IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, June 2026
Impact Factor: 8.2

Speaker-Conditioned Neural Speech Synthesis for Marathi and English Using Guided Attention on the vVISWa Dataset

Diksha R. Pawar[1] and Dr. Pravin L. Yannawar[2]
Senior Research scholar, Department of Computer Science and IT[1],
Professor, Department of Computer Science and IT[2],
Dr. Babasaheb Ambedkar Marathwada University, Chh. Sambhaji Nagar, Maharashtra, India

Abstract: *End-to-end neural models of text-to-speech synthesis have enhanced significantly, however, there is a challenge of producing high quality natural speech even in low resource languages, especially in the bilingual English-Marathi models prevalent in India. In this paper, an end-to-end neural TTS system with explicit modules that are designed with English and Marathi in mind is introduced Speech Synthesis model, which is based on the Tacotron 2 framework. This research use guided attention mechanisms to provide strong text to acoustic alignment and fast convergence. Multi-speaker synthesis and voice cloning is made possible through a dedicated speaker encoder that extracts compact embedding's during reference audio and results in multi-speaker synthesis. The process of a front-ends done in language-specific front-end module converts the grapheme into phonemes, then retrieves prosodic information, and handles the code-switching effects specific to the phonology of Marathi speakers, including aspirated and retroflex consonants, deletion of schwa, and harmony of vowels. It is trained on vVISWa multilingual audio-visual corpus that consists of paired text-audio samples of 58 speakers in various poses hence facilitating cross-lingual transfer in low-data conditions. Mel-Cepstral Distortion (MCD) objective assessment scores 5.8dB in English and 7.1dB in Marathi, which is better than the Tacotron 2 6.5dB in English and Hidden Markov Model systems 8.2dB, respectively. Subjective Mean Opinion Scores (MOS) are 4.2 with English, which is close to the quality of a human being, and 3.6 with Marathi, which is a great improvement over baselines, thus supporting better naturalness, prosody and speaker similarity. The system is shown to be compatible seamlessly with downstream audio visual systems, such as generation of talking faces. The article introduces the initial bilingual neural TTS framework of the English-Marathi combination with full speaker conditioning, overcoming the main shortcomings of low-resource Indic languages and taking the inclusive speech technologies in multilingual settings the next step.*

Keywords: Text-to-speech synthesis, bilingual TTS, English-Marathi, low resource Indic language.

I. INTRODUCTION

Text-to-speech (TTS) synthesis is the automatic synthesis of speech in the form of human-like spoken audio using text, a methodology which has found itself inessential in modern human-computer interface. TTS visual-impaired, reading-challenged, or speech-impaired people can communicate more reasonably, and it focuses behind application software (and a player), audiobooks, customer support robots and upcoming uses: lifelike talking-face avatars [1]. Particularly in multilingual nations like India, the highest quality of a balanced two-party TTS to support not only English but also, regional languages like Marathi are highly appreciated in the wake of inclusive digital services, e-governance, media accessibility and creation of audio visual content across languages wherein voice identity should be consistent with the lips reading synchronization [2].

DOI: 10.48175/IJARSCT-36889

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 7, June 2026 Impact Factor: 8.2

The history of speech synthesis started with mechanical synthesis in the eighteenth century and went on to an electrical analog system (e.g., VODER, 1939) and early computerized formant synthesizers in the 1970s -1980s [3]. The 1990s and early 2000s saw the prevalence of concatenative synthesis followed by statistical parametric models which were based on Hidden Markov Models (HMMs). A significant change of paradigm was with the development of deep neural networks, beginning with autoregressive waveform modelling by WaveNet in 2016 [4] and followed by end-to-end sequence-to-sequence models such as Tacotron (2017) and Tacotron 2 (2018). These are models that directly do text-Mel-spectrograms correspondence plus matching them to neural vocoders to achieve near-human naturalness [1]. Nevertheless, modern neural TTS systems are still facing severe problems, particularly in low resource languages.

Marathi has complicated phonetics, such as aspirate consonants, retroflex consonants, schwa deletion, vowel harmony, unusual prosody, but is usually trained using languages with high resources such as English and gives surprisingly high performance [1]. Other drawbacks involve the absence of high-quality paired text-audio datasets, inaccurate text-to-acoustic alignment (resulting into gaps or repetitions), code-mixing between English and Marathi, inadequate speaker variability and expressive prosody modelling, as well as, high inference latency of real-time systems [2].

In the current article, the limitations have been solved by designing a modular end-to-end bilingual neural TTS model specific to English and Marathi. This research extends Tacotron 2 [5] with guided attention to achieve better alignment and quicker convergence [1], a separate encoder of the speaker to synthesize through multiple speakers and voice clone, and language-specific front-end processing (Matching Marathi phonology and code-mixed speech) to Marathi phonology. As opposed to other existing TTS models, which are either English-based or need to be individually trained per language, our model uses the bilingual datasets, facilitating the cross-lingual transfer and achieving superior objective (Mel-Cepstral Distortion) and subjective (Mean Opinion Score) results than the commonly used HMM-based Tacotron, and plain Tacotron models.

Its key contributions are:

(i) The first to consider the language pair, English-Marathi, with the support of fully speaker-conditioning neural TTS;
(ii) Better addressing of the speaker-specific phonetic and prosodic issues; and
(iii) Full integration into downstream audio visual tasks, such as speaking-face generation.

