Paper

ControlVC [1] is a neural voice conversion system designed to provide time-varying control over pitch and speed while maintaining timbre and linguistic content. Unlike previous approaches that typically focus on global or utterance-level control, ControlVC introduces local control mechanisms for dynamic modifications of pitch and speed, enabling a more flexible and natural voice conversion process.

The system architecture consists of three primary stages: pre-processing, analysis, and synthesis. The pre-processing stage modifies the speed of the source utterance using the Time-Domain Pitch-Synchronous Overlap and Add (TD-PSOLA) algorithm based on an input speed control curve. The analysis stage extracts and modifies the pitch contour of the pre-processed source utterance, which is then encoded into a discrete representation using a Vector Quantization-Variational Autoencoder (VQ-VAE) based pitch encoder. A pre-trained HuBERT model extracts the linguistic embedding from the modified source utterance, ensuring linguistic content is maintained, while a speaker encoder extracts the speaker embedding from the target utterance. These embeddings are then passed to the synthesis stage, where a modified HiFi-GAN vocoder generates the waveform of the converted speech.

The pre-trained models play a critical role in the system's zero-shot capability. The speaker encoder is trained on a combination of VoxCeleb and LibriSpeech datasets to generalize across unseen speakers. The linguistic encoder, derived from HuBERT, is pre-trained on extensive speech datasets, ensuring robustness across different linguistic variations. The VQ-VAE pitch encoder is trained separately on the VCTK dataset to learn an effective pitch representation, and the HiFi-GAN vocoder is trained from scratch for high-fidelity speech synthesis.

ControlVC is evaluated on the CSTR VCTK Corpus, which contains 44 hours of English speech from 110 speakers. The training set includes 100 speakers, while 10 speakers (5 male, 5 female) are used for testing. The evaluation includes both objective and subjective tests. Objective metrics include word error rate (WER) for intelligibility assessment and cosine similarity scores for speaker similarity. Subjective evaluations are conducted through Mean Opinion Score (MOS) tests on naturalness and timbre similarity, as well as a controllability test to assess the effectiveness of pitch and speed modifications.

Baseline comparisons include two self-constructed methods: P-LPC, which applies TD-PSOLA for pitch and speed control with linear predictive coding for timbre transfer, and P-AutoVC, which integrates TD-PSOLA with AutoVC for timbre transfer. ControlVC consistently outperforms

Conference Name #1234

both baselines in terms of speaker similarity, intelligibility, and controllability across different control settings. MOS results indicate that ControlVC maintains naturalness and timbre similarity even under controlled modifications of pitch and speed.

The training process for ControlVC includes 350,000 steps for the HiFi-GAN vocoder using an Adam optimizer with an initial learning rate of 0.0002 and decay rate of 0.999. The pitch encoder is pre-trained for 40,000 steps on VCTK, and HuBERT embeddings are processed through K-means clustering for linguistic representation. The experiments involve zero-shot, non-parallel voice conversion across 90 speaker pairs, with control curves applied to manipulate pitch and speed dynamically.

ControlVC is the first VC method that achieves time-varying control on pitch and speed, and noexsting methods were found for direct comparison that had time-variying control on pitch and speed.

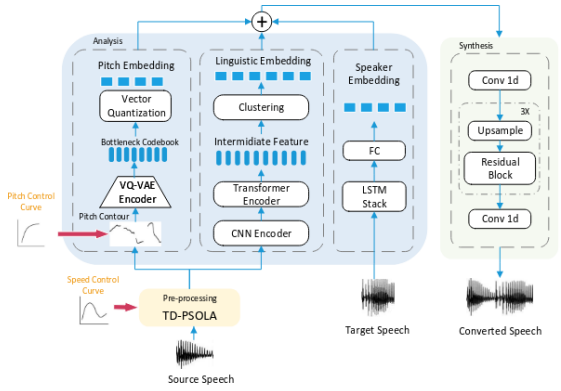


Figure 1: *System Overview.*

1.2. Results

The VC experiments are performed amng all 90 pairs of 10 test speakers. Each utterance of one test speaker is converted to each of the other 9 speakers' voices. Each test speaker reads a different set of sentences. All the 10 speakers and their sen-tences are unseen during training. In this experiment, the paper applies control curves for speed and/or pitch. Four control settings are tested: "No Control" - tra-ditional voice conversion without any explicit control; "Pitch Only" and "Speed Only" denote pitch or speed control but not both; "Speed+Pitch" means both aspects are controlled. The paper test two curves for pitch control: stressing (i.e., pitch rising abruptly then going down gradually) and rising, and three curves for speed control: parabola, speed up and slow down. The control settings and the control curves are drawn with equal probability for each conversion. The paper perform both subjective and objective evaluations to assess the conversion quality, intelligibility and controllability of their proposed system.

1.2.1 Objective Evaluation

The paper conducts an objective evaluation to assess the speech intel-ligibility, timbre similarity and controllability of the converted utterances. For speech intelligibility evaluation, they used IBM speech recognition service to transcribe converted speech into text and then calculate the word error rate (WER) against the ground-truth transcripts. For timbre similarity, They first use a pre-trained speaker encoder Resemblyzer. to ex-tract speaker embeddings of the converted and the target utter-ance. Then they scored the speaker similarity by calculat-ing the cosine distance between the embeddings on a scale of 0 to 1, the higher the more similar.

