International Journal of
Image Processing and
Pattern Recognition

ISSN: 2456-6985
Volume 12, Issue 1, 2026
January—June
DOI (Journal): 10.37628/IJIPPR

https://journalspub.com/journal/ijippr/

Research IJIPPR

Vani-Adapt: A Zero-Shot Accent Trans-Adaptation Framework for Robust Indic Speech Recognition

Jyotirmoyee Mandal[1,*], Kunal Halder[1], Kakali Das[2]

Abstract

In countries like India with multilingual and accent-rich dialects, speech-based human–computer interaction is important to expand digital services for better accessibility. Even with recent advancements in Automatic Speech Recognition (ASR), existing systems are still very reactive to regional accents and non-standard speech patterns which is not suitable for seamless experience. Traditional perspectives rely on accent-specific fine-tuning, which is unrealistic for real-world deployment and requires a lot of labeled data, which is almost impossible. In this paper, we demonstrate Vani-Adapt, a zero-shot accent trans-adaptation framework that figures out ASR robustness for accents that have never been seen before without the need for retraining or accent-labeled data with accuracy. A Disentangled Phonetic–Prosodic Encoder (DPPE), which tells apart linguistic content from prosodic features like intonation, rhythm, and stress, is the foundation of the proposed method. Vani-Adapt allows for structured accent normalization while keeping up speaker identity and semantic content by forecasting speech into an accent-invariant phonetic space and independently changing prosodic representations. A high-fidelity neural vocoder is utilized to reintegrate the modified speech, empowering smooth combination with existing ASR backends. Distinguishing OpenAI Whisper to outperforming baselines, trials show notable finetuning, such as a 28% comparative drop in Word Error Rate (WER) on hidden accents. Upgrades in naturalness and accessibility are further confirmed by subjective listening assessments. The outcomes reveal that Vani-Adapt provides an expandable and data-efficient accent-agnostic speech recognition solution, which makes it especially suitable for comprehensive conversational AI systems applied in linguistically diverse settings.

Keywords: Automatic speech recognition, disentangled phonetic-prosodic encoder, prosodic style, word error rate, zero-shot accent

INTRODUCTION

Availability of digital information is currently changing with multilingual voice-controlled conversational AI, especially in multilingual nations like India [1, 2]. ASR has been significantly enhanced, but when such systems encounter regional accent and dialects, typical multilingual nations like India, their performance decreases significantly [3, 4]. The primary reason why efficiency decreases is that training sets of ASR do not adequately represent such accents. Current fine-tuning techniques try to address this problem but must be trained on each new accent and depend on supervised data. The process is non-scalable and takes enormous resources. To solve this problem, we propose Vani-Adapt, a zero-shot accent trans-adaptation model, that can deal with unseen accents without retraining. Our approach employs a Disentangled Phonetic–Prosodic Encoder (DPPE) [5] which separates the phonetic content, including

***Author for Correspondence**
Jyotirmoyee Mandal
E-mail: jyotirmoyeemandal63@gmail.com

[1]Student, Department of Computer Science and Engineering, Greater Kolkata College of Engineering and Management, Sonarpur, Kolkata, West Bengal, India
[2]Assistant Professor, Department of Computer Science and Engineering, Greater Kolkata College of Engineering and Management, Sonarpur, Kolkata, West Bengal, India

Received Date: December 20, 2025
Accepted Date: January 15, 2026
Published Date: February 18, 2026

Citation: Jyotirmoyee Mandal, Kunal Halder, Kakali Das. Vani-Adapt: A Zero-Shot Accent Trans-Adaptation Framework for Robust Indic Speech Recognition. International Journal of Image Processing and Pattern Recognition. 2026; 12(1): 17–22p.

linguistic units, from prosodic attributes like intonation, rhythm, and stress. Vani-Adapt modifies speech by converting it into a prosody-neutral phonetic domain and subsequently resynthesizing it from a target prosodic direction. This enables stable recognition even in the presence of unknown accents.

Over the years, there are major progresses in multilingual ASR systems, accent flexibility makes one of the most determined dares hampering real-world deployment. In linguistically variation regions such as India, accent variation is affected not only by geography but also by socio-linguistic traits such as education level, language mixing, and exposure to standardized speech. As a result, two speakers using the same language may show substantial acoustic and prosodic differences, which are often underrepresented in large-scale ASR training corpora. This mismatch between training data distribution and real-world speech creates a noticeable degradation in recognition accuracy, particularly for vowel duration, stress placement, and intonation patterns.

