SBC Computing Reviews, 2026, 05:1,
ISSN: 2966-3938 • doi: 10.5753/reviews.2026.8617
This work is licensed under a Creative Commons Attribution 4.0 International License.

Systematic Literature Review

Accent Transfer: A Systematic Literature Review

Accent Transfer: A Systematic Literature Review

Leticia Gabriella de Souza [SiDi Intelligence & Innovation Center - TTS Team, Brazil | *l.gabriella@sidi.org.br*]
Arnaldo Candido Junior [Institute of Biosciences, Language and Physical Sciences, São Paulo State University, Brazil | *arnaldo.candido@unesp.br*]
Sandra Maria Aluisio [University of São Paulo at São Carlos – SP, Brazil | *sandra@icmc.usp.br*]

SiDi Intelligence & Innovation Center - TTS Team - Rua Aguaçu, 171, Ed. Jacarandá, Alphaville Empresarial, Campinas, SP, Brazil.

Resumo.
Contexto: A fala sempre foi essencial para a comunicação humana. No campo do Processamento de Línguas Natural (PLN), os avanços nas tecnologias de voz — como assistentes virtuais e tradução de fala para fala (STS) — visam tornar a interação entre humanos e máquinas cada vez mais natural. **Lacuna:** Entretanto, a variedade limitada de sotaques regionais em modelos de Text-To-Speech (TTS) frequentemente leva a uma percepção de falta de autenticidade. **Objetivos:** Este artigo apresenta uma Revisão Sistemática da Literatura (RSL) projetada para analisar avanços, mapear o estado da arte e identificar os desafios remanescentes na transferência de sotaque, distinguindo-a da Conversão de Voz (CV) geral. **Métodos:** Aderindo ao método PICOC e às diretrizes de Kitchenham, extraímos dados de 91 estudos revisados por pares e de acesso aberto, publicados entre 2010 e 2026 nas principais bibliotecas digitais. **Resultados:** A revisão delineia uma clara mudança arquitetural em direção a abordagens modernas impulsionadas por tokens discretos, Large Language Model (LLM) e Aprendizado Auto-Supervisionado (AAS). Descobrimos que, embora Word Error Rate (WER) e Mean Opinion Score (MOS) sejam as métricas mais utilizadas, e o L2-ARCTIC o conjunto de dados mais comum, eles apresentam limitações significativas para avaliar a fidelidade do sotaque com precisão. **Conclusão:** Apesar dos avanços arquiteturais, a Transferência de Sotaque ainda enfrenta desafios críticos, notavelmente um viés linguístico acentuado para o inglês, a falta de métricas de avaliação padronizadas focadas na pronúncia e o persistente emaranhamento entre identidade do falante e sotaque em cenários *zero-shot*.

Abstract. **Context:** Speech has always been essential for human communication. In the field of Natural Language Processing (NLP), advancements in voice technologies — such as virtual assistants and speech-to-speech (STS) translation — aim to make the interaction between humans and machines increasingly natural. **Gap:** However, the limited variety of regional accents in TTS models often leads to a perceived lack of authenticity. **Objectives:** This paper presents a Systematic Literature Review (SLR) designed to analyze advancements, map the state-of-the-art, and identify the remaining challenges in accent transfer, distinguishing it from general Voice Conversion (VC). **Methods:** Adhering to the PICOC method and Kitchenham guidelines, we extracted data from 91 open-access peer-reviewed studies published between 2010 and 2026 across major digital libraries. **Results:** The review outlines a clear architectural shift toward modern approaches driven by discrete tokens, LLM, and Self-Supervised Learning (SSL). We found that while WER and MOS are the most frequently used metrics, and L2-ARCTIC the most common dataset, they exhibit significant limitations for accurately evaluating accent fidelity. **Conclusion:** Despite architectural advances, Accent Transfer still faces critical challenges, notably a pronounced linguistic bias toward English, a lack of standardized pronunciation-focused evaluation metrics, and persistent entanglement of speaker identity and accent in *zero-shot* scenarios.

Palavras-chave: Transferência de Sotaque, TS, Conversão de Voz, CV, Conversão de Sotaque, CS, Processamento de Fala, Síntese de Fala

Keywords: Accent Transfer, AT, Voice Conversion, VC, Accent Conversion, AC, Speech Processing, Speech Synthesis

Recebido/Received: 02 July 2026 • **Aceito/Accepted:** 14 October 2026 • **Publicado/Published:** 23 November 2026

1 Introduction

Since the beginning of human evolution, speech has been the main instrument for exchange of information and promoting social cohesion. In the current context of Natural Language Processing (NLP) the development of voice technologies, such as virtual assistants and enhancing Speech-to-Speech (STS) translation, requires the establishment of increasingly natural interactions between humans and machines.