Table 1: *Objective evaluation results.*

		Sim. ↑	WER (%) ↓
GT		**1.00**	**9.82**
No Control	P-LPC	0.65	89.12
	P-AutoVC	0.66	76.51
	Proposed	**0.85**	**10.99**
Pitch Only	P-LPC	0.66	88.56
	P-AutoVC	0.65	72.83
	Proposed	**0.82**	**12.40**
Speed Only	P-LPC	0.65	88.64
	P-AutoVC	0.65	72.59
	Proposed	**0.84**	**16.37**
Pitch+Speed	Proposed	**0.83**	**22.46**

1.2.2 Subjective Evaluation

The paper reports on two subjective experiments conducted via a custom survey website that was publicly available to the University of Rochester community and its alumni, with no monetary incentives provided. The first experiment focused on assessing audio quality using the mean opinion score (MOS) methodology. In this test, each participant was presented with a set of utterances that included a source, a target, and several converted samples produced by ControlVC and two baseline systems. The order of the converted utterances was randomized. Participants rated each sample on a scale of 1 to 5 for both naturalness and timbre similarity (with higher scores indicating better quality), resulting in a total of 1,398 ratings from 233 samples. The results demonstrated that ControlVC achieved the best MOS scores among the three methods across all controlled settings, and the introduction of control features only slightly reduced the quality of the converted speech.

Conference Name #1234

The authors also compared their baseline P-AutoVC performance with previously reported results. While earlier work using AutoVC reported MOS scores of approximately 3.1 for naturalness and 2.9 for similarity (with another study reporting a naturalness MOS of 2.59), the P-AutoVC baseline in this study obtained scores of 2.17 and 2.68, respectively. The performance gap was attributed to factors such as the exclusive use of the VCTK dataset (in contrast to previous studies that also used LibriTTS) and the potential artifacts introduced by PSOLA preprocessing applied in earlier work.

The second experiment evaluated the controllability of the ControlVC system. In this test, participants compared uncontrolled and controlled conversion results alongside a visual representation of the control curve. They were asked to rate how accurately the curve reflected the changes in pitch or speed on a scale from 1 (“not at all accurate”) to 5 (“very accurate”). Each test round included evaluations for pitch control, speed control, and a combined pitch+speed control for the same source-target pair, with a total of 676 ratings collected from 169 rounds. To establish a baseline for uncontrollability, 15% of the single-factor control conversions were paired with a fake control curve (created by flipping or circularly shifting the actual curve). Statistical analysis using paired t-tests showed that ControlVC’s pitch control was rated significantly higher than the baseline (p ¡ 0.01), and although the difference in speed control was subtler, it remained statistically significant (p ¡ 0.01). Moreover, the MOS ratings between single-factor controls (pitch-only or speed-only) and the combined pitch+speed control did not differ significantly (p = 0.07 for pitch; p = 0.43 for speed), indicating that the system can effectively manage simultaneous control of both factors without compromising overall quality.

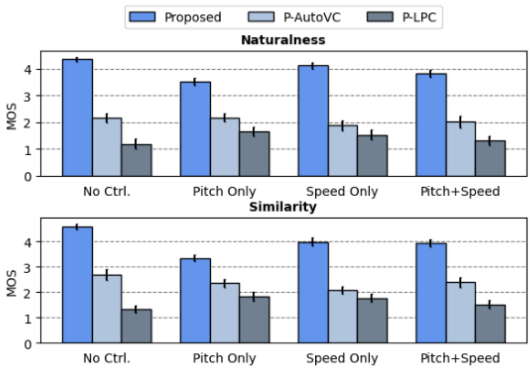


Figure 2: *MOS results on audio quality (naturalness and timbre similarity) with 95% confidence intervals.*

Table 2: *MOS results on controllability with 95% confidence intervals.*

Controllability Rating			
		Pitch	**Speed**
Pitch Only	Real Curve	**3.38 ± 0.15**	-
	Fake Curve	3.00 ± 0.19	-
Speed Only	Real Curve	-	**3.37 ± 0.25**
	Fake Curve	-	3.21 ± 0.19
Pitch+Speed	Real Curve	**3.18 ± 0.15**	**3.41 ± 0.14**

1.3. Strengths and Limitations

It achieves zero-shot, non-parallel voice conversion, allowing adaptation to unseen speakers while maintaining linguistic content and speaker identity through pre-trained encoders. The system outperforms self-constructed baselines in speech quality, intelligibility, and controllability, as demonstrated through objective metrics like word error rate (WER) and speaker similarity, along with subjective MOS ratings. Additionally, the use of a diverse dataset (VCTK corpus) strengthens the robustness of the model’s performance, making it more generalizable across different speech variations. It also is one of the first VC method that achieves time-varying control on pitch and speed.
ControlVC does not mimic the natural prosody of the target speaker, which may impact its effectiveness in high-fidelity voice cloning. The study lacks a direct comparison with state-of-the-art voice conversion systems, limiting the scope of performance benchmarking beyond self-constructed baselines. Additionally, TD-PSOLA preprocessing introduces artifacts, affecting speech quality in some cases, and when both pitch and speed controls are applied simultaneously, noticeable degradation in quality occurs. Finally, the subjective evaluation relies on a limited participant pool, which may affect the generalizability of user perception results.