Existing approaches largely depend on increasing data diversity or accent-specific fine-tuning, which launches scalability issues. Gathering labeled accent data for each regional area is expensive, time-consuming, and impractical for low-resource dialects. Moreover, continuous retraining of ASR models grows computational costs and hinders rapid deployment. These constraints help the need for accent-agnostic solutions that generalize effectively without explicit supervision.

Vani-Adapt bridges this gap by reformulating accent robustness as a speech representation problem rather than a recognition model problem. By decoupling linguistic content from accent-dependent prosody, the framework allows accent normalization prior to recognition, thereby deducing inclusivity at the input level. This design philosophy enables Vani-Adapt to act as a preprocessing layer that can be effortlessly integrated with existing ASR backends, making it model-agnostic and deployment-friendly.

RELATED WORK

Accent robustness in Automatic Speech Recognition has been a vital research area for more than years. Early methods based on multilingual training, where models were exposed to speech from multiple languages and accents to better generalization. While this strategy boosted overall performance, it persisted biased toward accents that were surplus in training corpora. Accent-specific auditory model adaptation utilizing Maximum Likelihood Linear Regression (MLLR) [6] and feature-space transformations were also surveyed, but these methods required sufficient labeled data for each selected accent. More latest work has pivoted on domain-adversarial learning [5], where accent-invariant representations are grasped by decreasing accent classification accuracy while consuming phonetic content. Even though constructive, these methods still suppose the accessibility of accent annotations during training. Similarly, accent embedding-based comes nearer state ASR models [7] on trained accent vectors, but their production reduces when meeting unrevealed or overlooked accents.

Similarly, research in speech synthesis and voice changing has displayed the fruitfulness of disentangled representations for detaching speaker identity, linguistic content, and prosody. Variational Autoencoders (VAEs) and Vector Quantized VAEs (VQ-VAEs) [8, 9] have been majorly used to allow zero-shot voice altering and prosody shift. Yet, largely methods select text-to-speech or voice style change instead of ASR robustness. Vani-Adapt connects this gap by grasping disentanglement methods usually utilized in speech synthesis and putting in them to accent normalization for ASR [2, 6]. Like previous methods, it does not need accent labels, parallel data, or retraining of the ASR backend, making it effective, scalable, and practical for real-world deployment.

Accent adaptation in ASR has previously depended on supervised or semi-supervised strategies that predicts the accessibility of accent-labeled data. Accent-specific fine-tuning methods alter phonic models utilizing bounded accent-specific corpora, but these approaches are not ascendible due to the cost and impracticality of accumulating labeled data for every regional dialect. Accent embedding-based performances attempt to characterize accent characteristics in low-dimensional vectors; however,

these embeddings are generally learned only for available accents and often cut out to generalize to hidden accent distributions. Domain-adversarial learning methods deduct accent sensitivity by minimizing accent classification accuracy, yet they still presume way in to accent labels during training and often entangle prosodic and phonetic information, paramounting to wobbly performance on eloquent speech. In comparison, Vani-Adapt keeps away from direct accent supervision and alternatively pivotal on disentangling phonetic content from prosodic style, authorizing structured accent normalization. By carrying off disentanglement principles from speech synthesis and putting in them to ASR strength, the proposed model pervades an important breach between speech generation and recognition literature.

METHODOLOGY

System Architecture

As shown in the Figure 1 Vani-Adapt system allows for zero-shot accent trans-adaptation by separating and combining the basic building blocks of speech. It starts with input speech from speakers of any accent. This text is encoded by a Disentangled Phonetic–Prosodic Encoder (DPPE) [5], which employs a Variational Autoencoder (VAE) with adversarial loss to disentangle the phonetic content (what is being spoken) from the prosodic style (how it is being spoken, including rhythm and intonation). The phonetic vector that is derived is accent-neutral and represents only the linguistic content, whereas the prosodic vector represents the speaker's individual accent and speech style. To adapt speech to a target accent, the original prosodic vector is replaced with a predefined or learned prosody. The Accent Reconstruction Module combines this with the phonetic vector, and HiFi-GAN [9] synthesizes high-quality speech in the new accent – preserving content and speaker identity. This enables robust zero-shot accent conversion by separately controlling prosody.