However, as pointed out by Azevedo *et al*. [2026], the way we communicate is not the same everywhere. Speech is full of variations and regional accents that must be taken into account in the development of these technologies, so that people can identify with them and better accept the use of voice systems.

Although VC and Accent conversion (AC) share similar functions in speech processing, their objectives are distinct. VC focuses on modifying the vocal identity, timbre and speaker characteristics of a source speaker to match that of a target speaker while preserving the underlying linguistic content [Sisman *et al*., 2020]. Meanwhile, Accent transfer (AT) often referred to as AC, works as a specialized subfield of VC. AT strictly targets the modification of segmental (phoneme pronunciation) and suprasegmental (rhythm, duration, stress

and intonation) acoustic features to match a target dialect or non-native accent and at the same time preserving the source speaker's original vocal timbre and speaker identity [Sisman *et al*., 2020; Quamer and Gutierrez-Osuna, 2024]. By considering vocal quality and accent as independent and decoupled factors, AT gets around the broad identity reconfiguration mechanisms of classical VC [Sisman *et al*., 2020].

TTS models have been trained with a limited variety of languages and accents; this lack of representation is a known issue in the literature. For example, Azevedo *et al*. [2026] indicates that the technology fails when interacting with most of the different users. In speech, the impact of these differences is so significant that TTS models trained solely on a single accent, such as European Portuguese, would cause great strangeness and even rejection if used as the voice of a virtual assistant for Brazilians [Azevedo *et al*., 2026]. Sociolinguistics shows that accent is an essential mark of a person's identity [Oushiro, 2021; Collet, 2020; Lorandi *et al*., 2011]. As a result, the main purpose of modeling and transferring accents in TTS models is, in fact, to synthesize identities. By developing artificial voices that include regional accents, we avoid imposing a single standard, begin to legitimize differences in speech, and validate the identity of marginalized populations.

Techniques VC and AC have gained importance in modern speech processing research. According to Sisman *et al*. [2020], Voice Conversion focuses on modifying the voice of a source speaker to sound like the voice of a target speaker while maintaining the original linguistic content. Quamer and Gutierrez-Osuna [2024] suggest that Accent Conversion provides a fine-grained separation of speaker characteristics by treating accent and voice quality as independent factors that must be decoupled. As shown by Shrestha *et al*. [2025], the conversion of an accent requires the extraction and modification of specific features, such as rhythm, pronunciation, stress, and intonation, while preserving the unique identity of the speaker. Furthermore, extracting and isolating vocal identities remains an active research problem, with recent studies exploring contrastive learning and adversarial training to disentangle speaker representations from environmental noise in languages such as Brazilian Portuguese [Matos *et al*., 2026].

This systematic review of the literature analyzed advances and understanding of the field of accent transfer. The primary goal of this review is to map the state-of-the-art methodologies in accent transfer to identify open challenges. The SLR was organized as follows: Section 2 details the structured methodology to carry out this SLR, including the search strategy and selection criteria. Section 3 provides an analysis of state-of-the-art architectures, features, evaluation metrics, datasets, and challenges and gaps identified in the field of accent transfer, followed by the conclusions in Section 4.

2 Methodology

The methodology of this SLR is designed to systematically map the current landscape of AC and VC within Speech Synthesis (SS). The primary goal is to identify the methods, techniques, architectures, and ongoing challenges related to accent adaptation within TTS models, highlighting current trends and literature gaps. Ultimately, this review aims to provide a consolidated analysis of the state-of-the-art and serve as a foundational guide for future developments in this field. To ensure reproducibility, the entire review protocol was structured using the Parsifal platform[1], strictly adhering to the systematic literature review guidelines proposed by Kitchenham and Charters [2007].