2. End-to-End Zero-Shot Voice Conversion with Location-Variable Convolutions

2.1. Paper

End-to-End Zero-Shot Voice Conversion with Location-Variable Convolutions [3] is an end-to-end zero-shot voice conversion model designed to convert speech from one speaker to another while directly synthesizing audio, eliminating the need for a separate vocoder. The model leverages location-variable convolutions to efficiently integrate content and speaker information, enabling high-quality conversion with relatively few parameters.
The architecture consists of three main components: a generator, a speaker encoder, and a set of discriminators. The generator, based on a fully convolutional neural network,

Conference Name #1234

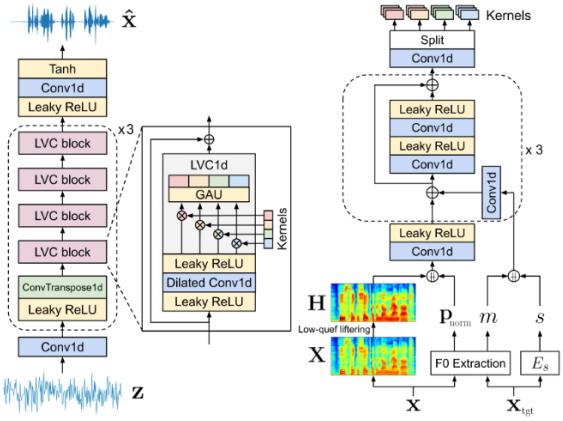
takes content and speaker embeddings as input and uses LVCs to generate the output waveform. The speaker encoder is a pre-trained ResNet-34 model trained on VoxCeleb2, which extracts speaker embeddings. The discriminators, including a multi-resolution spectrogram discriminator and a multi-period waveform discriminator, facilitate adversarial training to improve audio quality.

The model takes as input a set of carefully designed features to effectively separate content and speaker information. The content features include spectral envelopes obtained through low-quefrency liftering and per-frame normalized quantized log-F0 features. These features aim to capture linguistic content while minimizing speaker identity leakage. The speaker features are derived from the pre-trained speaker encoder, which extracts speaker embeddings from a given utterance. Additionally, a speaker's median log-F0 is quantized into bins and used as an extra condition. These features are fed into the LVC-VC generator, where LVCs efficiently combine them to produce natural-sounding speech.

During training, LVC-VC utilizes self-reconstruction and speaker similarity loss to learn robust voice conversion. Self-reconstruction loss ensures that the generated speech maintains intelligibility by modifying content embeddings with warping techniques to remove residual speaker characteristics. The speaker similarity criterion forces the generator to produce speech that closely matches the characteristics of the target speaker. The model is trained using a combination of LSGAN loss, multi-resolution short-time Fourier transform loss, and auxiliary reconstruction loss.

The dataset used for training and evaluation is the VCTK corpus, containing recordings from 109 speakers. Training is conducted on 99 seen speakers, while 10 unseen speakers are reserved for zero-shot evaluation. The system is trained using the AdamW optimizer with a batch size of 32, first focusing on self-reconstruction for 1.8 million iterations before incorporating speaker similarity constraints.

Experimental results demonstrate that LVC-VC achieves a well-balanced trade-off between voice style transfer and intelligibility compared to baselines such as AutoVC, AdaIN-VC, and NVC-Net. It outperforms other models in preserving speech clarity while maintaining high-quality voice conversion. Ablation studies confirm that each component, including Gaussian speaker embeddings, speaker similarity loss, and feature warping, contributes significantly to overall performance. Additional analysis of the synthesis process shows that LVC-VC effectively learns to disentangle and recombine speaker and content information, generating interpretable and high-fidelity audio.



(a) *Generator.* (b) *Kernel predictor for LVCs.*

Figure 1: *The components of the overall LVC-VC architecture. Content and speaker features are fed into the kernel predictors, which output kernels for the LVC layers in the generator. Each kernel predictor outputs the kernels for all four LVC blocks in a given transposed convolutional stack (shown in red, yellow, green, and blue at the right of (a) and top of (b)). $+\!\!+$ denotes stacking/concatenation in (b).*

2.2. Results

To evaluate the performance of LVC-VC, the paper conducted comparisons against several baseline models, including AdaIN-VC, AGAIN-VC, AutoVC, AutoVC-F0, Blow, and NVC-Net. For fairness, all baseline models were trained from scratch using the same dataset and spectrogram configurations as LVC-VC. Models requiring a vocoder were paired with the UnivNet-c16 vocoder, trained on the 99 seen speakers from the VCTK dataset and the train-clean-360 split of the LibriTTS dataset.