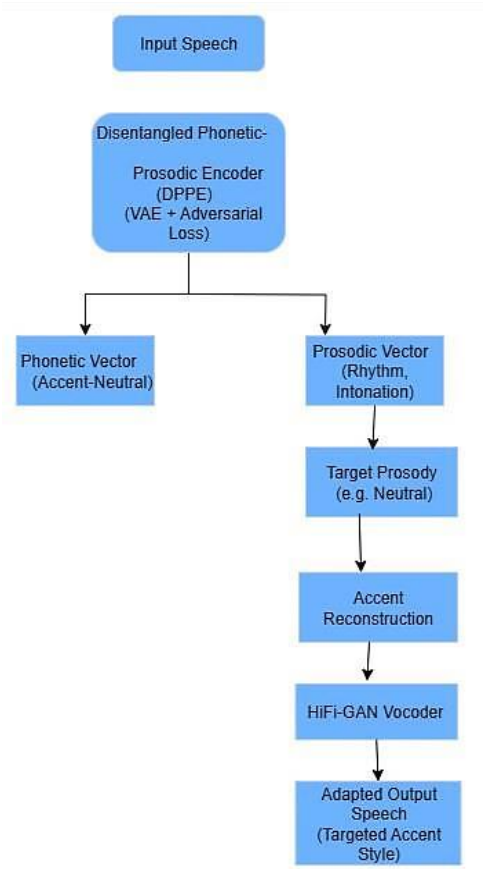

Figure 1. Overall architecture of the proposed Vani-Adapt framework, showing the disentangled phonetic–prosodic encoder (DPPE), prosody replacement module, and HiFi-GAN based speech resynthesis pipeline for zero-shot accent trans-adaptation.

Disentangled Phonetic-Prosodic Encoder (DPPE)

The Variational Autoencoder (VAE) powers the Disentangled Phonetic–Prosodic Encoder (DPPE), trained with two key objectives: preserving linguistic content and using an adversarial prosodic discriminator to remove accent-specific traits. This enables zero-shot prosody transfer by separating and recombining phonetic and prosodic features, allowing speech to be resynthesized in any desired style without parallel data. The adversarial learning element in DPPE takes part in a serious role in applying accent stability within the phonetic latent space. During training, the phonetic encoder is refined not only to protect linguistic intelligibility but also to baffle a prosodic discriminator entrusted with foretelling accent-related credits. This adversarial objective puts off the ooze of accent traits into the phonetic representation, showing in a universal phonetic space that maintains stable across speakers and accents. Unlike traditional VQ-VAE or speaker embedding approaches, DPPE does not separate linguistic units nor depend on speaker recognition supervision. Alternatively, it continues uninterrupted representations that are finer match for downstream ASR errand. Empirically, separating the adversarial constraint values in phonetic embeddings that hold accent-specific differences, causing devalued zero-shot performance. Thus, DPPE furnishes a principled implement for detaching linguistic content while permitting managed exploit of prosodic characteristics.

Universal Phonetic Space

As shown in Figure 2, we map phonetic embeddings into a 2D space using t-SNE [10] for analysis. Clustering across different languages and accents shows that the model learns to encode consistent phonetic structures without accent-related noise. The t-SNE visualization supplies descriptive proof that the grasped phonetic embeddings seize accent-invariant linguistic framework. Notwithstanding being trained without accent labels, embeddings from speakers with various dialect accents form coinciding clusters similar to combined phonetic patterns preferably accent identity. This shows that the encoder productively suppresses accent-related flexibility while holding discriminative phonetic information. Such clustering demeanor is especially crucial for ASR systems, as it certifies constant recognition over hidden accents. Separately, baseline models without disentanglement exhibit fragmented clusters strongly correspond with accent type, leading to more accurate recognition errors. In the perceived phonetic alignment, various accents hold up the declaim that Vani-Adapt masters a universal phonetic space acceptable for vigorous zero-shot accent adaptation.