Additionally, to ensure rigor despite resource constraints, the first author conducted the initial literature review through a comprehensive review of titles and abstracts. Following this initial filtering, the full-text selection and data extraction phases were performed. To mitigate potential selection and reviewer biases arising from this primary single-author involvement, all extracted data, conflicts, and boundary cases were systematically cross-verified by the two co-authors, bringing the total number of reviewing authors to three. Disagreements regarding study inclusion or quality assessment were resolved through consensus meetings, in which the full texts of ambiguous cases were analyzed collectively. Given that the initial selection and quality assessment were not conducted in a fully independent and blinded manner by multiple reviewers from the outset, the calculation of an inter-rater reliability measure, such as Cohen's Kappa, is not statistically applicable to our workflow.

2.1 Research Questions

To track recent developments, six research questions were formulated. These questions aim to consolidate the scope of the review and guide the extraction of relevant data from the following literature.

Architectures & Models:

- **(RQ1)** What are the main architectures that support AT and how are they adapted?

Techniques:

- **(RQ2)** What training techniques and embeddings are used to represent accent conversion?

Evaluation:

- **(RQ3)** What objective and subjective metrics are used to evaluate the quality of AT?

Challenges & Impact:

- **(RQ4)** What are the main challenges and limitations observed in current approaches?
- **(RQ5)** What research trends and gaps were identified in the literature between 2010 and 2026?
- **(RQ6)** How does AT impact performance in low-resource languages?

2.2 Search Strategy

To systematically answer the research questions pre-defined previously, the search was structured using PICOC (Population, Intervention, Comparison, Outcome and Context) method. This ensures that the retrieved literature aligns with the defined scope.

[1] https://parsif.al/

- **Population:** SS, TTS models, AC, VC, and End-to-End (E2E) architectures that address variations in accents, styles, or vocal identities, including low-resource languages.
- **Intervention:** Methods and approaches for AT.
- **Comparison:** TTS or VC models without accent adaptation or based on monolingual/traditional training used as baselines.
- **Outcome:** Identification of how AT is implemented in E2E architectures, the metrics used to evaluate naturalness, similarity and intelligibility, and the existing gaps.
- **Context:** Research published between 2010 and 2026 in conferences, journals, and workshops related to TTS, SS, and Deep Learning (DL).

2.3 Data Sources & Search String

The automated searches were executed between November 2025 and April 2026, in the following digital libraries: (1) Google Scholar, (2) IEEE Digital Library, and (3) Scopus. These specific databases were selected because IEEE and Scopus are the leading high-quality repositories for peer-reviewed engineering, computer science, and speech processing literature. Although ACL Anthology is a primary repository for computational linguistics, it was not searched as a standalone database because its entire catalog is comprehensively indexed within both Scopus and Google Scholar, ensuring that no major studies were omitted. The query strings varied slightly between databases due to their distinct search mechanisms. Thus, the Google Scholar query was intentionally constrained to exact keyphrases (e.g., *"accent transfer"* OR *"accent conversion"*) to maintain search precision.

To ensure that no key foundational study was left out, this was supplemented with a (4) Manual Search. This manual phase consisted of targeted queries on arXiv using isolated keyphrases ("accent conversion", "accent transfer", and "voice conversion"). Additionally, to identify significantly cited and topically related literature, we selected the most frequently recurring articles from our initial database searches to serve as seed papers. Specifically, the foundational works authored by Gutierrez-Osuna and his research group were selected as seeds, as they are widely recognized as the most frequently cited, relevant, and pioneering literature establishing the core concepts of accent conversion. These seed papers were inputted into the visual citation graph tool Connected Papers[2] to perform forward and backward snowballing, allowing us to manually separate and extract articles closely connected to our core keywords. This rigorous snowballing approach also reduced potential inconsistencies — such as database indexing bias or the inadvertent exclusion of relevant preprints — ensuring a robust and exhaustive coverage of the literature.

String for (2) IEEE Digital Library and (3) Scopus: *("Speech Synthesis" OR "TTS" OR "VC" OR "Voice Conversion" OR "text-to-speech") AND ("accent Transfer" OR "accent Conversion" OR "accent Modification" OR "accent Style Transfer" OR "accented speech" OR "speaker conversion").*

String for (1) Google Scholar and (4) Manual Search: *"accent Transfer" OR "accent Conversion"*.

The search strings were adapted according to the specificities of each database. The comprehensive string was applied to IEEE and Scopus to maximize precision using complex Boolean operators. Conversely, the string for Google Scholar and the manual search was simplified to accommodate the platform's character limits and distinct parsing algorithms, preventing search syntax errors while maintaining a broad retrieval scope.

