



iJRASET

International Journal For Research in
Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** X **Month of publication:** October 2026

DOI: <https://doi.org/10.22214/ijraset.2026.85090>

www.ijraset.com

Call: ☎ 08813907089

E-mail ID: ijraset@gmail.com

A Survey on Transfer Learning for Low-Resource Language Technologies: A Focused Survey of Transformer-Based Code-Mixed Language Identification and Zero-Shot Multilingual Speech Synthesis

Bhuvana S R¹, Niranjana S S², Mrs. Jipsy Jose³, Mrs. Viji C⁴, Mrs. Sagara M R⁵

^{1,2}Department of Computer Science, (Data Science), St. Thomas Institute for Science and Technology, Trivandrum, India

^{3,4,5}Assistant Professor, Department of Computer Science, (Data Science), St. Thomas Institute for Science and Technology, Trivandrum, India

Abstract: Most progress in natural language and speech processing has been achieved for a small group of languages with abundant labelled data, while languages with limited corpora and tooling continue to be underserved. This survey examines how transfer learning is used to narrow that gap in two complementary settings: word-level language identification of romanized Malayalam–English code-mixed social media text, and zero-shot multilingual multi-speaker text-to-speech synthesis for Indonesian, Javanese, and Sundanese. Two peer-reviewed studies are analysed in depth and placed in context using the literature they cite. A taxonomy of transfer strategies is proposed that distinguishes data-centric corpus construction, full fine-tuning of pretrained transformer encoders, shape-matched and partial parameter transfer, frozen pretrained modules used as feature extractors or perceptual critics, and explicit conditioning on language, speaker, and style. The historical development of word-level language identification and of speaker adaptation in neural speech synthesis is traced, and the evaluation practices of both lines of work are compared. The analysis shows that both studies report large gains over their in-study baselines, but that these gains depend on heuristically labelled or pooled small datasets, single evaluation splits, and, in the speech case, an objective metric that shares its encoder with the training loss. Open problems include reliable gold-standard annotation, class imbalance, orthographic variation, decoupled evaluation, and responsible use of voice cloning. Research directions such as code-mixed speech synthesis driven by token-level language tags are outlined.

Index Terms: Code-mixing, language identification, low-resource languages, Malayalam, text-to-speech, transfer learning, transformers, zero-shot speaker adaptation.

I. INTRODUCTION

Language technology is unevenly distributed. Research effort, benchmark datasets, and pretrained models are concentrated on a few widely used languages, whereas many languages spoken by large populations still lack annotated corpora, evaluation resources, and dedicated software [1]. Because modern neural models generally need large amounts of training data, this imbalance tends to widen rather than shrink as methods become more data-hungry.

The problem appears in different forms depending on the modality. In written communication, speakers of Indian languages frequently type in the Latin script and switch between their native language and English inside a single sentence, and sometimes inside a single word [2], [3]. Any downstream analysis of such text, for example sentiment analysis or information retrieval, first needs to know which language each token belongs to; this task is called word-level language identification (WLLI). In spoken language, the challenge is to produce natural speech in a given voice for languages that offer only a few hours of transcribed recordings, ideally for speakers who were never seen during training [4].

Although these tasks differ in input, output, and evaluation, they raise the same methodological question: how can knowledge acquired on a well-resourced source be reused when the target language offers little data? Thara and Poornachandran [2] answer it by fine-tuning several pretrained transformer encoders on a newly collected Malayalam–English corpus. Azizah and Jatmiko [4] transfer an English single-speaker synthesis model and an English speaker-verification encoder to Indonesian, Javanese, and Sundanese, and add a style encoder and a speaker-level training loss to improve zero-shot speaker similarity.

This paper reviews both lines of work together. Its contributions are as follows:

- a taxonomy of transfer strategies for low-resource language technologies that covers both text and speech;
- a concise account of the evolution of WLLI methods and of speaker adaptation in neural text-to-speech (TTS);
- a critical reading of the two core studies, including points where reported claims and reported numbers do not fully agree;
- comparative tables of methods, datasets, and evaluation metrics; and
- a list of open challenges and research directions grounded in the limitations of the reviewed work.

The discussion is organised around four guiding questions. **Q1:** Which forms of transfer are used, and what does each require? **Q2:** How do decisions about data collection and labelling influence the reliability of the reported results? **Q3:** What do the chosen evaluation protocols actually establish? **Q4:** Which problems remain unresolved?

Section II defines the scope of the survey. Section III introduces the necessary background, and Section IV traces the development of both research areas. Section V presents the taxonomy of transfer strategies. Section VI analyses the two core studies, and Section VII compares them. Applications, challenges, future directions, and the limitations of this survey are discussed in Sections VIII–XI, and Section XII concludes the paper.

II. SCOPE AND REVIEW METHOD

This work is a focused technical survey and not a systematic literature review. No database search, screening protocol, or inclusion count is claimed. Two peer-reviewed journal articles form the core evidence: a study of transformer-based WLLI for Malayalam–English code-mixed text [2] and a study of zero-shot multilingual multi-speaker TTS for three Indonesian languages [4]. They were chosen because they represent the text and speech sides of the same underlying problem and both report quantitative results with enough methodological detail for comparison.

Additional works are cited only for context. They were identified through the reference lists of the two core studies, and their bibliographic details are reproduced from those lists; they were not independently screened. No new experiments were carried out for this survey. All numerical values are quoted as reported by the original authors. Where a difference between two reported values is given, it is a simple arithmetic derivation and is marked as such. Because the two core studies address different tasks, datasets, and metrics, cross-study comparisons are qualitative and do not rank one approach above the other.

III. BACKGROUND

A. Low-Resource Languages and Transfer Learning

A language is usually described as low-resource when it lacks large annotated corpora, standard evaluation sets, or mature processing tools, regardless of how many people speak it [1]. Remedies discussed in the literature fall into three broad groups: obtaining more data through collection, augmentation, or automatic labelling; exploiting language-specific linguistic knowledge; and using learning techniques that reuse information from other tasks or languages [1], [4].