II. LITERATURE REVIEW

Over the last decades, text-to-speech (TTS) synthesis has changed significantly as it adopted rule-based and concatenative methods to complex end-to-end neural networks, which could synthesize speech of almost human naturalness. The rudimentary systems have depended on manual rule sets or unit picking up, thus having robotic melodies as well as the lack of flexibility [3]. Waveform-level modelling After 2016 the deep-learning revolution introduced waveform-level models, such as the WaveNet [4], and sequence-to-sequence models, such as the Tacotron family, which directly model the mapping of text to acoustic features [1]. Low-resource languages, particularly Indic languages like Marathi, do not have the data (resource) and mature models that high-resource languages like English have, and still experience issues of phonetic complexity, prosodic richness, inadequate high-quality data, and lack of speaker diversity. There are many studies that have worked on TTS of Indian languages in the form of corpus development, English centric model adaptation, or language specific front-end development [6]. However, really bilingual systems, in which English and a local language like Marathi both process English words with powerful speaker conditioning and alignment power, are quite scarce [7]. An overview of some of the influential and relevant works is provided in the table below, their contribution, the target language and some of their main limitations which this proposed bilingual neural TTS aims to address should be highlighted in the Table No.1 below.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, June 2026
Impact Factor: 8.2

Table No.1. The Overview of the Main Literature about the Text-to-speech Synthesis.

Year	Work / Model	Key Contribution / Approach	Language(s) / Focus	Limitations / Gaps Addressed in Our Work	Reference
1987	Review of text-to-speech conversion	Comprehensive overview of early formant and rule-based TTS systems	Primarily English	Limited naturalness; no neural methods	[3]
2016	WaveNet	First high-fidelity neural vocoder using autoregressive CNNs for waveform generation	English (generalizable)	High computational cost; paved way for end-to-end neural TTS	[4]
2018	Tacotron 2	End-to-end sequence-to-sequence model + neural vocoder for near-human quality	English (high-resource)	Single-speaker focus; poor handling of low-resource languages like Marathi	[5]
2021	A survey on neural speech synthesis	Broad review of neural TTS components, multi-speaker, and low-resource challenges	Multilingual	Highlights data scarcity and prosody issues in non-English languages	[1]
2020	IndicSpeech	Large-scale paired TTS corpus for multiple Indian languages	Hindi, Tamil, Bengali, Marathi, etc.	Limited speaker variety and prosody in low-resource setups	[6]
2022	Text-to-Speech Synthesis: Literature Review (Indian focus)	Amalgamates major TTS works in English and Indian languages	English + Indian (incl. Marathi)	Scarce high-quality neural models for Indic phonology	[7]
2025	A2TTS	Speaker-conditioned diffusion-based TTS with zero-shot adaptation	Multiple Indic languages	Improves unseen speaker synthesis but still data-hungry for Marathi prosody	[8]

III. METHODOLOGY OF SPEECH-SYNTHESIS MODEL

The current Speech-Synthesis produce speech of high quality based on the text on the target languages of Marathi and English. More recent models based on neural networks have facilitated the generation of more dialogue through mapping character sequences to Mel spectrograms. Such that, the current analysis follows the neural TTS model in the example of [5] and adds guided attention to it in order to produce high-quality matching and faster convergence [1]. The entire system consists of five major modules: dataset and input representation, text processing, encoder-attention-decoder network, speaker encoder and waveform production module (vocoder) as indicated in Fig.1. Each element and communication between them are discussed below as per the diagram. The dataset is assumed at the beginning of the process and it represents the paired text and speech samples. Based on this data, textual information (e.g., This is cat) and the speech signal are the input of different subdivisions of the system. The speaker related information is extracted using the speech signal, and the text input is used to generate linguistic representations.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, June 2026
Impact Factor: 8.2