Table 1 presents the quantitative evaluation results across different conversion settings. The evaluation criteria included Mean Opinion Score (MOS) for naturalness, speaker similarity, Character Error Rate (CER) for intelligibility, and Equal Error Rate (EER) for speaker verification. The results indicate that most baseline models exhibited a trade-off between voice style transfer (VST) and audio quality. Specifically, models such as AdaIN-VC, AGAIN-VC, and NVC-Net demonstrated strong voice style transfer capabilities but generated lower-quality audio. Conversely, AutoVC and AutoVC-F0 maintained higher audio quality but underperformed in VST accuracy.

LVC-VC effectively balances this trade-off, achieving competitive performance across MOS and similarity metrics while consistently producing the lowest CER, particularly in unseen-to-seen (u2s) and unseen-to-unseen (u2u) conversion settings. These results highlight that LVC-VC maintains intelligibility exceptionally well. Notably, despite its compact model size, LVC-VC delivers one of the most balanced performances overall, ensuring high-fidelity speech

Conference Name #1234

Table 1: *Model sizes and voice conversion evaluation results on the three conversion settings. We include 95% confidence intervals for MOS. Bold and underlined values indicate the best and second best scores in a given metric, respectively.*

Model	# Params	Seen-to-Seen				Unseen-to-Seen				Unseen-to-Unseen			
		MOS	Sim	CER	EER	MOS	Sim	CER	EER	MOS	Sim	CER	EER
Ground Truth	–	4.61 ± 0.11	91.88	3.81	0.00	4.74 ± 0.08	95.63	2.93	0.00	4.62 ± 0.11	92.50	3.58	0.00
UnivNet	–	4.33 ± 0.12	90.63	4.97	5.00	4.51 ± 0.11	97.50	3.59	0.00	4.58 ± 0.10	92.50	4.68	0.00
LVC-VC	5.97M	3.54 ± 0.17	51.88	**11.00**	17.50	3.51 ± 0.18	43.75	**7.03**	20.00	3.24 ± 0.18	38.75	**8.29**	**26.25**
AdaIN-VC	4.89M	2.35 ± 0.16	63.75	22.78	28.75	2.57 ± 0.17	53.13	16.41	36.25	2.41 ± 0.16	**49.38**	20.60	35.00
AGAIN-VC	7.93M	2.13 ± 0.16	48.75	25.12	18.75	2.39 ± 0.16	46.25	23.94	31.25	2.26 ± 0.15	42.50	25.79	31.25
AutoVC	40.68M	$\mathbf{3.84 \pm 0.15}$	30.63	11.15	30.00	$\mathbf{3.71 \pm 0.16}$	31.88	10.65	26.25	$\mathbf{3.61 \pm 0.17}$	13.13	12.07	63.75
AutoVC-F0	41.21M	3.44 ± 0.16	32.50	12.53	28.75	3.39 ± 0.16	37.50	10.54	32.50	3.31 ± 0.17	20.00	14.20	65.00
Blow	62.11M	1.78 ± 0.15	29.38	18.33	52.50	–	–	–	–	–	–	–	–
NVC-Net	15.13M	2.96 ± 0.19	**76.88**	31.46	**12.50**	3.14 ± 0.19	**66.25**	26.91	**11.25**	3.10 ± 0.20	40.00	26.27	37.50

Table 2: *Unseen-to-unseen conversion results for ablations.*

Model	CER	EER	NISQA
LVC-VC	8.29	26.25	3.50 ± 0.13
w/o Gaussian embeddings	11.11	25.00	2.89 ± 0.14
w/o SSC loss	6.64	68.75	3.83 ± 0.13
w/o warping $\mathbf{H}$	7.39	51.25	3.62 ± 0.17
w/o $\mathbf{p}_{\text{norm}}$	8.60	32.50	3.36 ± 0.18
w/o m	9.90	28.75	3.47 ± 0.14

synthesis and accurate voice style conversion.

2.2.1 Ablation Studies

The paper also conducted ablation studies on various aspects of LVC-VC the results are shown in Table 2. As a proxy measure for subjective MOS, the paper used NISQA, which provides an estimate of an utterance's speech quality on a scale from 1 to 5. They found that each of the ablated components contributed meaningfully to the model's performance. Training on fixed speaker embeddings instead of sampling from a Gaus-sian caused audio quality to degrade, suggesting that training on more diverse embeddings helps the model generalize better to new speakers. Training without the SSC loss or without warp-ing H caused VST performance to decrease. Finally, pnorm and m contributed to general performance gains in all metrics.

2.3. Strengths and Limitations

Its end-to-end approach to zero-shot voice conversion, eliminating the need for a separate vocoder. By using location-variable convolutions (LVCs), the model efficiently captures and combines content and speaker features, leading to better intelligibility and voice style transfer balance compared to other models. The architecture is compact, requiring fewer parameters than comparable systems like AutoVC or Blow, while maintaining high-quality speech synthesis. Furthermore, the model's training strategy, including warping techniques to remove residual speaker information, enhances its generalization ability to unseen speakers.