Transfer learning belongs to the third group. Following the classical formulation [5], a domain $\mathcal{D}$ consists of a feature space and a marginal distribution, and a task $\mathcal{T}$ consists of a label space and a predictive function $f(\cdot)$. Given a source pair $(\mathcal{D}_s, \mathcal{T}_s)$ and a target pair $(\mathcal{D}_t, \mathcal{T}_t)$, transfer learning seeks to improve the target predictor f_t using knowledge obtained from the source, typically when target data are scarce. In deep transfer learning the predictors are neural networks, and the transferred knowledge is most often a subset of trained parameters.

B. Code-Mixing in Romanized Social Media Text

Code-mixing refers to the use of two or more languages within one utterance [2], [3]. Malayalam is a Dravidian language with official status in Kerala and in the union territories of Lakshadweep and Puducherry; it is highly agglutinative and has its own alphasyllabic script [2]. On social media, however, many users write Malayalam in the Latin script for ease of typing. As a result, script cannot be used as a cue for language, and several difficulties arise [2]:

- *Homographs.* A romanized Malayalam word may coincide with an English word, so the correct tag depends on context.
- *Intra-word mixing.* An English stem may carry a Malayalam inflectional suffix, producing a token that belongs to neither language alone.
- *Noisy spelling.* Elongated characters, contractions, and non-standard abbreviations are common.
- *Skewed distributions.* English and Malayalam tokens dominate, whereas acronyms and mixed tokens are rare.

C. Pretrained Transformer Encoders

The transformer replaces recurrence with multi-head self-attention, which lets every token attend to every other token in a sequence and allows parallel computation [6]. BERT stacks transformer encoder layers and is pretrained bidirectionally on unlabelled text before being fine-tuned for a target task [7]. For WLLI, fine-tuning is framed as token classification. If $\mathbf{h}_i \in \mathbb{R}^d$ is the final hidden state of token i and $\mathcal{Y}$ is the label set, the classifier computes

$$p(y_i | x) = \text{softmax}(\mathbf{W}\mathbf{h}_i + \mathbf{b}), \quad \hat{y}_i = \underset{y \in \mathcal{Y}}{\operatorname{argmax}} p(y | x), \quad (1)$$

with $\mathbf{W} \in \mathbb{R}^{|\mathcal{Y}| \times d}$, and is trained with token-level cross-entropy.

Several BERT variants differ in how or on what they were pretrained. As summarised in [2], DistilBERT is a smaller model obtained by knowledge distillation, and ELECTRA trains a discriminator to detect tokens that a small generator has replaced. CamemBERT follows the RoBERTa recipe but is pretrained on French [8]. XLM-RoBERTa is pretrained on filtered CommonCrawl text in many languages [9]. The open-source Transformers library makes these models accessible through a common interface [2].

D. Neural Text-to-Speech

Tacotron-2 is an attention-based sequence-to-sequence model that maps a character sequence to a mel-spectrogram [4], [10]. A convolutional and recurrent text encoder produces hidden states, a location-sensitive attention module aligns decoder steps with encoder steps, an autoregressive decoder predicts one mel frame and a stop token per step, and a convolutional post-net adds a residual refinement. A separate neural vocoder, such as the flow-based WaveGlow [4], converts the predicted mel-spectrogram into a waveform. Learning a monotonic, approximately diagonal attention alignment is critical: when alignment fails, the output is unintelligible [4]. Multi-speaker variants condition the decoder on a speaker representation, and expressive variants add a prosody representation such as global style tokens (GST) [11].

E. Speaker Representations

Speaker embeddings condense the identity of a voice into a fixed-length vector. The i-vector approach uses factor analysis on statistical speaker models. Neural alternatives, reviewed in [4], include the d-vector, obtained by averaging hidden activations of a network trained to discriminate speakers, and the x-vector, obtained from a time-delay network with statistics pooling [12]. The generalized end-to-end (GE2E) loss trains a d-vector encoder so that each utterance embedding moves towards the centroid of its own speaker and away from the others [13]. For utterance j of speaker i with embedding $\mathbf{e}_{ij}$ and a scaled cosine similarity $S_{ij,k}$ to the centroid of speaker k , the per-utterance loss is

$$\mathcal{L}(\mathbf{e}_{ij}) = -S_{ij,i} + \log \sum_{k=1}^N \exp(S_{ij,k}), \quad (2)$$

where N is the number of speakers in the batch and the centroid of speaker i excludes utterance j when $k = i$. Azizah and Jatmiko argue, citing earlier speaker-recognition work, that d-vectors are better suited than x-vectors when the target domain is low-resource [4].

IV. EVOLUTION OF THE TWO RESEARCH AREAS

A. Word-Level Language Identification

Early WLLI systems relied on dictionaries, character n-gram statistics, and hand-written rules. These were followed by statistical sequence and token classifiers, including conditional random fields [14], logistic-regression and related classifiers with contextual features, and extended Markov models, as reviewed in [2]. Shared tasks on code-switched data then standardised label inventories such as *lang1*, *lang2*, *mixed*, *named entity*, *ambiguous*, and *other*, and reported results for language pairs such as Spanish–English and Modern Standard Arabic–Dialectal Arabic [15].

Neural approaches subsequently learned character-level representations directly. Examples include a character-to-word convolutional encoder combined with a bidirectional LSTM [16]; the review in [2] additionally covers character embeddings for Hindi–English, LSTM taggers for multilingual code-switching, multichannel networks mixing 1-D convolutions and LSTMs, and deep models for English–Hindi–Bengali corpora. Pretrained transformers mark the most recent stage. According to Thara and Poornachandran, their study was the first to apply BERT-family models to Malayalam–English WLLI, a pair for which no public gold-standard corpus had been available [2]. Table I summarises the four families.