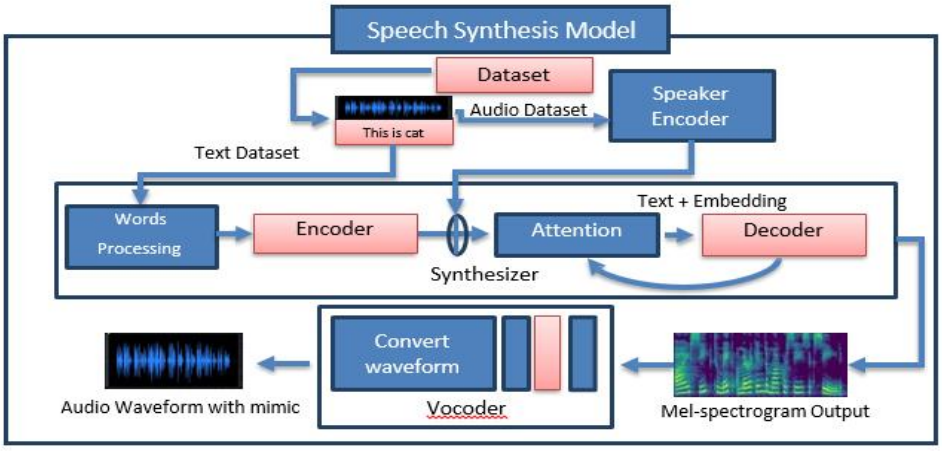


Fig.1. Speech Synthesis Model

3.1 Speaker Encoder

Firstly, the reference speech waveform is sent to the Speaker Encoder which serves to generate a small representation, known as a speaker embedding, which captures the identity of the speaker and his or her vocal attributes. The Speaker Encoder is a critical element in multi-speaker or voice-cloning text-to-speach (TTS) systems and is a component that is expected to capture and encode the unique vocal identity of a specific speaker based on reference most of the audio samples. It does not rely on the main text-processing pipeline and also conditions the synthesis process to produce speech with certain voice qualities so that the synthesis is able to be used to produce speech with varying voices without the system having to re-train itself when used with a new speaker. Raw audio waveforms (or a representation of them in Mel-spectrogram form) are typically input to the encoder of the speaker, and the encoder returns a fixed-dimensional embedding or, more commonly called a d-vector or x-vector or speaker embedding. This speaker-specific feature in this vector, timbre, pitch range, accent, rate of speaking and prosodic issues, is independent of the linguistic content being spoken. A classic implementation, which is based on inspiration of speaker verification architectures such as those of Google, based on a deep neural network, is known as SV2TTS [10]. The reference speech waveform, along with features obtained with the help of the Audio Encoder in Section 3.1, are fed through a three-layer bidirectional long short term memory (LSTM) network with 256 hidden units in each layer. The neural network combines the series of convolutional neural network outputs to learn temporal variations, including time-varying intonation. Afterwards, the sequential representations are summarized by another(global) pooling layer into a single vector, which is then projected to a 256512 dimensional space by a fully connected layer; L2 normalization provides the final speaker embedding. This embedding is then provided to the main synthesis network where it conditions the synthesized speech on the target voice characteristics as shown in Fig.2.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, June 2026
Impact Factor: 8.2

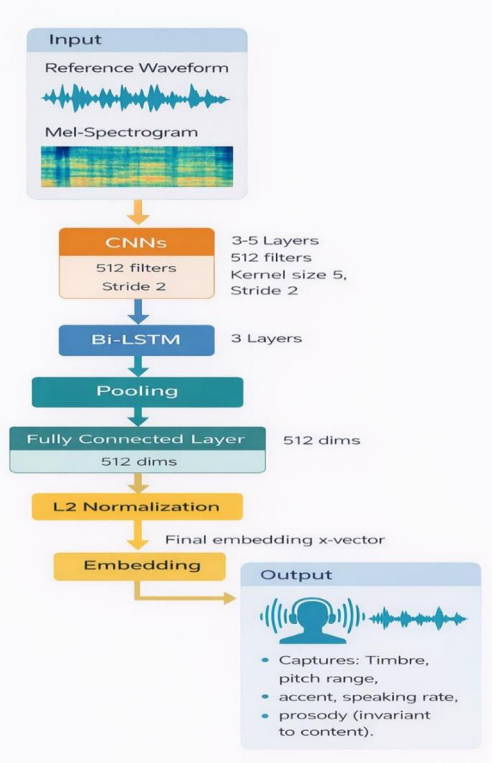


Fig.2. Speaker Encoder

3.1.1 Integration:

The embedding is added into the decoder by concatenation or addition by projection layers to either the hidden states of the decoder or to the decoder attention context. This incorporation affects the Mel spectrogram created, guiding it towards the target voice desirable. This design is modular to support multi-speaker synthesis such that it can be trained on a heterogeneous set of speakers and selective embedding type can be chosen at run-time. Voice mimicry, especially of audio waveforms that recreate that of a particular speaker, can therefore be allowed to be done with high fidelity, so it is not necessary to have per-speaker models. In addition, the embedding's themselves are low dimensional, making them easier to store and use on call, without calling on recomposition.

3.2 Words Processing

At the same time, the text that is being inputted is processed by the Words Processing bit The module forms the first step or textual analysis phase in front end, in which raw textual input is converted into a 2D representation, easily readable by the machine, which can further assert itself in the neural encoder. The transformation deals with the variability and ambiguity of natural language, producing a sequence of linguistic features, including phonemes or characters, and to the model is able to interpret efficiently. It is a process that includes a set of in-built sub-steps that are adapted to particular languages like English and Marathi [5].

- ***Text Normalisation (TN):*** This step transforms non-orthodox textual documentation into an understandable format. As an example: Numbers: "123" to Hundred and twenty-three. Abbreviations: "Dr. to doctor, etc. to et cetera. The system normalizes dates and money terms; such as the token 23 Feb 2026 is changed to February twenty-third 2026. Sections of multilingual text using the Devanagari script are processed using the native

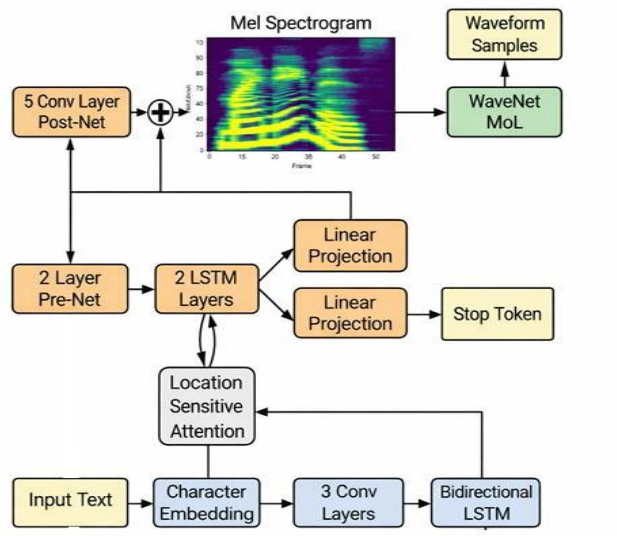
IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, June 2026
Impact Factor: 8.2

orthography, and orthographic assortment with the English language is settled down based on agreed-upon transliteration rules.

