Hitachi Review

January 2025 Technical Information

Innovation for Inclusive Fintech

Speech AI for Digital Systems and Services

#Innovation Creation #Generative AI #IoT/Data Utilization #Digital Solutions

Author

Mayurakshi Mukherji
R&D Centre, Hitachi India Pvt. Ltd.
Current work and research: Speech and Acoustics team activities.
Society memberships: IEEE.

Shreyas Kulkarni
R&D Centre, Hitachi India Pvt. Ltd.
Current work and research: Speech AI, Generative AI and Optimization activities.

Vivek Kumar
R&D Centre, Hitachi India Pvt. Ltd.
Current work and research: Generative AI, NLP, Cloud Computing, and full stack development related projects.

Rahul Mishra, Ph.D.
R&D Centre, Hitachi India Pvt. Ltd.
Current work and research: Speech and Acoustics team activities.

Munender Varshney, Ph.D.
R&D Centre, Hitachi India Pvt. Ltd.
Current work and research: Speech and Acoustics team activities.
Society memberships: IEEE.

Thiruvengadam Samon
R&D Centre, Hitachi India Pvt. Ltd.
Current work and research: Artificial Intelligence Solutions Architect focusing on fintech and sustainable mobility solutions.

Senthil Raja G, Ph.D.
R&D Centre, Hitachi India Pvt. Ltd.
Current work and research: Speech and Acoustics team activities.
Society memberships: IEEE.

Kingshuk Banerjee, Ph.D.
R&D Centre, Hitachi India Pvt. Ltd.
Current work and research: Director of HIL R&D Centre.

Highlight

The Digital India initiative by the Government of India, combined with smartphone penetration in all sections of society, has led to the need for AI in building inclusive digital systems for critical public services like finance, healthcare and agriculture. The R&D team in Hitachi India is actively engaged in researching speech (the most natural medium of communication) as a means to democratize access to those essential public services, particularly in banking.

This article discusses three critical aspects of speech-based financial inclusion - voice-based authentication, vernacular speech recognition, and embedding of automated speaker verification in edge devices. This latter is helpful for spilt-second inferencing in high-latency and low-bandwidth situations where a miniaturized speech comprehension engine is available locally in a smartphone app, which may even be offline.

We also discuss the development of our AI model architecture from scratch with focus on understanding of Indian vernacular languages; appropriate quantization of the neural network model for miniaturized footprint; and the challenges related to accuracy. Finally, it touches upon the company's approach of building a relevant dataset, focusing on connected numbers, a key need for "financial amount" articulation in speech-based transactions.

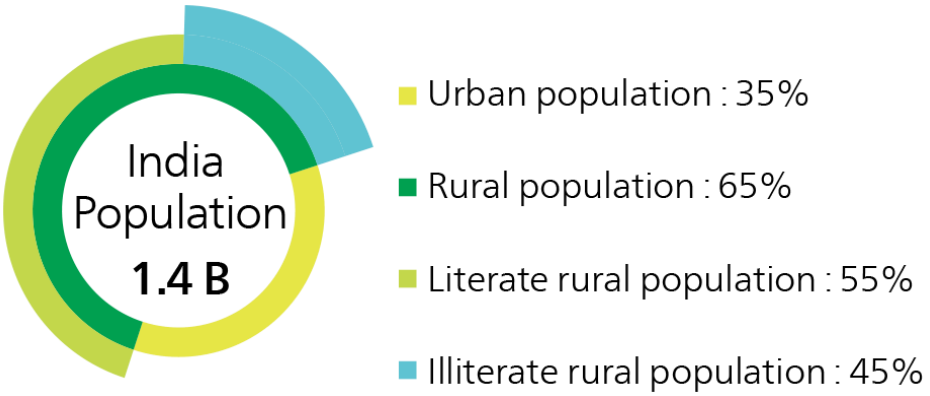
1. Introduction

In recent years, India has seen widespread adoption of digital payment systems such as Unified Payments Interface (UPI) QR Code*1 and e-wallets, with smartphone apps making it easier for citizens to transfer money, pay bills, and make purchases, without having to visit a bank or use cash. However, as

represented by the chart in Figure 1, India has a significant rural population, many of whom face challenges with literacy. Therefore, it is essential to introduce speech-based assistance in native languages integrated into banking applications to make digital banking and finance truly inclusive under the Digital India initiative.

*1 QR Code is a registered trademark of Denso Wave Inc.