TABLE I. FAMILIES OF WORD-LEVEL LANGUAGE IDENTIFICATION APPROACHES

Family	Working principle	Strengths	Limitations	Representative work
Lexicon and rule based	Look up tokens in word lists; apply character n-gram and length rules	Transparent; no labelled training data required	Fails on homographs, unseen spellings, and intra-word mixing; lists must be curated	Pre-annotation step in [2]
Statistical classifiers	Hand-crafted character and context features fed to CRF, HMM, or linear classifiers	Captures some context; cheap to train	Feature engineering effort; limited long-range context	[2], [14]
Neural sequence models	Learned character or word embeddings with CNN and (Bi-)LSTM encoders	Learns spelling patterns automatically; robust to noise	Needs more labelled data; trained from scratch per task	[2], [16]
Pretrained transformers	Fine-tune a pretrained encoder with a token classification head	Contextual representations from large-scale pretraining; little task-specific engineering	Pretraining language may not cover romanized target text; larger compute cost	[2]

B. Speaker Adaptation in Neural TTS

Azizah and Jatmiko group earlier multi-speaker TTS systems into three families [4]. In the first, a speaker encoder is trained jointly with the synthesis network; this requires speaker labels and only supports speakers present in training. In the second, a separately pretrained speaker encoder supplies the embedding, but the synthesis model is fine-tuned on the target speaker. In the third, the pretrained encoder is used without any fine-tuning, which enables zero-shot synthesis for unseen speakers [17], [18]. Earlier zero-shot studies nevertheless reported low similarity for unseen voices and inaccurate accent and prosody [4].

Two further ideas shaped later systems. Perceptual losses, originally proposed for image transformation, compare outputs in the feature space of a pretrained auxiliary network rather than only at the frame level; this principle has been applied to speech, including multi-speaker synthesis with a speaker-verification feedback constraint [19] and expressive TTS with a style reconstruction loss, as discussed in [4]. Global style tokens offer unsupervised control of prosody [11]. In parallel, low-resource TTS has been addressed through semi-supervised training [4], speech-chain training with speech recognition [4], cross-lingual transfer [4], and hierarchical transfer learning [4]. Table II compares the speaker adaptation families.

TABLE II. SPEAKER ADAPTATION FAMILIES IN NEURAL TTS (SUMMARISED FROM [4])

Family	Key property / limitation	Zero-shot
Jointly trained encoder	Needs speaker labels; acts as a lookup table at inference	No
Pretrained encoder + fine-tuning	Adapts per target speaker; requires target data and retraining	No
Pretrained encoder, no fine-tuning [17], [18]	Embedding extracted from a short sample; similarity for unseen voices often low	Yes
Zero-shot + style and perceptual loss [4]	Adds GST prosody conditioning and an utterance-level speaker loss	Yes

V. A TAXONOMY OF TRANSFER STRATEGIES

Fig. 1 organises the techniques found in the two core studies along two questions: what is transferred, and how is it reused in the target model. Table III lists the trade-offs.

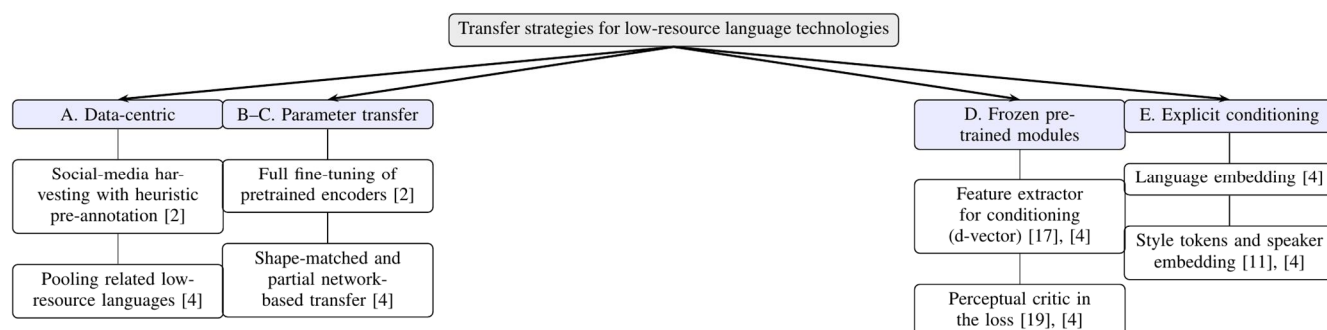


Fig. 1. Proposed taxonomy of transfer strategies observed in the reviewed studies. Letters refer to the subsections of Section V.

A. Data-Centric Strategies

When no suitable corpus exists, building one is the first transfer step. Thara and Poornachandran collected YouTube comments through the platform's API, retained romanized bilingual sentences with a regular-expression filter, normalised elongated spellings and contractions, and assigned labels with a rule cascade (an English dictionary, an emoji pattern, stop-word and acronym lists, symbol and digit sets, a length threshold, and a mixed-word test), after which exceptions were corrected manually [2]. In effect, heuristic knowledge is transferred into supervision. Azizah and Jatmiko instead pooled three small corpora of related languages into one multilingual training set, so that each language benefits from acoustic and phonetic regularities shared with the others [4].

B. Full Fine-Tuning of Pretrained Encoders

In [2], all parameters of each pretrained encoder are updated on the target corpus together with a newly initialised token classifier. The approach requires no architectural change, but it assumes that the pretraining data provide useful representations for the target text, an assumption examined in Section VI.

C. Shape-Matched and Partial Parameter Transfer

Conventional network-based transfer copies a source sub-network into the target model only when the corresponding layers have identical shapes. Extending a single-speaker model to a multilingual, multi-speaker, style-conditioned one changes the input width of several layers, because new embeddings are concatenated to the encoder output and to the decoder input. Azizah and Jatmiko therefore copy a lower-dimensional source tensor element-wise into the matching positions of the larger target tensor and leave the remaining entries at their initial values [4]. For a weight matrix this can be written as

$$\mathbf{W}^t[a, b] \leftarrow \begin{cases} \mathbf{W}^s[a, b], & a \leq m_s, b \leq n_s, \\ \mathbf{W}_0^t[a, b], & \text{otherwise,} \end{cases} \quad (3)$$

where $\mathbf{W}^s \in \mathbb{R}^{m_s \times n_s}$, $\mathbf{W}^t \in \mathbb{R}^{m_t \times n_t}$ with $m_s \leq m_t$ and $n_s \leq n_t$, and $\mathbf{W}_0^t$ is the target initialisation. Fig. 2 illustrates the idea. The new rows or columns that process language, speaker, or style inputs are learned during fine-tuning, while the text-processing part starts from trained values.