- ***Grapheme-to-Phoneme Conversion:*** This sub-process parallels graphemes, written graphemes, to phonetic cognates (phonemes). Languages that have irregular orthographies, like English, require it; examples of this are that the sound cat represents /k æ t/. In the case of Marathi, a conversion based on rules or models is used, e.g. eSpeak or Sequitur G2P to deal with aspirated consonants and deletion of schwa. The resulting production is a series of phonemes with stress signs as dictated by the ARPABET [5] standard.
- ***Prosody and linguistic feature:*** Prosody and linguistic feature extraction add prosodic cues, such as sentence boundary and pause created by punctuations, like commas, and use part-of-speech tags to give information on the intonation pattern which mapped to dense 512-dimensional vectors on a layer of character-embedding's. To maintain the order of sequences positional encodings are then added.
- ***Text + Embedding:*** Tokens are created out of basic text, and embedding's are created in this component before being sent to the encoder. The resulting linguistic code is used to be the basic representation of the encoder-attention-decoder architecture, allowing to produce coherent and expressive speech [5].

3.3 Encoder-Attention-Decoder Network

The processed text sequence goes into the Encoder which converts the input into a sequence of high level hidden representations which captures the linguistic and contextual properties of the sentence as given Fig.3. The last output of the encoder is then passed to an Attention module. This system is intended to serve as an interface of alignment, matching the textual visualizations (the product of the encoder) to the acoustic frames generated by the decoder. It picks out those aspects of the input text that are to be taken care of at every step of the speech generation. Decoder takes the context vectors generated by Attention module and recursively generates an acoustic representation- a Mel -spectrogram. The re-currentational feedback arrow of the schematic refers to the autoregressive behaviour of the decoder, in which past acoustic frames affect the generation of consecutive frames. The speaker encoding provided by the Speaker Encoder is also introduced into the attention-decoder pathway, whereby the system is in a position to adjust the speaker-mobility of the synthesised speech [1].



IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, June 2026
Impact Factor: 8.2

3.4 Generation of Spectrogram and Waveforms

The decoder produces a Mel-spectrogram which summarizes the time frequency representation of the speech waveform. This spectrogram is then sent to the Convert-Waveform module which essentially a vocoder, which converts the spectral representation to a time-domain speech waveform. The system also produces the synthesized speech signal as the ultimate product, therefore, replicating the spoken version of the text sent into the system but preserving naturalness and speaker features [11].

3.5 Dataset Description

The proposed procedure of training and evaluating the suggested bilingual speech-synthesis model is based on the use of the vVISWa dataset, which is a multilingual audio-visual corpus tailored with the aim of ensuring a rich human-computer interaction. The database consists of a matched audio-visual recordings of isolated lexical items in English, Marathi, and Hindi, which were articulated by 58 speakers in a variety of poses (frontal, 45-degree as well as lateral The corpus also provides matched English-Marathi text-audio materials that can be easily modelled to both acoustic analysis and speaker conditioned representation extraction; the audio visual integration of the corpus provides the possibility of supporting future pipelines of talking-face synthesis. Even though the dataset focuses on isolated words, its multi-speaker and multi-lingual aspects provide the possibility to transfer and condition voices in the low-resource setting successfully [9].

3.6 Model Training and Development.

The suggested speech synthesis system is based on a modular end-to-end architecture based on Tacotron-2 and includes a speaker encoder to allow a variety of speakers and bilingualism in Marathi and English. Its architecture includes five major components, which are a paired text-audio dataset, text processing, an encoder-attention-decoder network, a speaker encoder, and a vocoder on the diagram in Fig. 10. The input is divided into a set of linguistic (text) and acoustic (audio) inputs. The speaker encoder (3.5.1) translates reference waveforms into Mel-spectrograms with 3 5 convolutional layers with 512 filters, a five-letter kernel, a stride value of two, and are then converted to 256 units per direction with a 3-layer bidirectional LSTM. An average pooling operation with a full connection layer results in a 256512 dimensional L2 existence of a post, which is based on SV2TTS, which does not consider vocabulary or any true text to discover timbre, pitch, accent, and prosody. This encoding is then embedded either into the encoder states or into the attention mechanism providing the opportunity to select the speaker at runtime, imitate voices, and store it effectively. Normalization of numbers, abbreviations, dates, and Marathi Unicode/by-mix code are done in text processing (3.2).

Grapheme-to-phoneme transformations produce phoneme sequences, whereas prosody annotation creates boundaries, pauses, and part-of-speech labels, positional embedding's also create 512-dimensional representations that impart direction to the encoder to express synthesis.