However, weaknesses include the fact that while LVC-VC achieves a balance between voice style transfer and intelligibility, it does not outperform the best competing models in either category individually. Its Mean Opinion Score (MOS) is lower than ground truth and traditional vocoding methods, indicating that the audio quality still has room for improvement. Additionally, while the model is compact, training complexity and inference time could still be a concern due to the use of kernel predictor networks in LVCs. Finally, despite its ability to generalize, there is still some degradation in similarity and intelligibility when applied to completely unseen speakers, as shown by the lower similarity scores in unseen-to-unseen conversions.

3. Diff-HierVC: Diffusion-based Hierarchical Voice Conversion with Robust Pitch Generation and Masked Prior for Zero-shot Speaker Adaptation

3.1. Paper

Diff-HierVC [2] is a diffusion-based hierarchical voice conversion (VC) system designed to improve pitch accuracy and speaker adaptation in zero-shot voice conversion scenarios. The model introduces a two-stage architecture that hierarchically converts speech components, enhancing voice style transfer performance. The system comprises two primary modules: DiffPitch, responsible for generating the target speaker's fundamental frequency (F0), and DiffVoice, which synthesizes a high-quality Mel-spectrogram using the generated F0 and other disentangled speech representations. The introduction of a masked prior further refines speaker adaptation, making the model more robust in unseen speaker scenarios.

The architecture of Diff-HierVC follows a hierarchical structure that efficiently processes and disentangles speech components. The model consists of multiple encoders, diffusion models, and a vocoder to reconstruct the final waveform. The content encoder extracts speaker-independent linguistic information using intermediate representations from XLS-R, a large-scale self-supervised model. The pitch encoder processes the fundamental frequency (F0), which is extracted from the source speech using the YAAPT algorithm at four times the resolution of the Mel-spectrogram. The style encoder captures speaker characteristics from the Mel-spectrogram and provides guidance for voice style adaptation. The hierarchical structure enables precise separation of speech attributes and ensures effective transformation into the target voice.

The DiffPitch module is responsible for generating an accurate F0 contour for the target speaker. It utilizes a WaveNet-based conditional diffusion model, where an iterative denoising process refines the predicted pitch representation. The pitch encoder transforms the normalized F0 of the source speaker into a latent representation, which is then regularized using pitch reconstruction loss. The forward diffusion process gradually adds noise to the pitch representation, while the reverse process removes noise to generate a clean, high-quality F0 sequence.

The DiffVoice module synthesizes speech by constructing

Conference Name #1234

a Mel-spectrogram conditioned on content, target F0, and speaker style. It follows a source-filter theory-based approach, where speech is decomposed into source (pitch) and filter (vocal tract characteristics) representations. The source-filter encoder, which consists of separate source and filter encoders, reconstructs an intermediate Mel-spectrogram Zm from disentangled speech features. This intermediate representation acts as a data-driven prior, guiding the diffusion process to improve voice style transfer accuracy. DiffVoice employs a 2D-UNet structure with three blocks, processing features hierarchically at different resolutions. The generated Mel-spectrogram is then converted to a waveform using a HiFi-GAN vocoder.

For training, the model utilizes the LibriTTS dataset, specifically the train-clean-360 and train-clean-100 subsets, which together contain 245 hours of speech from 1,151 speakers. The VCTK dataset is used for evaluating zero-shot VC performance. The audio is processed into 80-bin log-Mel spectrograms with a hop size of 320 and a window size of 1,280. Training is conducted for 2 million steps on two NVIDIA A100 GPUs using the AdamW optimizer with a decay factor of 0.9991/8. DiffPitch employs a DiffWave-based structure, while DiffVoice follows a 2D-UNet architecture with hierarchical feature processing.