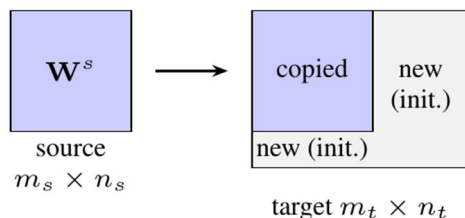


Fig. 2. Concept of partial network-based transfer: trained source weights fill the matching block of a larger target tensor; extra entries for new inputs remain at their initial values (drawn from the description in [4]).

D. Frozen Pretrained Modules

A pretrained module can be reused without updating its weights. In zero-shot TTS, a frozen speaker-verification encoder maps a short reference recording to a d-vector that conditions synthesis [4], [17]. The same frozen encoder can also act as a perceptual critic: the synthesized mel-spectrogram is passed through it, and the distance between its embedding and the embedding of the reference speech is added to the training loss [4], [19]. The first use transfers a representation; the second transfers a supervision signal.

E. Explicit Conditioning

Multilingual and multi-speaker models need inputs that tell the network which language, voice, and speaking style to produce. In [4], a one-layer fully connected language encoder maps a one-hot language index to a continuous embedding, a GST-based style encoder summarises prosody from the reference mel-spectrogram, and the d-vector speaker embedding is supplied both to the attention module and to the decoder pre-net. These embeddings are concatenated with the text encoding, which is precisely why the partial transfer of Section V-C is needed.

TABLE III. COMPARISON OF TRANSFER STRATEGIES IDENTIFIED IN THE REVIEWED STUDIES

Strategy	What is transferred	Advantages	Limitations	Used in
Heuristic pre-annotation	Lexical and orthographic knowledge as labels	Rapid creation of large labelled corpora	Labels inherit rule errors; models may learn the rules rather than the languages	[2]
Pooling related languages	Shared phonetic and acoustic regularities	More training data per model; supports cross-lingual synthesis	Benefit is smaller for languages with divergent pronunciation rules	[4]
Full fine-tuning	All encoder parameters	Simple; strong in-domain results	Depends on pretraining coverage of the target text; risk of overfitting small data	[2]
Shape-matched transfer	Sub-networks with identical dimensions	No architectural constraints on copied parts	Cannot reuse layers whose input width changed	[4]
Partial transfer	Lower-dimensional weights into larger tensors	Reuses layers after architectural extension	Benefit shown qualitatively; theory of why it works not analysed	[4]
Frozen feature extractor	Speaker representation	Enables zero-shot synthesis without speaker labels	Encoder trained on another language may misrepresent target voices	[4], [17], [18]
Frozen perceptual critic	Speaker-level supervision	Optimises a property that frame losses ignore	Couples training objective to evaluation metric if the same encoder is used	[4], [19]

VI. ANALYSIS OF THE CORE STUDIES

A. Transformer-Based WLLI for Malayalam–English

1) Corpus and Labels

The corpus contains about 50 000 romanized sentences and 775 430 tokens gathered from comments posted under entertainment, film, food-review, and news videos on YouTube [2]. Six labels are used; their reported frequencies are shown in Table IV. Fig. 3 summarises the construction and classification pipeline.

TABLE IV. LABEL INVENTORY OF THE MALAYALAM–ENGLISH CORPUS [2]

Label	Meaning	Tokens	Share
eng	English word	320 353	41.31%
mal	Malayalam word	307 858	39.7%
univ	Symbols, emoji, numbers	89 385	11.5%
mix	Mixed morphemes in one word	24 820	3.2%
undef	Undefined/unintelligible	23 229	2.99%
acr	Acronym	9 785	1.26%

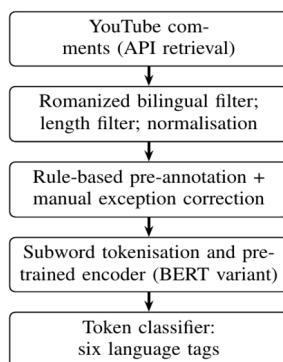


Fig. 3. Conceptual pipeline of transformer-based WLLI for Malayalam–English text, redrawn from the description in [2].

2) Models and Training

Five encoders were fine-tuned: BERT-base-uncased, DistilBERT, ELECTRA-base discriminator, XLM-RoBERTa-base, and CamemBERT-base [2]. After tuning, the authors used a batch size of 16, a learning rate of 2×10^{-5} , dropout of 0.1, 10 epochs, and a maximum sequence length of 128, with an 80/20 train–validation split.

3) Reported Results

Table V reproduces the main results. ELECTRA obtained the highest F1-score (0.9933) and the lowest misclassification rate (0.0058). The *mix* label was the hardest category for every model, ranging from 0.8305 for CamemBERT to 0.9734 for ELECTRA. Confusions occurred mostly between *eng* and *mal*. On a Hindi–English dataset [20], the same approach produced F1-scores between 0.8774 and 0.9038, and the authors summarise the difference relative to Malayalam–English as an increase of about 9% [2].

TABLE V. REPORTED WLLI RESULTS ON MALAYALAM–ENGLISH [2]

Model	P	R	F1	Acc.	mix acc.
BERT	0.9934	0.9901	0.9917	0.9893	0.9649
ELECTRA	0.9937	0.9931	0.9933	0.9941	0.9734
CamemBERT	0.9793	0.9653	0.9718	0.9839	0.8305
XLM-RoBERTa	0.9883	0.9811	0.9846	0.9899	0.9161
DistilBERT	0.9893	0.9902	0.9897	0.9880	0.9687

4) Critical Assessment

Several aspects of the study bear on how its results should be interpreted.

Label provenance. Because labels were produced mainly by a deterministic rule cascade and then corrected for exceptions, a model can reach very high agreement by approximating the rules rather than by recognising language. In the rule cascade, *mal* is the fallback for any word longer than two characters that matches none of the other rules, so a rare English word missing from the dictionary would also become *mal*. The paper does not report how many labels were changed manually or any inter-annotator agreement, so the share of genuinely human-verified labels is unknown.

Ceiling effects. Four of the five models lie within 0.0087 F1 of each other (a derived difference between 0.9933 and 0.9846). No repeated runs, variance estimates, or significance tests are reported, so the ranking among these models is not established with statistical confidence.