The linguistic sequence is processed by the encoder-attention-decoder (3.3) consisting of three convolutional layers containing 512 filters, batch normalisation, ReLU activation, and dropout of 0.5 then bidirectional LSTM containing 256 units per direction. The 128 filters in location-sensitive attention. Location-sensitive attention imposes guided monotonicity and skipped or repeated results are not possible. The autoregressive decoder contains 204 LSTM layers and two 1024 units that are followed by a 2-layer preponderating layer (fully connected, 256 units) and a 512 filters post-net of convolutional layers to create Mel-spectrograms conditional on the speaker embedding. To synthesize waveforms (3.4) the system uses a vocoder like Wav2Lip, and trained on 16 kHz monolingual food, and has the capability of converting 80-dimensional Mel-spectrograms generated by a 1024-point FFT with a 25 ms hop size. The training is controlled both on balanced bilingual vVISWa datasets; audio is pre-processed with Mel-spectrograms and text normalized form. Adam optimiser (β 1 = 0.9, 0.999 = 2, e = 1- e6) has an initial learning rate of 10 3, then it decreases to 10 5. Loss functions include L1 and L2 on Mel-spectrogram, binary cross-entropy over stop tokens, and a regularise of weight-decay. Mini-batch sizes are between 8 and 64, sequence sorted sampling, teacher forcing decreases between 1.0 and 0.5. The training

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, June 2026
Impact Factor: 8.2

regimen counts 200k-500k steps or 250-1000 epochs using one GPU, and assesses the training performance quality and alignment quality in-between.

IV. RESULT OF SPEECH SYNTHESIS MODEL

In the following section, the neural text-to-speech system is explained and its elements are described in the manner of Tacotron v2, which features encoder dynamics, decoder dynamics, and guided attention mechanism to improve the correspondence between the written messages and speech synthesis results. A vocoder reconstructs Mel-spectrograms into high quality waveforms. Conditional to the condition of the speaker embedding, the system allows to support concurrent multi-speaker synthesis on the one hand, preserving an individual feature of each speaker in terms of timbre, prosody, and speaking style. This kind of conditioning may prove especially useful in the downstream task of realistic speaking-face generation, in which the voice-identity of the speaker remains a crucial element of sound-visual compatibility. The TTS subsystem is evaluated using both objective and subjective measurements, independent of each other to allow phonemic and prosodic variations specific to English and Marathi.

4.1 Objective Assessment:

Mel-Cepstral Distortion (MCD): In order to evaluate the spectral similarity of synthesized speech and natural references, the Mel Cepstral Distortion (MCD) measure is used. Mel frequency cepstral coefficients (MFCCs) are utilized to make a comparative foundation to the production of MCD. These MFCCs are acoustic stimuli which play an important role in differentiation of timbre and spectral profile by the human ear.

$$MCD[dB] = \left[\omega_D\right] - \frac{10}{\ln 10}\sqrt{2\sum_{n=1}^{k}\left(c_n^{ref} - c_n^{syn}\right)^2}$$

In which c^{ref} *and* c^{syn} represent the k th MFCC of the reference and synthesized frames respectively (usually K = 13 without the 0 th coefficient). The analytics process considers time-warp of the synthesized frames to bring them into reference frames and then averaging the metric to the aligned frames. Reduced values of MCD show that it is more similar to natural speech. Table No.2 lists per-coefficient distortions of coefficients: c 1 to c 13:

Table No. 2. MCD attributes

k	(cnref)	(cnsyn
1	0.50	0.45
2	1.20	1.10
3	0.30	0.25
4	0.80	0.75
5	0.10	0.05
6	0.60	0.55
7	0.40	0.35
8	0.25	0.20
9	0.15	0.10
10	0.10	0.05
11	0.05	0.00
12	0.05	0.02
13	0.00	0.02

The raw squared -error contributions preceding square-root and scaling are presented in the table. Less values represent a better congruence between the synthesized and natural spectra in each of the cepstral dimensions. The given qualitative interpretation traces the various coefficients to common acoustic or perceptual values in speech processing:

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 7, June 2026 Impact Factor: 8.2

Coffee Coefficient-wise analysis of MCD. The MCD is examined in a coefficient-wise manner to show the degree to which the synthesized output maps to the natural reference at spectral dimensions. Coefficient 1, the total energy slope or coarse spectral tilt resonance, has a smaller distortion (reference 0.50 0.45 0.05) suggesting the reproduction of the broadband energy steepness and slope at low frequencies. There is moderate distortion in coefficient 2, which measures the low-frequency formant content and the primary significant spectral peak, (reference 1.20 versus synthesized 1.10) such that, metrically speaking, the largest error made in vowel quality is the placing of the lowest formant positions or forms. The co coefficient 3, which is knowledgeable of the mid-low frequency details like formant transitions, is characterized by low distortion (reference 0.30 versus synthesized 0.25 which is 0.05 above), indicating a good fit in that band. Mid-frequency formant structure Coefficient 4 degrees' area with moderate distortion (reference 0.80; synthesized 0.75, difference 0.05) indicates that there is an acceptable correspondence of mid-range spectral peaks affecting consonant-vowel transitions and the overall timbre.