Figure 1. Distribution of Indian Population



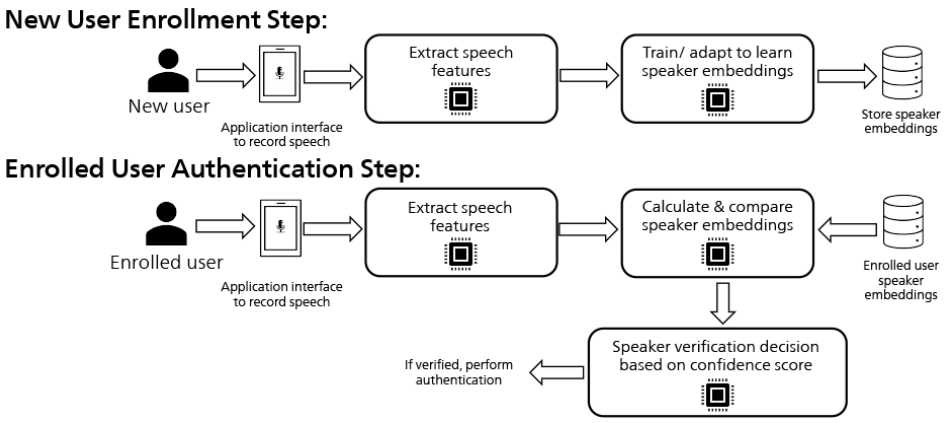
A significant portion of the Indian population is illiterate or semi literate. Inclusive fintech is especially beneficial for them as speech commands in vernacular languages will help them access banking services and apps using speech, the most natural form of communication. Additionally, a large portion of the Indian population is also located in rural areas where there is limited bandwidth, thus posing a restriction on the amount of information that can be exchanged over the internet. This requires deployment of speech-based solutions on edge devices (smartphones). Our vernacular language automatic speech recognition (ASR) and edge-based automatic speaker verification (ASV) support the mission of inclusive fintech in India.

A combination of AI technologies—speech recognition, speaker verification and identification, and text-to-speech technology—can be used to build efficient and smooth banking experiences as described in sections 2 and 3 of this article. However, AI models are computationally expensive and require high network bandwidth when deployed in cloud APIs, which limits the user experience. AI-on-edge deployment addresses this issue by exploiting the native computational resources of smartphones, which reduces AI system latency significantly. Further, it improves data security and privacy as described in section 4 of this article.

2. Voice-Based Authentication

Voice authentication (VA) typically consists of two steps namely, enrollment and authentication (see Figure 2). In the enrollment step, users register their voice by providing an audio sample to the VA system. The AI model converts the audio sample to a user-specific voice signature, which is subsequently stored by the system. During authentication, users provide voice samples that are compared against their stored voice signatures and a similarity score is generated to quantify how similar or dissimilar the voice is compared to the stored voice signature of the user. If the similarity score is higher than the predetermined similarity score threshold value, the user is authenticated.

Figure 2. ASV Pipeline Involving Enrollment Step and Authentication Step



ASV includes two major steps - enrollment and authentication. During enrollment, users register their voice by submitting a voice sample which is used for extracting speech features. This is done by calculating speaker embeddings (i.e., the voice signatures). The embedding calculated at the time of enrollment is stored in the system. The calculation of speaker embeddings is done using a ResNet34 model trained on speech samples of multiple speakers. During authentication

when the enrolled user submits a new voice sample, another embedding is calculated using the same ResNet34 model and compared with the stored speaker embedding of the user. If the comparison score (i.e., confidence score) is higher than a threshold value, the user is authenticated.

VA has applications in a variety of domains; however, our focus has been on its applications in banking apps. In retail-banking apps, VA can be used to substitute for the log-in password, to transfer small payment amounts, and to check transaction records. In India, where UPI payments using QR Codes prevail, most merchants choose to display their banking account QR Codes to customers at the time of payment. There are two types of QR Codes, (1) static QR Codes, which only carry the payee banking account information, where the customer must enter the payment amount themselves and (2) dynamic QR Codes, which carry the payee banking account information and the payment amount, where the customer does not have to enter the payment amount. VA combined with automatic speech recognition (ASR) can help make payments faster by allowing merchants to use voice commands such as “generate a QR Code of Rs. 300,” to authenticate themselves, and subsequently generate dynamic QR Codes to collect payments. This simplifies the process by executing it in a single step.

Faster payment for reduced user friction implies shorter audio samples at the time of authentication. However, short audio samples carry comparatively less user-specific voice information, resulting in poor accuracy. This can be mitigated using domain-specific audio samples during the enrollment step to create larger areas of overlap between the stored voice signature and the authentication audio samples. Using the open-source automatic speaker verification model WeSpeaker[1] based on the ResNet34 architecture[2], we improve the results on use-case authentication audio by carefully designing the textual content of the enrollment audio samples. Additionally, a tunable similarity score threshold based on the use-case requirements, i.e., frictionless or high-security, can be used to improve use-case-specific accuracy. Improved performance can also be achieved on test datasets comprised of use-case specific audio samples by fine-tuning the base WeSpeaker model. The finetuning is performed using 2-3-second connected number audio samples in Hindi language. An Equal Error Rate (EER) of 1.96% has been achieved using the finetuned models, compared to an EER of 4.2% of the base WeSpeaker models when tested on a use-case specific dataset of 2-second Hindi audio samples.