Cross-dataset comparison. The Hindi–English corpus uses an eight-label scheme with a different distribution; for example, its *mixed* class covers 0.04% of tokens, compared with 3.2% for *mix* in the Malayalam–English corpus [2], [20]. The 9% gap therefore reflects differences between datasets and annotation schemes at least as much as differences between methods, and it should not be read as evidence that the approach is better suited to Malayalam.

Pretraining coverage. BERT, DistilBERT, and ELECTRA were pretrained mainly on English, CamemBERT on French, and XLM-RoBERTa on many languages [7], [8], [9]. Romanized Malayalam is probably rare in all of these sources. The weak *mix* accuracy of CamemBERT is consistent with this concern, but tokenizer fragmentation was not analysed in [2], so this remains a hypothesis.

Reporting details. The model description in [2] lists XLM-RoBERTa-base with eight attention heads and 125 M parameters and describes its pretraining corpus as 2.5 gigabytes. These values do not appear to match the configuration described by the original XLM-R authors [9] and should be verified before reuse. Proper nouns are not given a separate tag and are placed under *mal*, which the authors acknowledge as a limitation [2].

B. Zero-Shot Multilingual Multi-Speaker TTS for Indonesian Languages

1) Training Pipeline

The approach of [4] proceeds in three stages, shown in Fig. 4. A Tacotron-2 model is first trained on the English LJSpeech corpus of about 24 hours from one female speaker. A GE2E d-vector encoder is then trained on LibriSpeech-360, which contains 921 speakers and about 363.5 hours of speech. In the third stage, the target model inherits the text encoder and post-net through shape-matched transfer, the attention module, decoder, and pre-net through partial transfer, and the speaker encoder through shape-matched transfer, after which all parts except the frozen speaker encoder are fine-tuned on the target data.

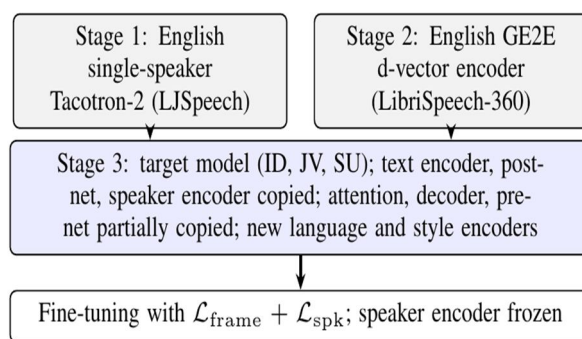


Fig. 4. Three-stage transfer pipeline for zero-shot multilingual multi-speaker TTS, redrawn conceptually from the description in [4].

2) Training Objective

Let Y^{gt} be the ground-truth mel-spectrogram, Y and Y' the decoder output before and after the post-net, and Z^{gt}, Z the true and predicted stop tokens. The frame-level and speaker-level terms are

$$\mathcal{L}_{\text{frame}} = \text{MSE}(Y^{gt}, Y) + \text{MSE}(Y^{gt}, Y') + \text{BCE}(Z^{gt}, Z), \quad (4)$$

$$\mathcal{L}_{\text{spk}} = \text{MSE}(\mathbf{q}, \mathbf{q}'), \quad (5)$$

where $\mathbf{q}$ and $\mathbf{q}'$ are the d-vectors of Y^{gt} and Y' , and the model minimises $\mathcal{L}_{\text{frame}} + \mathcal{L}_{\text{spk}}$ [4].

3) Data and Evaluation

The target domain combines TITML-IDN for Indonesian (14.5 hours, 20 speakers) with OpenSLR corpora for Javanese (7 hours, 41 speakers) and Sundanese (5.5 hours, 39 speakers) [4]. Thirty speakers form the seen set and 24 the unseen set. Evaluation covers naturalness through MOS, speaker similarity through DMOS, spectral distortion through mel-cepstral distortion (MCD) [4], speaker cosine similarity of d-vectors, and UMAP visualisation of embeddings [4]; 15 to 30 listeners took part in the subjective tests [4].

4) Reported Results

Training from scratch and shape-matched transfer alone did not learn a usable attention alignment, whereas partial transfer produced a diagonal alignment and intelligible speech; this evidence is presented through loss curves and alignment maps [4]. Table VI shows the aggregated speaker-similarity results. In the unseen native condition, cosine similarity rose from 0.196 to 0.664, and in the unseen foreign condition from 0.094 to 0.373. Listener ratings of naturalness (MOS) did not separate the proposed model from the baselines in a statistically meaningful way, and the MCD differences between models were likewise not significant. Javanese had the highest distortion (MCD 8.064 for seen and 12.007 for unseen native speakers with the proposed model), which the authors attribute to grapheme-to-phoneme rules that differ from those of Indonesian and Sundanese [4].

TABLE VI. REPORTED SPEAKER SIMILARITY, ALL LANGUAGES AGGREGATED [4]

Condition	Metric	Tdv_mne	Tdv_mse	Proposed
Native, seen	Cosine	0.707	0.731	0.893
Native, unseen	Cosine	–	0.196	0.664
Foreign, seen	Cosine	0.381	0.115	0.470
Foreign, unseen	Cosine	–	0.094	0.373
Native, seen	DMOS	3.369	3.401	3.536
Native, unseen	DMOS	–	2.642	3.014
Foreign, seen	DMOS	2.863	2.505	2.894
Foreign, unseen	DMOS	–	2.387	2.677

Tdv_mne cannot synthesize unseen speakers. DMOS scale: 1–4.

5) Critical Assessment

Metric-objective coupling. The speaker loss is computed with the same d-vector encoder that produces the cosine-similarity metric. A large cosine gain is therefore partly expected by construction. DMOS, which is independent of the encoder, also improves, which supports a genuine effect; however, DMOS for foreign-language synthesis remains below 3.0 on a four-point scale, so listeners were often unsure whether the voices matched.