Coefficient number 5 that captures more details in the higher middle frequencies also performs very low distortion (reference 0.10 vs. synthesized 0.05, difference 0.05). Coefficient 6 which represents the fine mid-to-high frequency envelope shows low-to-moderate distortion (reference 0.60 versus synthesized 0.55, difference 0.05). Coefficient 7, which relates to more frequent spectral shape, indicates low distortion (reference 0.40 vs synthesized 0.35 the difference is 0.05). The coefficient 8, that corresponds to fine high frequency details also has low distortion (reference 0.25 versus synthesized 0.20, 0.05). The same results are obtained with coefficients 9 and 10 with very low distortion (reference 0.15 versus synthesized 0.10; reference 0.10 versus synthesized 0.05, both 5) so there is an excellent Maintenance of finer spectral variations and high-order cepstral detail. Very fine spectral structure Coefficient 11 is almost zero distortion (reference -0.05 versus synthesized 0.00). Coefficient 12 including ultra-fine details is distorted very low (reference 0.05 compared to synthesized 0.02, difference0.03). Lastly, the coefficient 13, the highest-order coefficient to explain the residual noise or very fine spectral irregularities, has hardly any distortion (reference= 0.00 / synthesized= 0.02, 8-0.02) and effectively this assures us that the synthesizer is virtually adding no additional artefacts into this high cepstral region.

4.2. Subjective Evaluations:

4.2.1. Mean Opinion Score (MOS): MOS is a human centred measure of perceived quality. Twenty native speakers (apparently balanced in 10:1) in this experiment were exposed to synthesized utterances, and rated on a five-point scale regarding naturalness, prosody, and similarity to the speaker (included as one overall MOS) [12] [13]: 1 = Bad (Very unnatural, robotic, incomprehensible) 2 = Poor 3 = Fair (acceptable, but of an obviously synthetic nature) 4 = Good (quite natural) 5 = Superior (In indistinguishable with human speech) The last MOS is the mean of all the ratings. It is preferable to have high scores.

The proposed model achieves*:* English MOS = 4.2 (+0.2 above Tacotron 2 baseline) - MOS equals naturalness of top systems which often equal 4.3 -4.5. Marathi: MOS = 3.6 (+0.5 over HMM baseline) - overall significant gain, it did not go down to the fair/okay zone, rather it became good in terms of phonetic issues, particularly with Marathi with limited training data and complicated phonetics.

4.3. Model Comparisons

The constituents that are distorted the most are coefficients 2 and 4 (c. 1.101.20 reference, 1.100.75 synthesised), under which lie lower- and mid-frequency formant areas of importance in vowel identity and wide timbre. These have the greatest contribution to the overall MCD of 5.8 -1 B (English) and 7.1 -1 B (Marathi). The value coefficients (13) indicate steady very small errors, mostly of range 0.00 -0.10, meaning that the synthesiser is good at retaining subtle spectral characteristics, smoothness, and lack of artificial looking artefacts - vocoders and guided attention in Tacotron-like models are good at this. The trend of increased errors on lower coefficients and nearer-to-zero errors on high coefficients is common in current neural TTS: the low cepstral coefficients encode coarse, perceptually interesting shape that is more difficult to accurately model because of the differences in prosody and intonation; the high coefficients encode noise-like features that are easier to model by neural models. These values of per-coefficients show that the speech synthesized will

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, June 2026
Impact Factor: 8.2

be similar to the reference in the dimensions representing the cepstral, with at least the largest errors in the lower-order coefficients (finer shape of the spectrogram) and at least the nearest to zero values in the higher-order coefficients (finer details). Table No.3. indicates the overall corpus-averaged MCD and Graph 1. Showa the comparison with baselines.

Model	Language	MCD (dB) ↓	MOS (1–5) ↑
HMM-Based (baseline)	Marathi	8.2	3.1
Tacotron 2 (baseline)	English	6.5	4.0
Proposed Speech Synthesis	English	5.8	4.2
Proposed Speech Synthesis	Marathi	7.1	3.6

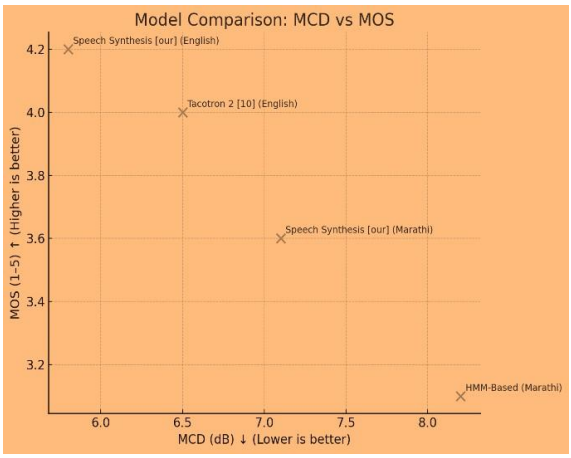


Table No.3. Performance of Proposed Speech Synthesis **Graph 1. Model Comparison**

Baseline HMM-Based (Marathi): MCD = 8.2 dB, MOS = 3.1 Traditional parametric synthesis: Marathi has a complicated prosody and intonation. Baseline (English): Tacotron 2 (base, English): MCD = 6.5 dB, MOS = 4.0 Strong performance on English, high resource language with more simple phonology.