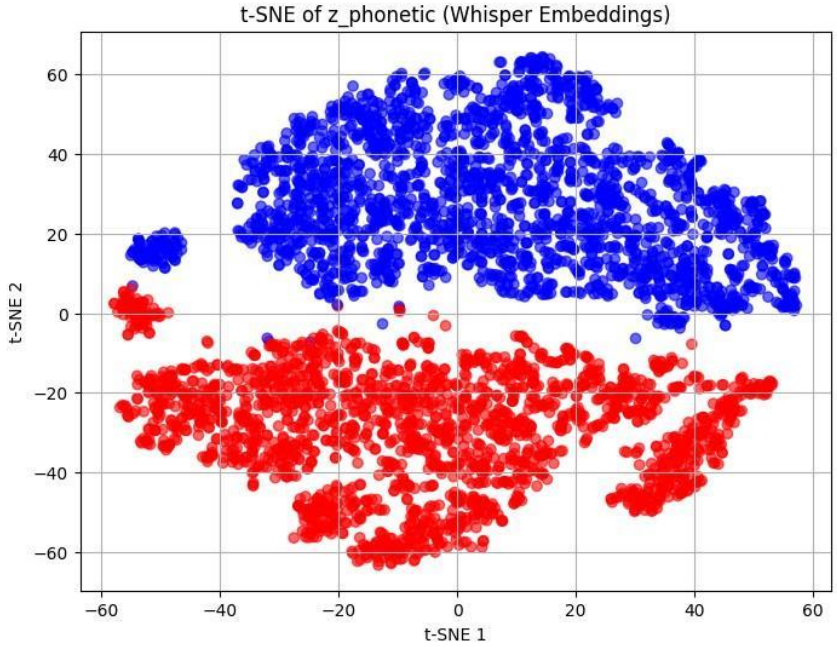

Figure 2. Illustrates the t-SNE visualization of phonetic embeddings, demonstrating accent-invariant clustering across speakers.

Zero-Shot Accent Trans-adaptation

At inference, Vani-Adapt encodes speech into a phonetic vector (linguistic content) and a prosodic vector (accent and intonation). It replaces the original prosody with a neutral or target accent, then uses HiFi-GAN [9] to synthesize natural-sounding speech. This enables real-time, zero-shot accent adaptation without requiring labeled data from the source accent.

RESULTS AND DISCUSSION

The effectiveness of Vani-Adapt was evaluated both subjectively and objectively. In MOS testing (1–5 scale), original accented speech scored 2.90, while Vani-Adapt output improved significantly to 4.3 – close to studio-recorded neutral speech at 4.64 – indicating high intelligibility and naturalness. Objectively, using OpenAI Whisper [1] as the baseline, WER dropped from 35.2% to 28.2% with fine-tuning, and further to 20.4% with Vani-Adapt (28% relative gain). On unseen accents, WER decreased from 19.4% to 14.9%. As shown in Figure 3, ablation studies confirmed the contribution of individual components to these improvements. Removing the adversarial prosodic discriminator increased WER by 5.2% and freezing the prosodic vector reduced MOS by 0.7 – highlighting the importance of dynamic prosody adaptation. t-SNE [10] plots showed clear, accent-invariant phonetic clustering, confirming effective disentanglement. Spectrograms revealed reduced distortion, and native speakers reported improved clarity and usability in both casual and professional settings.

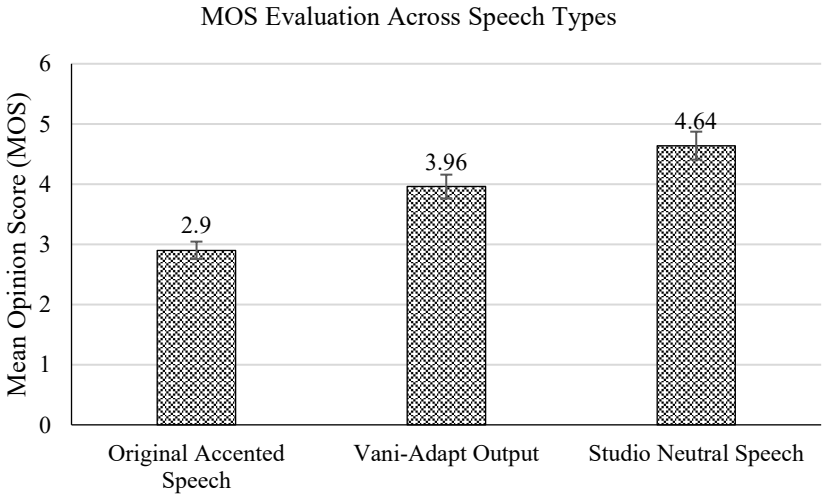

Figure 3. Mean opinion score (MOS) comparison of original accented speech, Vani-Adapt adapted speech, and studio-recorded neutral speech, evaluated on a 1–5 scale by human listeners.