Uneven gains. The proposed model is not the best in every cell. For seen Javanese speakers in the foreign condition, the cosine similarity of the jointly trained baseline (0.503) exceeds that of the proposed model (0.412). The authors also state that the baseline with a pretrained encoder performs worse than the jointly trained baseline; this holds for the foreign seen aggregate (0.115 versus 0.381) but not for the native seen aggregate (0.731 versus 0.707) [4].

Transfer evidence. The superiority of partial transfer is shown qualitatively. Objective intelligibility measures, such as the character error rate of an automatic speech recogniser applied to the synthesized audio, are not reported and would make the comparison with the failing schemes more convincing.

Foreign-language references. Because none of the speakers is multilingual, foreign-language outputs are compared with the speaker’s native-language recordings and, for MCD, with a different speaker reading the same text [4]. Both choices limit what the foreign-language scores can show.

VII. COMPARATIVE ANALYSIS

A. Study-Level Comparison

Table VII sets out the two core studies in parallel. Both construct or assemble a target dataset, both transfer from English-centric sources, and both report large improvements over their own baselines on a single data split. They differ in how much of the pretrained model is updated: the WLLI study fine-tunes entire encoders, whereas the TTS study combines fine-tuning, partial copying, and a frozen module.

TABLE VII. CROSS-STUDY COMPARISON OF THE CORE STUDIES

Aspect	WLLI for Malayalam–English [2]	Zero-shot multilingual TTS [4]
Task and modality	Token classification of written text (six tags)	Mel-spectrogram generation for unseen voices
Languages	Romanized Malayalam and English (code-mixed)	Indonesian, Javanese, Sundanese (native and cross-lingual)
Target data	About 50 000 sentences / 775 430 tokens from YouTube, rule-labelled	27 hours in total from three public corpora; 54 speakers split into seen and unseen
Transfer source	English, French, and multilingual pretrained encoders	English TTS (LJSpeech) and English speaker encoder (LibriSpeech-360)
Transfer mechanism	Full fine-tuning	Shape-matched and partial transfer; frozen encoder; perceptual loss
Main metrics	Precision, recall, F1, accuracy, confusion matrix	MOS, DMOS, MCD, cosine similarity, UMAP
Headline result	F1 = 0.9933 (ELECTRA)	Unseen native cosine 0.196 → 0.664
Statistical reporting	No variance or significance tests	95% confidence intervals; significance discussed for MCD and MOS
Main caveat	Rule-derived labels; no human agreement measure	Training critic and evaluation metric share the encoder

B. Evaluation Metrics

The two studies measure different properties, summarised in Table VIII. For classification, precision P and recall R combine into

$$F_1 = \frac{2PR}{P + R}, \quad (6)$$

which is more informative than accuracy for skewed label distributions [2]. For speech, speaker similarity between the d-vector $\mathbf{d}_u$ of a synthesized utterance and the speaker model $\mathbf{d}_{\text{spk}}$ is measured by

$$\cos(\mathbf{d}_u, \mathbf{d}_{\text{spk}}) = \frac{\mathbf{d}_u^\top \mathbf{d}_{\text{spk}}}{\|\mathbf{d}_u\| \|\mathbf{d}_{\text{spk}}\|}, \quad (7)$$

and spectral distortion over T aligned frames with K cepstral coefficients $c_{t,k}$ (reference) and $c'_{t,k}$ (synthesized) by

$$\text{MCD}_K = \frac{1}{T} \sum_{t=0}^{T-1} \sqrt{\sum_{k=1}^K (c_{t,k} - c'_{t,k})^2}, \quad (8)$$

with $K = 13$ and dynamic time warping for alignment in [4].

TABLE VIII. EVALUATION METRICS USED IN THE CORE STUDIES

Metric	Measures / caveat	Study
Accuracy	Overall correctness; inflated by majority classes	[2]
F1	Balance of precision and recall; per-class view needed for rare tags	[2]
Confusion matrix	Error structure between tags	[2]
MOS	Perceived naturalness; small listener panels	[4]
DMOS	Perceived speaker match; independent of encoder	[4]
MCD	Spectral distance; mixes speaker and language effects	[4]
Cosine sim.	Embedding similarity; coupled to training loss	[4]
UMAP	Qualitative view of speaker space	[4]

C. Synthesis

With respect to Q1, the reviewed work shows that transfer in low-resource settings is rarely a single operation: data construction, parameter reuse, frozen modules, and conditioning are combined. With respect to Q2, the quality of the target data shapes what the results mean; heuristic labels in the WLLI study and pooled, monolingual-speaker corpora in the TTS study both bound the conclusions. With respect to Q3, neither study uses an evaluation that is fully independent of its own design decisions, although the human DMOS ratings in [4] come closest. No ranking between the two studies is possible, since they share no task, data, or metric.

VIII. APPLICATIONS

A. Text Processing

Thara and Poornachandran position WLLI as a prerequisite for downstream analysis of code-mixed text, mentioning sentiment analysis, information retrieval, and the detection of security threats on social media [2]. Their related-work discussion also lists code-mixed question answering, aspect extraction, polarity identification, named entity recognition, and parsing as tasks that depend on knowing the language of each token [2]. Sentiment analysis of code-mixed tweets has been organised as a shared task [2].

B. Speech Synthesis

Zero-shot multilingual TTS can provide voices for languages that lack large single-speaker recordings; Azizah and Jatmiko note that Indonesia alone has more than 700 low-resource ethnic languages [4]. Because synthesis requires only a few seconds of reference audio, the same technique supports personalised voices and cross-lingual synthesis in which a speaker's voice is used for text in another language.

C. Prospective Combination

Code-mixed text, if it were to be read aloud, would require token-level knowledge of which pronunciation rules to apply. A WLLI tagger could serve as the front end of a multilingual synthesiser conditioned on a per-token language embedding. Neither core study implements this combination; it is listed here as a prospective application connecting the two lines of work.

IX. CHALLENGES AND OPEN ISSUES

A. Reliable Annotation

The Malayalam–English corpus is large but largely rule-labelled [2]. A smaller, independently annotated test set with reported inter-annotator agreement would make reported scores far more informative. Named entities, currently merged into *mal*, need their own label, as in established code-switching label schemes [15].