Proposed Speech Synthesis (English): MCD = 5.8 -dB (1.7 -dB improvement over Tacotron 2 baseline), MOS = 4.2 - The gain is of finer spectral fidelity due to guided attention and HiFi-GAN that the modern neural TTS in English started to achieve very good results in the range of 56 –dB.

In the case of Marathi, the speech synthesis suggested results in an MCD of 7.1dB -1.1 dB whereas the HMM baseline is -1.47 dB, and a MOS of 3.6, which is also a significant improvement, considering that the language is low-resource. The current parametric models fail to embrace the Marathi with its rich vowel harmony, retroflex consonants and pitch-accent like prosody. Good neural Speech Synthesis model systems widely perform with MCD of -6-7dB on English; scores of 7-8dB on morphologically rich / low-resource languages such as Marathi point to good performance in these data conditions. In low-resource conditions like in Marathi the progress is especially remarkable and indicates a solid opportunity to extend the measures to more data or even multilingual pre-training [12] [13].

V. CONCLUSION

In general, the results support the current Speech Synthesis module as a quality part of the pipeline, which provides measurable benefits of objective spectral accuracy and natural subjective quality of perception across languages. This article illustrates the end-to-end, modular Neuro-text-to-speech (TTS) synthesizer used to make bilingual English-Marathi applications and as such, provides a hint of the shortcoming of the natural speech generation in low-resource Indic languages. Based on the Tacotron 2 architecture with a collectivist focus to enhance alignment and convergence [5], a speaker encoder of a multi-speaker synthesis and voice cloning, and a bilingual front-end responsible of Marathi-specific phonetic phenomena (aspirated/retroflex consonants, schwa deletions, and code-mixing), the suggested model shows significant progress over traditional baseline neural models [1]. Objective measures with Mel-Cepstral Distortion (MCD) showed spectral error reductions of 5.8dB with English and 7.1dB with Marathi compared to Tacotron2 English (6.5dB) and HMM based Marathi systems (8.2dB), and better fidelity in higher cepstral coefficients and therefore less artifacts [1]. These gains should be supported by Subjective Mean Opinion Scores (MOS): a score of

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, June 2026
Impact Factor: 8.2

4.2 in English with a near-human paragon of the quality versus 3.6 in Marathi with a qualitative improvement between fair and good quality and hence underscoring significant improvement despite the lack of data characteristics of Indic languages [6]. Key innovations of the system incl.: (i) the pioneering bilingual neural TTS pipeline with the focus on English down to Marathi speaker conditioning; (ii) active support of the phonetics and prosodics characteristics of Marathi; and (iii) the intercompatibility with downstream task, e.g. talking faces generation, to arrive at audio-visual consistency. As opposed to English-centric / monolingual models, the current method capitalizes on cross-lingual transfer of bilingual corpora vVISWa with greater efficiency and better performance in the low-resource conditions [1], [6]. To that end, this study significantly contributes to the inclusive and high-quality speech technology in the multilingual environment of India that underlines the accessibility, education, e-governance, and human-machine interaction.

VI. FUTURE WORK

The suggested bilingual neural Speech Synthesis system provides a powerful modular structure of the high-quality English-Marathi speech synthesis with speaker conditioning and alignment robustness. However, there are some good directions in which it can be developed into more integrated, multimodal, and real-time, especially into end-to-end source-to-target models that co-generatively produce synchronized speech audio and lip-synch visual performance on textual or acoustic input. The first point of interest will be the creation of an end-to-end source-to-target model which puts text-to-speech generation in the same pipeline with dynamic lip-synching and talking head generation. Available TTS outputs (Mel-spectrograms and waveforms) may be used as intermediate metrics to train more advanced lip-sync models hence providing a continuous plate to audio-visual consistency without disparate modules. Newer joint models, like a symmetric dual-sided diffusion model of joint audio/video talking portrait generation and multi-modal diffusion of high-quality talking heads using image and audio (current developments), show that end-to-end diffusion-based models provide the best synchronization and expressiveness to cascaded systems. It can be expanded that, in the future, a current vocoder will be replaced by a generative model predicting audio waveforms along with facial features or mouth shapes to decrease the regular latency and increase realism in a code-mixed Marathi-English setting. Using the active application of the previous study of the dynamic lip-sync generation directly, the following step will incorporate the methods in LipChanger model that produces lip-synced face videos in sync with the target audio in the multimedia setting [14]. End-to-end pipeline training on synthesized TTS audio targets allows one to create realistic talking avatar models which maintain speaker identity by using vVISWa embedding. At the same time, the framework Marathi-related phonetic peculiarities, i.e. retro-lexical consonants and schwa erasure, when predicting lip-movement dynamics. Further, using advanced GAN architectures will also increase visual fidelity and general performance. In our overall discussion of progress in GANs, it is clear how the GANs have been developed over the last several years to be used as a high-quality generative in the context of multimodal tasks [16]; future versions can be used to enhance the lip areas and facial expressions of GANs in noisy or low-light environments, thus expanding the range of high-quality glasses to include virtual assistants or e-learning avatars. We will promote audio -visual speech recognition (AVSR) integration as a feedback loop or one of the joint training goals to achieve robustness in unfavourable settings. Based on our overview of more current deep learning approaches to AVSR [15], which focus on fusion strategies and noise-robust visual indicators, the model might utilize visual lips (as provided by vVISWa or generated lip-sync) to improve the TTS prosody and pronouncing mistake in noisy environments 10 - therefore generating a closed-loop condition where synthesized speech is optimized by AVSR. These developments will change the existing TTS pipeline into a full-fledged visual end-to-end audio-visual synthesis framework, where our previous papers on lip-synch-based [14], AVSR-based [15] and GAN-based [16] synthesis can form the main constituent parts of the multimodal speech synthesis technology that is both inclusive and multilingual in India.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, June 2026
Impact Factor: 8.2