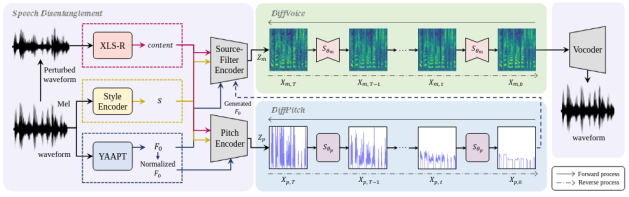


Figure 1: *Overall framework*

3.2. Results

Table 1: *Zero-shot VC results on unseen speakers from VCTK dataset*

Method	iter.	nMOS (↑)	sMOS (↑)	CER (↓)	WER (↓)	EER (↓)	SECS (↑)	Params.
GT	-	3.68±0.09	3.59±0.03	0.21	2.17	-	-	-
GT (Mel + Vocoder)	-	3.70±0.09	3.42±0.04	0.21	2.17	-	0.989	13M
AutoVC [1]	-	3.56±0.09	2.63±0.07	5.14	10.55	37.32	0.715	30M
VoiceMixer [2]	-	3.59±0.09	2.98±0.06	1.08	3.31	20.75	0.797	52M
SR [3]	-	3.51±0.10	2.83±0.06	5.14	10.55	37.32	0.715	15M
DiffVC*[5]	6 / 30	3.39±0.09 / 3.48±0.09	2.81±0.06 / 2.88±0.06	6.86 / 7.51	13.77 / 14.42	9.25 / 10.05	0.826 / 0.842	127M
DiffVC [5]	6 / 30	3.63±0.09 / 3.63±0.09	2.98±0.06 / 2.94±0.06	5.82 / 6.92	11.76 / 13.19	25.30 / 24.01	0.786 / 0.785	123M
Diff-HierVC (Ours)	6 / 30	**3.70±0.09 / 3.74±0.09**	3.03±0.06 / 3.02±0.06	**0.83 / 1.19**	**3.11 / 3.58**	3.29 / 3.66	0.861 / 0.860	18M
Diff-HierVC-Finetune (Ours)	6 / 30	3.65±0.09 / 3.66±0.09	**3.04±0.05 / 3.07±0.05**	0.97 / 1.34	3.15 / 3.75	**1.50 / 1.26**	**0.894 / 0.894**	18M

TWe conduct various subjective and objective evaluation on the zero-shot VC scenario with three models: (1) autoencoder based VC model, AutoVC [1], (2) GAN based VC model, VoiceMixer, (3) unit-based end-to-end speech model, Speech Resynthesis (SR) [3], and (4) diffusion-based VC model, Dif-fVC1. Following [31], we conduct the naturalness and similar-ity mean opinion score (nMOS and sMOS, respectively). Table 1 depicts that our model has a better nMOS and sMOS than the others. Specifically, our model achieves significantly im-proved content consistency2 and the speaker adaptation performance. In addition, we conducted cross-lingual VC to demonstrate zero-shot conversion performance in unseen languages. Figure 4 shows the robust generalization performance of our model in both resynthesis and VC scenarios, even for unseen languages. Furthermore, we fine-tune the model using only one sample per speaker. Fine-tuning with small steps (1,000 steps) can improve the performance of speaker adaptation. However, the model fine-tuned with more steps shows a lower robustness of content consistency by decreasing the CER and WER

Table 2: *Results of ablation study on zero-shot VC tasks with unseen speakers from VCTK dataset. For all methods, the number of sampling iterations is 6.*

Method	nMOS	sMOS	CER	EER	SECS
Diff-HierVC	**3.86±0.06**	**3.02±0.09**	0.83	**3.29**	**0.861**
Denorm. + DiffVoice	3.81±0.06	3.00±0.10	2.67	5.25	0.850
F_0 Encoder + DiffVoice	3.83±0.06	3.00±0.09	0.89	4.09	0.857
w.o Masked Prior	3.83±0.06	2.91±0.10	0.82	4.52	0.852
w.o Data-driven Prior	3.81±0.06	2.90±0.10	0.56	12.77	0.823
w.o SF Encoder	3.83±0.06	3.01±0.10	0.68	6.75	0.847
DiffPitch + SF Encoder	3.77±0.06	2.95±0.10	**0.30**	5.26	0.854

Table 3: *Results of ablation study on different masking ratio*

Metric	0%	10%	30%	50%	70%	90%
CER (↓)	0.82	0.70	0.83	0.86	0.89	0.96
EER (↓)	4.52	4.55	3.29	3.75	3.74	3.75

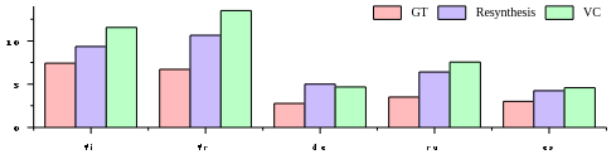


Figure 4: *CER results for zero-shot cross-lingual VC on CSS10*

3.3. Strengths and Limitations

Its hierarchical structure, utilizing DiffPitch and DiffVoice, enhances pitch generation and speaker adaptation, leading to more natural and expressive speech synthesis. The diffusion-based approach allows for more accurate pitch modeling compared to traditional normalization techniques, improving pronunciation and intonation accuracy. Additionally, the masked prior mechanism increases generalization capability, reducing reliance on noisy training data and improving robustness in zero-shot scenarios. Compared to prior models, Diff-HierVC achieves superior speaker similarity and phonetic intelligibility, while maintaining a relatively compact parameter size.

However, weaknesses remain. The reliance on diffusion models increases computational complexity, making inference slower compared to GAN-based or autoencoder-based

Conference Name #1234

VC systems. Additionally, while the model effectively adapts speaker styles, it may capture unwanted noise as part of the speaker's voice characteristics, reducing audio quality in noisy environments. The training process is also more resource-intensive, requiring significant GPU powr and long training times. While it performs well in zero-shot cross-lingual voice conversion, the performance could degrade when fine-tuned on very few samples, leading to potential over fitting to specific speakers.

4. Phoneme Hallucinator: One-shot Voice Conversion via Set Expansion

4.1. Paper

Phoneme Hallucinator [4] is a novel one-shot voice conversion (VC) model designed to achieve high speaker similarity and intelligibility while requiring only a short target speaker sample. The model addresses the key challenge of balancing content intelligibility and speaker similarity by hallucinating target speaker phonemes from as little as a 3-second speech sample. Unlike traditional speaker embedding-based VC systems, which struggle with intelligibility or require large target speaker datasets, Phoneme Hallucinator expands the available target phoneme set using a conditional generative model, thereby improving conversion quality without sacrificing naturalness.