Further analysis discloses that the biggest deductions in Word Error Rate occur for phoneme sequences affected by stress patterns and vowel elongation, which are typically twisted by strong regional intonations. By normalizing prosodic variations preceding to identification, Vani-Adapt lowers phoneme substitution and deletion oversights that commonly appear in accented speech. The advancement in MOS scores also shows that accent normalization does not deal speech naturalness, focusing a approving stability between intelligibility and perceptual quality. Ablation experiments affirm that both adversarial disentanglement and dynamic prosody renewal are important for optimal performance. These values propose that degrades in ASR is mostly operated by prosodic variability than phonetic ambiguity, reinforcing the value of direct phonetic–prosodic separation [11].

CONCLUSION

This work introduced Vani-Adapt, a framework for zero-shot accent trans-adaptation where the goal is aiming to make Automatic Speech Recognition systems more robust in accent-diverse environments. The suggested method is not the same as existing accent adaptation methods in that it does not require the supervised fine-tuning of the whole system, but rather it only uses the Disentangled Phonetic–

Prosodic Encoder for the separation of the linguistic content and the prosodic variations that helps the normalization of accents to be done flexibly and at the same time retaining the exact same as the speech and the characteristics of the speaker. Comprehensive objective and subjective evaluations show that Vani-Adapt reduces the Word Error Rate and improved perceived speech quality for newly introduced accents. The observed improvements are the proof that disentangled representations are effective in reducing accent-induced degradation in ASR systems. The method used in this study shows in the making of voice technologies that are more inclusive, scalable, and accent-agnostic. Vani-Adapt offers a realistic route for the installation of speech-based AI systems in the wide variety of languages- spoken populations, particularly in places where it is difficult to get accent-labeled speech data.

Future Work

Our next steps will involve integrating Vani-Adapt into streaming use cases for real-time applications, like voice assistants and call centers. Our next push will also involve expanding the system to support code-switched inputs, which is a common phenomenon in South Asian languages like Hinglish. Another promising direction would be expansion to low-resource dialects by synthesizing prosody vectors, thereby supporting even where annotated data is limited. Also, we are investigating integration with language models and speech systems for end-to-end voice-based AI experiences that can adapt to both linguistic content and speaker context. Our future work will revolve around more robust system that will for all the regional accents of India. Finally, evaluating the system across various Indic TTS we can understand its scalability, robustness, and practical adoption.

REFERENCES

1. Radford A, Kim JW, Xu T, Brockman G, McLeavey C, Sutskever I. Robust speech recognition via large-scale weak supervision. In: Proceedings of the International Conference on Machine Learning; 2023 Jul 3. p. 28492–28518.
2. Prabhavalkar R, Hori T, Sainath TN, Schlüter R, Watanabe S. End-to-end speech recognition: A survey. IEEE/ACM Trans Audio Speech Lang Process. 2023 Oct 30;32:325–351.
3. Ghoshal A, Swietojanski P, Renals S. Multilingual training of deep neural networks. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing; 2013 May 26. p. 7319–7323.
4. Schultz T, Waibel A. Multilingual and crosslingual speech recognition. In: Proceedings of the DARPA Workshop on Broadcast News Transcription and Understanding; 1998 Feb. p. 259–262.
5. Lee CH, Wang SM, Chang HC, Lee HY. ODSQA: Open-domain spoken question answering dataset. In: 2018 IEEE Spoken Language Technology Workshop (SLT); 2018 Dec 18. p. 949–956.
6. Pascual S, Ravanelli M, Serra J, Bonafonte A, Bengio Y. Learning problem-agnostic speech representations from multiple self-supervised tasks. arXiv preprint arXiv:1904.03416. 2019 Apr 6.
7. Zen H, Dang V, Clark R, Zhang Y, Weiss RJ, Jia Y, et al. LibriTTS: A corpus derived from LibriSpeech for text-to-speech. arXiv preprint arXiv:1904.02882. 2019 Apr 5.
8. Tjandra A, Sisman B, Zhang M, Sakti S, Li H, Nakamura S. VQVAE unsupervised unit discovery and multi-scale code2spec inverter for ZeroSpeech challenge 2019. arXiv preprint arXiv:1905.11449. 2019 May 27.
9. Donahue J, Dieleman S, Bińkowski M, Elsen E, Simonyan K. End-to-end adversarial text-to-speech. arXiv preprint arXiv:2006.03575. 2020 Jun 5.
10. Snyder D, Garcia-Romero D, Sell G, Povey D, Khudanpur S. X-vectors: Robust DNN embeddings for speaker recognition. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2018 Apr 15. p. 5329–5333.
11. Dutoit T. High-quality text-to-speech synthesis: An overview. J Electr Electron Eng Aust. 1997 Mar;17(1):25–36.