B. Class Imbalance and Rare Categories

The *mix* and *acr* labels together account for less than 5% of tokens [2], yet mixed words are exactly the phenomenon that makes code-mixing difficult. Overall scores are dominated by the frequent classes, and per-class reporting with macro-averaging should be standard.

C. Orthographic and Phonological Variation

Romanization of Malayalam has no single standard, so the same word may be spelled in several ways [2]. On the speech side, grapheme-to-phoneme ambiguity, for example the Javanese vowel letter that can correspond to more than one phoneme, increased distortion for all models [4].

D. Evaluation Validity

The WLLI study reports a single train–validation split without variance estimates [2]. The TTS study uses small listener panels and an objective similarity metric that is coupled to its own training loss [4]. Both issues make it difficult to separate method effects from evaluation artefacts.

E. Generalisation

Data come from one platform in the WLLI study and from three related languages with monolingual speakers in the TTS study. Behaviour on other platforms, other Indian language pairs, unrelated language families, and genuinely multilingual speakers is untested.

F. Ethics and Privacy

Social-media comments are written by identifiable users, and zero-shot voice cloning can reproduce a person’s voice from seconds of audio. Neither core study discusses consent, data protection, or safeguards against misuse; these questions become pressing as the methods move towards deployment.

X. FUTURE RESEARCH DIRECTIONS

The following directions follow from the limitations above.

- 1) *Gold-standard code-mixed benchmarks.* Human-annotated Malayalam–English test sets with agreement statistics and a named-entity tag would allow fair comparison of WLLI methods.
- 2) *Romanization-aware pretraining.* Encoders pretrained on Indian languages and on romanized text, together with an analysis of subword fragmentation, could test whether pretraining coverage explains differences between models.
- 3) *Cross-platform and cross-pair evaluation.* Testing on other platforms and on other Dravidian–English pairs would show whether the reported performance generalises.
- 4) *Decoupled speech evaluation.* An independently trained speaker-verification model for scoring, together with recognition-based intelligibility measures, would separate genuine improvements from metric coupling.
- 5) *Linguistic knowledge in TTS.* Pronunciation lexicons and language-specific grapheme-to-phoneme rules, proposed as future work in [4], could address the Javanese distortion problem.
- 6) *Multilingual speakers.* Corpora with speakers recorded in several languages would provide proper references for cross-lingual synthesis [4].

- 7) *Code-mixed speech synthesis*. Combining token-level language identification with multilingual TTS, as outlined in Section VIII-C, is an unexplored meeting point of the two lines of work.
- 8) *Statistical rigour*. Multiple random seeds, confidence intervals, and significance tests should accompany results, especially when differences are small.
- 9) *Efficiency*. The small loss of DistilBERT relative to larger encoders [2] suggests that compact models deserve systematic study for low-resource deployment.
- 10) *Responsible deployment*. Consent procedures, data-handling policies, and misuse safeguards should be designed alongside the models.

XI. LIMITATIONS OF THIS SURVEY

This survey rests on two core studies and on the works they cite; it does not cover the full literature on code-mixing or low-resource speech synthesis. Secondary references were taken from the core bibliographies and were not independently verified. No experiments were reproduced, and results that appear only in figures of the original papers, such as the ablation bar chart in [4], were not converted into numbers. The interpretations offered in Section VI, including the label-provenance and pretraining-coverage concerns, are hypotheses that require empirical testing.

XII. CONCLUSION

This survey examined transfer learning for low-resource language technologies through two complementary studies: transformer-based word-level language identification for romanized Malayalam–English text and zero-shot multilingual multi-speaker speech synthesis for Indonesian, Javanese, and Sundanese. A taxonomy of transfer strategies was proposed, ranging from data-centric corpus construction to full fine-tuning, partial parameter transfer, frozen pretrained modules, and explicit conditioning. Both studies show that knowledge from high-resource sources can make low-resource tasks tractable: pretrained encoders reached very high agreement with the corpus labels, and partial transfer combined with style conditioning and a speaker-level loss enabled intelligible speech with markedly higher speaker similarity for unseen voices. At the same time, the analysis identified factors that limit how far these results generalise, including heuristically derived labels, single evaluation splits, coupling between the training objective and an evaluation metric, and narrow linguistic and platform coverage. Progress will depend as much on better benchmarks and evaluation practice as on new architectures, and the combination of code-mixed language identification with multilingual speech synthesis offers a natural next step.