Currently, the use of VA for banking use-cases is limited to phone-banking and IVR systems that collect multiple sets of longer audio samples from users to generate user voice signatures along with longer voice samples at the time of authentication. This leads to high friction and reduced adoption of digital banking using voice technology. Hence, highly accurate VA systems with short audio samples are the need of the hour to implement frictionless Speech AI fintech solutions.

3. Vernacular-Language Connected Number Speech Recognition

Connected Number Recognition (CNR) is a subset of ASR that focuses on recognizing sequences of numbers spoken in continuous speech. For example, the Tamil phrase *Toḷāyirattu aintu* corresponds to the number 905. Connected number speech recognition is becoming increasingly important in Indian banking because it simplifies user interactions. Instead of navigating through complex menu systems, customers can complete transactions and inquiries using natural speech, enhancing convenience. For this system to reach a larger area in India, it must be able to function in local languages. Enabling users to interact in their native language not only improves accessibility, but also promotes cultural preservation, making the technology more inclusive. It also supports education and economic opportunities by allowing more people to engage with digital services, bridging communication gaps in this growing digital age. This research is motivated by the frequent need to recognize connected numbers in sectors like finance, where accuracy in speech recognition is crucial.

In the company’s research, it analyzed the performance of existing state-of-the art ASR models on the task of CNR in Tamil and Hindi. It found that all the models exhibited significant performance degradation when handling connected numbers in these languages. To create a relevant dataset, it collected speech samples in Tamil from various districts across Tamil Nadu and in Hindi from multiple northern states of India. Participants were randomly shown numbers and asked to speak those numbers as they would in everyday situations, and these recordings were used for training and testing the models. Such samples have been collected in different languages from ten regions in India.

The main model, LSTM-TDNN (LT-Kaldi), is part of the Kaldi Speech Toolkit. It consists of six convolutional layers and 15 time-delay neural networks, with 31 million parameters[3]. The model was trained using high-resolution MFCCs (features extracted from speech samples) with cepstral mean normalization, following Kaldi’s standard training recipe. Additionally, the company experimented with fine-tuned versions of two pre-trained models—Wav2Vec2.0[4] and Whisper[5]—on publicly available datasets for Tamil and Hindi.

In terms of performance, the baseline LT-Kaldi model achieved a Word Error Rate (WER) of 15% for Tamil and 7% for Hindi. Preprocessing methods like spectral gating, spectral subtraction, and diarization were also explored, but these led to only minor improvements (2% for Tamil and 1% for Hindi) [6].

Comparing LT-Kaldi to Wav2Vec2.0 and Whisper, it was found that Wav2Vec2.0 had WERs of 98% for Tamil and 71% for Hindi, while Whisper exhibited WERs of 93% and 85%, respectively. These results suggest that while LT-Kaldi is a solid baseline for CNR, state-of-the-art models like Wav2Vec2.0 and Whisper require additional fine-tuning on specific datasets to improve their performance for connected number recognition tasks.

4. Voice-Based AI on Edge Devices

Voice-based AI solutions on edge devices are key to inclusive fintech due to their ability to process speech data locally, enabling fast, secure, and accessible services. By moving speech processing to the edge rather than relying on the cloud, these solutions reduce latency, minimize bandwidth use, and enhance privacy. This is particularly valuable in financial services, where quick, accurate responses and data security are critical. Edge-based processing also enables access in regions with limited internet, helping to bridge digital divides.

AI-based solutions on the edge include metrics such as latency, accuracy, and model size. In the case of voice-based AI solutions on edge, low latency is essential for real-time responses, especially in transactions and service inquiries, while high accuracy ensures reliable understanding across various accents and environments. A smaller model size allows it to fit within the limited storage of a smartphone or embedded systems. Together, these metrics help assess whether a model is practical and efficient for edge deployment. Real-world applications show the promise of voice AI on the edge. For example, virtual assistants on smartphones—Google Assistant*2 and Apple's Siri*3 — process common commands on the device to reduce the need for cloud offloading.

In the course of its research, the company focuses on applying the concepts of the edge implementation of AI models on VA models discussed in section 2. Primarily, it explores the technique known as quantization of models to achieve high accuracy, low-latency, and reduced model-size. Quantization as shown in Figure 3, reduces the bit-width of the model's weights, reducing both the model's size and inference latency[7]. However, there is a trade-off relationship in the case of accuracy. On quantizing the original 32-bit model to an 8-bit model, the WeSpeaker model shows a four-fold reduction in model size, two-fold reduction in inference latency and 2% increase in EER when evaluated on an edge device, i.e., Raspberry Pi 4*4. The company uses static quantization, a technique where both the model weights and model activations are quantized using a small representative dataset known as calibration dataset. To mitigate the issue of increasing accuracy, it proposes finetuning the original model using domain specific data such that the increase in EER can be compensated for in the case of quantization.