The model architecture follows a neighbor-based voice conversion (VC) pipeline, inspired by kNN-VC. The system first extracts frame-level speech representations from a pre-trained WavLM model, which provides self-supervised phoneme-aware embeddings for both the source and target speakers. These representations are unordered, as temporal alignment is not required for conversion. The Hallucinator module is then applied to the target speech representations, expanding the available target phoneme space by generating synthetic but realistic phoneme features. The expanded target set is then used in a k-nearest neighbors (kNN) regressor, which replaces each source speech frame with the closest neighbor from the hallucinated target phoneme set. Finally, a HiFi-GAN vocoder synthesizes the waveform from the converted speech representations.

The detailed model architecture consists of multiple key components: (1) Feature Extractor, (2) Hallucinator Module, (3) kNN Regressor, and (4) Vocoder. The Feature Extractor uses a pre-trained WavLM-large model, which is a self-supervised speech representation model trained for speech processing tasks such as speaker verification, separation, and recognition. It extracts 1024-dimensional frame-level embeddings every 20 ms from both the source and target speech. The Hallucinator Module is a probabilistic set expansion model built using Set Transformers, which ensures permutation invariance while generating synthetic phoneme representations conditioned on the limited target speech input. The Hallucinator consists of a Conditional Variational Autoencoder (VAE) that samples new target speaker phonemes while maintaining diversity and speaker identity. It is further refined using normalizing flows, ensuring high-quality speaker adaptation. The kNN Regressor replaces each source phoneme embedding with the nearest match from the hallucinated target phoneme set, ensuring phoneme alignment while retaining the source speech's linguistic content. Finally, a HiFi-GAN vocoder converts the generated speech representations into a waveform, synthesizing realistic speech output.

The model is trained using the LibriSpeech dataset, specifically the train-clean-100 split for training and the test-clean split for evaluation. The dataset contains 40 unseen speakers, each with approximately 8 minutes of 16 kHz speech data. During training, speech representations are extracted from WavLM-large, where a 1024-dimensional vector is produced for every 20ms of speech. The Hallucinator is trained by randomly masking phoneme embeddings from the target set and learning to regenerate them using Maximum Likelihood Estimation (MLE). The system is optimized using the Adam optimizer with a learning rate of 0.0001, a batch size of 50, and trained for 250 epochs.

Phoneme Hallucinator is evaluated against state-of-the-art (SOTA) voice conversion baselines, including kNN-VC, FreeVC, VQMIVC, and YourTTS. The evaluation metrics include Word Error Rate (WER) and Character Error Rate (CER) for intelligibility, Equal Error Rate (EER) for speaker similarity, and Mean Opinion Score (MOS) for naturalness. The results show that Phoneme Hallucinator significantly outperforms other models in both intelligibility and speaker similarity in the one-shot setting, where only a 3-second target speaker sample is available. Specifically, the model achieves a WER of 5.10%, which is lower than all baselines, while maintaining a high EER of 44.62%, indicating strong speaker adaptation. Subjective evaluations confirm that the model produces more natural-sounding and intelligible speech compared to existing methods.

Conference Name #1234

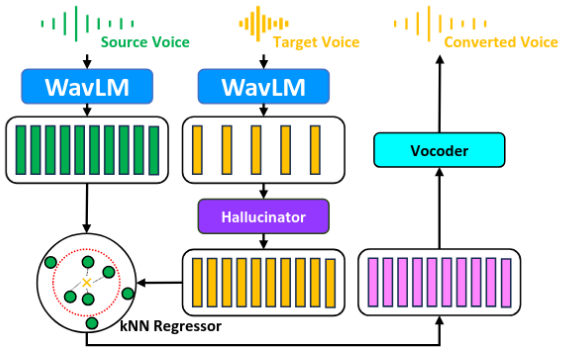


Figure 2: The VC pipeline of our method. A pre-trained WavLM model (Chen et al. 2022) extracts the source representation sequence (green) from the source voice and the target representation set (yellow) from the target voice respectively. Then, the target set is expanded by our hallucinator. Afterward, every source representation is replaced by its neighbors in the expanded target set, resulting in the converted sequence (pink). Finally, a pre-trained vocoder transforms the converted sequence to voice.

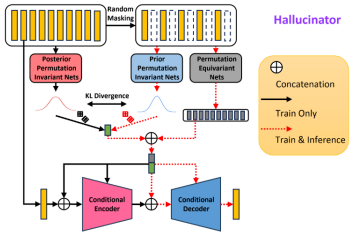


Figure 3: The detailed structure of the hallucinator. Posterior Permutation Invariant Nets, Prior Permutation Invariant Nets, and Permutation Equivariant Nets are all implemented by Set Transformer. Conditional Encoder and Conditional Decoder are implemented by multilayer perceptrons (MLP).