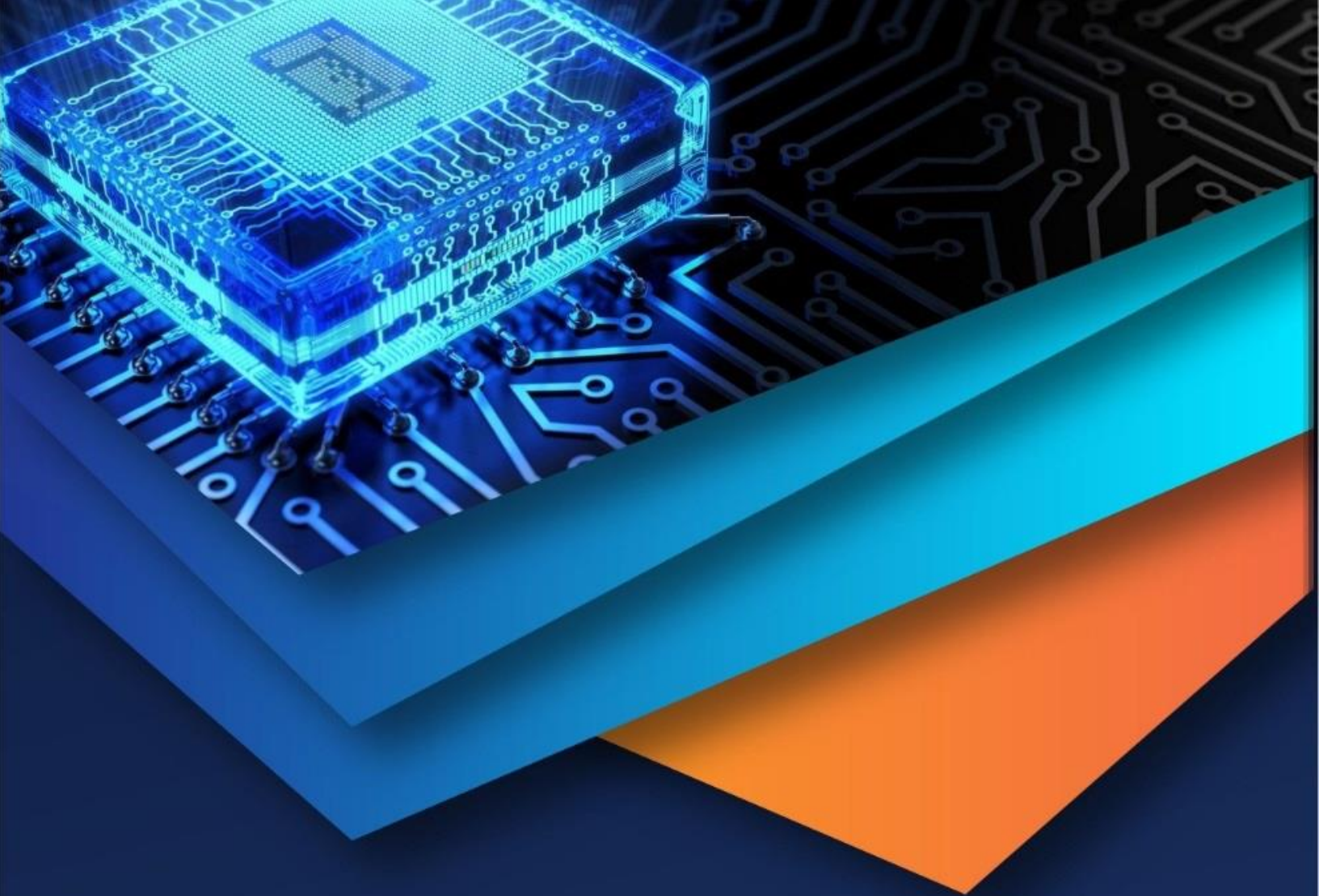
XIII. ACKNOWLEDGMENT

An AI assistant (Claude, Anthropic) was used to help draft and organise parts of this manuscript. All content was checked by the authors, who take responsibility for its accuracy.

REFERENCES

- [1] M. A. Hedderich, L. Lange, H. Adel, J. Strötgen, and D. Klakow, “A survey on recent approaches for natural language processing in low-resource scenarios,” in Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Lang. Technol. (NAACL-HLT), 2021, pp. 2545–2568.
- [2] S. Thara and P. Poornachandran, “Transformer based language identification for Malayalam-English code-mixed text,” IEEE Access, vol. 9, pp. 118 837–118 850, 2021.
- [3] U. Barman, A. Das, J. Wagner, and J. Foster, “Code mixing: A challenge for language identification in the language of social media,” in Proc. 1st Workshop Comput. Approaches Code Switching, 2014, pp. 13–23.
- [4] K. Azizah and W. Jatmiko, “Transfer learning, style control, and speaker reconstruction loss for zero-shot multilingual multi-speaker text-to-speech on low-resource languages,” IEEE Access, vol. 10, pp. 5895–5911, 2022.
- [5] S. J. Pan and Q. Yang, “A survey on transfer learning,” IEEE Trans. Knowl. Data Eng., vol. 22, no. 10, pp. 1345–1359, 2010.
- [6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2017, pp. 5998–6008.
- [7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” arXiv:1810.04805, 2018.
- [8] L. Martin, B. Müller, P. J. O. Suárez, Y. Dupont, L. Romary, É. V. de la Clergerie, D. Seddah, and B. Sagot, “CamemBERT: A tasty French language model,” arXiv:1911.03894, 2019.
- [9] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, “Unsupervised cross-lingual representation learning at scale,” arXiv:1911.02116, 2019.
- [10] J. Shen et al., “Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), 2018, pp. 4779–4783.

- [11] Y. Wang, D. Stanton, Y. Zhang, R. J. Skerry-Ryan, E. Battenberg, J. Shor, Y. Xiao, Y. Jia, F. Ren, and R. A. Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in Proc. 35th Int. Conf. Mach. Learn. (ICML), 2018, pp. 5180–5189.
- [12] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), 2018, pp. 5329–5333.
- [13] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized end-to-end loss for speaker verification," in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), 2018, pp. 4879–4883.
- [14] G. Chittaranjan, Y. Vyas, K. Bali, and M. Choudhury, "Word-level language identification using CRF: Code-switching shared task report of MSR India system," in Proc. 1st Workshop Comput. Approaches Code Switching, 2014, pp. 73–79.
- [15] G. Molina, F. AlGhamdi, M. Ghoneim, A. Hawwari, N. Rey-Villamizar, M. Diab, and T. Solorio, "Overview for the second shared task on language identification in code-switched data," in Proc. 2nd Workshop Comput. Approaches Code Switching, 2016, pp. 40–49.
- [16] A. Jaech, G. Mulcaire, M. Ostendorf, and N. A. Smith, "A neural model for language identification in code-switched tweets," in Proc. 2nd Workshop Comput. Approaches Code Switching, 2016, pp. 60–64.
- [17] Y. Jia, Y. Zhang, R. J. Weiss, Q. Wang, J. Shen, F. Ren et al., "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2018, pp. 4485–4495.
- [18] E. Cooper, C.-I. Lai, Y. Yasuda, F. Fang, X. Wang, N. Chen, and J. Yamagishi, "Zero-shot multi-speaker text-to-speech with state-of-the-art neural speaker embeddings," in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), 2020, pp. 6184–6188.
- [19] Z. Cai, C. Zhang, and M. Li, "From speaker verification to multispeaker speech synthesis, deep transfer with feedback constraint," in Proc. INTERSPEECH, 2020, pp. 3974–3978.
- [20] D. Mave, S. Maharjan, and T. Solorio, "Language identification and analysis of code-switched social media text," in Proc. 3rd Workshop Comput. Approaches Linguistic Code-Switching, 2018, pp. 51–61.



10.22214/IJRASET



45.98



IMPACT FACTOR:
7.129



IMPACT FACTOR:
7.429



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089  (24*7 Support on Whatsapp)