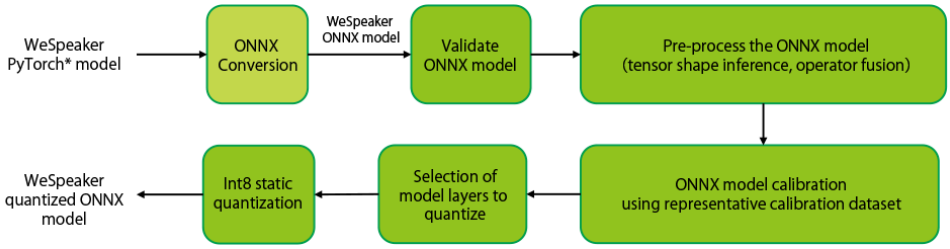
Additionally, it uses the ONNX (Open Neural Network Exchange) format for deployment. ONNX supports compatibility across various frameworks, making it ideal for edge environments where different devices and hardware constraints are common[8]. Using the ONNX format, it is possible to explore the deployment of the company's quantized VA solution on Android*2 smartphones. This combination of edge processing, efficient model formats, and device compatibility makes voice AI a strong tool for accessible, inclusive fintech.

*2 Google Assistant and Android are registered trademarks or trademarks of Google LLC in the USA and other countries.

*3 Siri is a trademark of Apple Inc., registered in the United States and other countries.

*4 Raspberry Pi is a trademark of the Raspberry Pi Foundation.

Figure 3. Pipeline for Building Quantized ONNX Model for Opensource RawNet3 PyTorch Model



* PyTorch is a registered trademark or trademark of The Linux Foundation in the United States and other countries.

An opensource PyTorch model is converted to ONNX format and validated for functionality. This ONNX model is pre-processed by fusing model operators and inferring the model tensor shapes. The pre-processed model is calibrated with a representative dataset and the model layers to be quantized are selected. Finally model quantization is performed to reduce its size to 1/4th of the original size. The model's inference latency is also reduced when model is implemented on edge devices such as Raspberry Pi 4 and smartphones.

5. Conclusions

Inclusive finance, powered by speech-based AI, combines voice authentication, connected number speech recognition, and edge-based voice AI solutions to enhance financial access. Voice authentication offers secure, seamless access; connected number speech recognition simplifies interactions for users with low literacy or limited digital experience; and edge processing allows services to function reliably in connectivity-limited areas. Together, these technologies create a more accessible financial ecosystem that caters to diverse user needs.

Key challenges remain, particularly in enhancing quantization accuracy and optimizing model deployment on some devices, where limited computational resources affect performance. Inclusive finance initiatives benefit end-users, especially in underserved regions, while fintech companies gain access to a broader market and demonstrate leadership in financial innovation. As these technologies advance, fintech firms have strong incentives to pursue these solutions to reach wider audiences and establish a more inclusive digital economy.

REFERENCES

1) H. Wang et al., "WeSpeaker: A Research and Production Oriented Speaker Embedding Learning Toolkit," ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, pp. 1-5 (2023).
2) He, K. et al., “Deep residual learning for image recognition,” 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016).
3) D. Povey et al., “The Kaldi speech recognition toolkit,” IEEE 2011 Workshop on Automatic Speech Recognition and Understanding (2011).
4) A Baevski, et al., “wav2vec 2.0: A framework for self-supervised learning of speech representations,” Advances in neural information processing systems (2020).
5) A. Radford, et al, "Robust speech recognition via large-scale weak supervision." In International conference on machine learning, pp. 28492-28518. PMLR (2023).
6) R. Mishra, et. al., “Revisiting Automatic Speech Recognition for Tamil and Hindi Connected Number Recognition,” In Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages, pp. 116-123 (2023).
7) Quantize ONNX Models
8) Open Neural Network Exchange

Hitachi Review

Hitachi Review is a technical medium that reports on Hitachi's use of innovation to address the challenges facing society.

The *Hitachi Review* website contains technical papers written by Hitachi engineers and researchers, special articles such as discussions or interviews, and back numbers.

Hitachi Hyoron (Japanese) website	*Hitachi Review* (English) website
https://www.hitachihyoron.com/jp/	https://www.hitachihyoron.com/rev/

Hitachi Review Newsletter

Hitachi Review newsletter delivers the latest information about Hitachi Review when new articles are released.