4.2. Results

The paper compares the Phoneme Hallucinator to text-free any-to-any VC methods, such as kNN-VC1 (Baas, van Niek-erk, and Kamper 2023), FreeVC2 (Li, Tu, and Xiao 2023), VQMIVC3 (Wang et al. 2021) and YourTTS4 (in text-free conversion mode) (Casanova et al. 2022) using their official implementations. Among these methods, FreeVC, YourTTS, and VQMIVC use speaker encoders to extract target speaker information for conversion, while kNN-VC adopts the classical concatenative VC approach. FreeVC, YourTTS and VQMIVC respectively use data augmentation, text transcriptions during training, and vector quantization to separate content and speaker information.

4.2.1 Objective Evaluation

To report objective metrics, The paper uses all 40 speakers from LibriSpeech test-clean split and randomly select 5 utterances for each speaker. For each utterance, they convert it to all the utterances of other speakers, resulting in a total of 39,000 (40×5×39×5) converted speech utterances. The paper follows kNN-VC to use Equal Error Rate (EER) to objectively measure speaker similarity. With a trained speaker verification system that computes an x-vector (Snyder et al. 2018) of speech to represent speaker identity, they compute cosine similarities between a pair of x-vectors to measure speech similarity. For all converted utterances, they compute their similarity scores to their corresponding target utterances. Then they compute an equal number of ground-truth similarity scores, which are computed from pairs of two randomly sampled utterances from the same target speaker. Finally, an EER is computed by combining the above two sets of scores, assigning a label of 1 to the ground-truth pairs and 0 to pairs containing converted speech. A higher EER means better speaker similarity and the maximum EER is 50%, meaning that the converted speech is indistinguishable from the real speech from the target speaker. To objectively measure intelligibility, they use an ASR system (Radford et al. 2023) to transcribe the converted speech and compute the word/character error rate (WER/CER). Lower error rates indicate better intelligibility

4.2.2 Subjective Evaluation

They use Amazon Mechanical Turk to subjectively measure naturalness and speaker similarity. For naturalness, we adopt mean opinion scores (MOS) in the range 1–5, where 1 is the lowest perceived quality, and 5 is the highest perceived quality. We randomly sample 50 utterances for each method. For speaker similarity, given a pair of utterances, we ask listeners to judge speaker similarity in the range 1-4. We randomly sample 20 source utterances from the LibriSpeech test-clean split and convert each of them to 4 target speakers. We report the average of this speaker similarity metric and denote it as SIM

#Samples	5k	10k	15k	20k	30k	50k
WER↓ (%)	6.14	5.62	5.35	5.24	5.10	**4.84**
EER↑ (%)	**49.1**	47.2	46.2	44.4	44.6	43.3

Table 2: The influence of the number of hallucinated features on Word Error Rate (WER) and Equal Error Rate (EER), which respectively measure intelligibility and speaker similarity.

4.3. Strengths and Limitations

Its core strength lies in hallucinating high-fidelity phonemes from a short target speech sample, allowing for accurate and expressive voice conversion with as little as 3 seconds of target speech. Unlike prior models, which require extensive target speaker data for high-quality conversion,

Conference Name #1234

this model expands the phoneme set to better match the source speech, improving intelligibility without compromising speaker identity. Additionally, its text-free, any-to-any conversion capability, combined with an efficient neighbor-based approach, results in state-of-the-art (SOTA) performance in both objective and subjective evaluations for intelligibility and speaker similarity.

However, the model has some limitations. While efficient, the hallucination process relies on self-supervised representations, which may introduce subtle artifacts or distortions in highly expressive speech. Additionally, despite outperforming kNN-VC and FreeVC in low-data settings, it still struggles with extreme variations in speaker expressiveness and emotional tone, which could limit its performance in highly expressive VC tasks. The model also leans on a pre-trained vocoder, which, if not optimized for the hallucinated phonemes, might introduce inconsistencies in generated speech.

5. Open Problems and Research Opportunities in Zero-Shot Voice Conversion

1. Zero-Shot VC for Low-Resource and Indian Languages.
2. Handling Noisy and Low-Quality Speech Inputs
3. High Quality Cross-Lingual and Multi-Speaker Adaptation

References

[1] Meiying Chen and Zhiyao Duan. Controlvc: Zero-shot voice conversion with time-varying controls on pitch and rhythm. *CoRR*, 2022. 1

[2] Ha-Yeong Choi, Sang-Hoon Lee, and Seong-Whan Lee. Diff-hiervc: Diffusion-based hierarchical voice conversion with robust pitch generation and masked prior for zero-shot speaker adaptation. *International Speech Communication Association*, pages 2283–2287, 2023. 5

[3] Wonjune Kang, Mark Hasegawa-Johnson, and Deb Roy. End-to-end zero-shot voice conversion with location-variable convolutions. *arXiv preprint arXiv:2205.09784*, 2022. 3

[4] Siyuan Shan, Yang Li, Amartya Banerjee, and Junier B Oliva. Phoneme hallucinator: One-shot voice conversion via set expansion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 14910–14918, 2024. 7