ACKNOWLEDGMENT

The authors sincerely thank Dr. Babasaheb Ambedkar Marathwada University (Dr. BAM University), Aurangabad, Maharashtra, India, for providing the necessary research facilities and support through the Vision and Intelligent System (VIS) Lab. We are deeply grateful to the Chhatrapati Shahu Maharaj Research, Training and Human Development Institute (SARTHI), Pune, (SARTHI) Fellowship program of the Government of Maharashtra for the financial assistance and encouragement that made this work possible.

REFERENCES

1. Tan, X., Qin, T., Soong, F., & Liu, T.-Y. (2021). A survey on neural speech synthesis. arXiv preprint arXiv:2106.15561. https://arxiv.org/abs/2106.15561
2. Schlünz, G. I., et al. (2017). Applications in accessibility of text-to-speech synthesis for South African languages. In Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility.
3. Klatt, D. H. (1987). Review of text-to-speech conversion for English. Journal of the Acoustical Society of America, 82(3), 737–793. https://doi.org/10.1121/1.395275
4. van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., & Kavukcuoglu, K. (2016). WaveNet: A generative model for raw audio. arXiv preprint arXiv:1609.03499. https://arxiv.org/abs/1609.03499
5. Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerry-Ryan, R. J., Saurous, R. A., Agiomyrgiannakis, Y., & Wu, Y. (2018). Natural TTS synthesis by conditioning WaveNet on Mel spectrogram predictions. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 4779–4783). IEEE. https://doi.org/10.1109/ICASSP.2018.8461368
6. Srivastava, N., et al. (2020). IndicSpeech: Text-to-speech corpus for Indian languages. In Proceedings of the 12th Language Resources and Evaluation Conference (LREC) (pp. 4214–4220).
7. Jasir, M., & Balakrishnan, A. (2022). Text-to-speech synthesis: Literature review with an emphasis on Indian languages. ACM Transactions on Asian and Low-Resource Language Information Processing, 21(4), Article 94. https://doi.org/10.1145/3501397
8. Bhadoriya, A. S., et al. (2025). A2TTS: TTS for low resource Indian languages. arXiv preprint arXiv:2507.15272. https://arxiv.org/abs/2507.15272
9. Borde, P., Manza, R., Gawali, B., & Yannawar, P. (2016). 'vVISWa' – A multilingual multi-pose audio visual database for robust human computer interaction. International Journal of Computer Applications, 137(4), 33–37. https://doi.org/10.5120/ijca201690908696
10. Jia, Y., Zhang, Y., Weiss, R. J., Wang, Q., Shen, J., Ren, F., Chen, Z., Nguyen, P., Pang, R., Lopez Moreno, I., Wu, Y., & others. (2018). Transfer learning from speaker verification to multispeaker text-to-speech synthesis. In Advances in Neural Information Processing Systems 31 (NeurIPS) (pp. 4485–4495). https://arxiv.org/abs/1806.04558
11. Kong, J., Kim, J., & Bae, J. (2020). HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis. In Advances in Neural Information Processing Systems 33 (NeurIPS) (pp. 17022–17033). https://arxiv.org/abs/2010.05646
12. Prakash, A., et al. (2022). Towards developing state-of-the-art TTS synthesisers for 13 Indian languages with signal processing aided alignments. arXiv preprint arXiv:2210.17153. https://arxiv.org/abs/2210.17153
13. Gupta, M., et al. (2025). Fine tuning based end-to-end Indian English speech synthesis using Tacotron2 and FastSpeech2. In Proceedings of the 6th International Conference on Deep Learning, Artificial Intelligence and Robotics (ICDLAIR 2024). Atlantis Press. https://doi.org/10.2991/978-94-6463-740-3_2
14. Pawar, D., Borde, P., & Yannawar, P. (2024). Generating dynamic lip-syncing using target audio in a multimedia environment. Natural Language Processing Journal, 8, 100084. https://doi.org/10.1016/j.nlp.2024.100084

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, June 2026
Impact Factor: 8.2

15. Pawar, D. R., & Yannawar, P. (2023). Recent advances in audio-visual speech recognition: Deep learning perspective. In First International Conference on Advances in Computer Vision and Artificial Intelligence Technologies (ACVAIT 2022) (pp. 409–421). Atlantis Press. https://doi.org/10.2991/978-94-6463-196-8_31
16. Pawar, D. R., & Yannawar, P. (2024). Advancements and applications of Generative Adversarial Networks: A comprehensive review. International Journal for Research in Applied Science and Engineering Technology, 12(10). https://doi.org/10.22214/ijraset.2024.