Applying the generic search string across the databases combined with the manual search resulted in a total of *357 records*. The distribution of the studies retrieved per data source is detailed in Figure 1.

2.4 Selection Criteria

To identify relevant studies and filter the results obtained from the search strategy, inclusion (I) and exclusion (E) criteria were defined, as detailed in Table 1.

Table 1. Inclusion (IC) and Exclusion Criteria (EC)

Inclusion Criteria	
IC1	Studies that address accent transfer, adaptation, or modeling in TTS.
IC2	Studies that deal with accent transfer methodologies.
IC3	Studies utilizing E2E models focused on accent transfer.
Exclusion Criteria	
EC1	Articles published outside the 2010–2026 range (e.g., 1900–2009).
EC2	Studies partially out of the research scope.
EC3	Studies completely out of the research scope.
EC4	Duplicate studies across databases.
EC5[a]	Studies that are not open access or not published in English.
EC6	Theses, dissertations, abstracts, technical reports, and non-systematic reviews.

[a] Specifically regarding EC5, the exclusion of non-open-access studies was established to ensure the full reproducibility of this review, guaranteeing that the extracted architectural methodologies and datasets could be thoroughly verified by the community without paywall restrictions. We acknowledge this as a potential threat to validity, as it may exclude relevant proprietary research or classical methodologies published in closed-access journals, slightly narrowing the scope of historical context.

To align with our English-language search strings, we established an eligibility criterion restricting the selected literature exclusively to English publications. Since English is the lingua franca of global scientific communication in computer science and speech processing, the vast majority of state-of-the-art architectural advancements in leading databases (e.g., IEEE, Scopus) are published in English. Restricting the review to this language ensures a uniform baseline for peer-reviewed methodological quality and avoids inconsistencies in the extraction of highly technical architectural concepts and metrics that could arise from cross-lingual translation nuances.

[2] `https://www.connectedpapers.com/`

2.5 Quality Assessment

The selected studies went through a quality assessment process to ensure methodological quality and data reliability. A checklist was applied to evaluate the completeness and relevance of each study. The assessment questions included:

- (Q1): Does the study provide a clear description of the accent transfer method?
- (Q2): Are the experiments detailed enough to allow for understanding and reproducibility?
- (Q3): Is accent transfer addressed as the central theme?
- (Q4): Is the proposed model based on a E2E architecture?
- (Q5): Is the proposed model separate Acoustic model (AM) or unified with a vocoder?
- (Q6): Is there a comparison with baseline models (without accent transfer)?
- (Q7): Does the study include objective and/or subjective evaluations (e.g., MOS, speaker similarity, intelligibility)?
- (Q8): Does the study contribute to low-resource languages or under-studied accents?
- (Q9): Does the article clearly describe the datasets used?
- (Q10): Does the study address cross-lingual transfer or transfer between different accents of the same language?
- (Q11): Are limitations, challenges, or future research directions discussed?
- (Q12): Is the model or code reproducible (publicly available or detailed)?

The scoring for each of the 12 questions followed this scale: *1.0 = Meets the requirements*, *0.5 = Partially meets the requirements*, or *0.0 = Does not meet the requirements*. Therefore, the maximum possible score is strictly 12.0 points per paper. Studies that achieved a total score equal to or greater than 6.0 (out of 12.0) were accepted and included in the final list of retrieved reports.

2.6 Screening & Inclusion

Initial search strategies across all databases and manual registers returned a total of 357 records, as illustrated in Figure 1. Analyzing by data source, Scopus provided the highest volume with 138 records, followed by the IEEE Digital Library with 94, Google Scholar with 68, and manual searches contributing 57 records. The deduplication procedures removed 108 overlapping records from these sources, specifically eliminating 46 from Scopus, 25 from Google Scholar, 19 from IEEE, and 18 from manual searches.

Following the removal of duplicates, 249 unique records advanced to the title and abstract selection phase. Articles evaluated in this phase were composed of 92 records from Scopus, 75 from IEEE, 43 from Google Scholar, and 39 from manual searches. The exclusion criteria at this stage resulted in the rejection of 158 records. These exclusions contained 71 records from Scopus, 54 from IEEE, 21 from Google Scholar, and 12 from manual searches. Finally, 91 studies successfully passed the initial screening and were referred for full-text quality assessment. This included 21 from Scopus, 21 from IEEE, 22 from Google Scholar, and 27 manually sourced studies. Quality assessment protocols evaluated

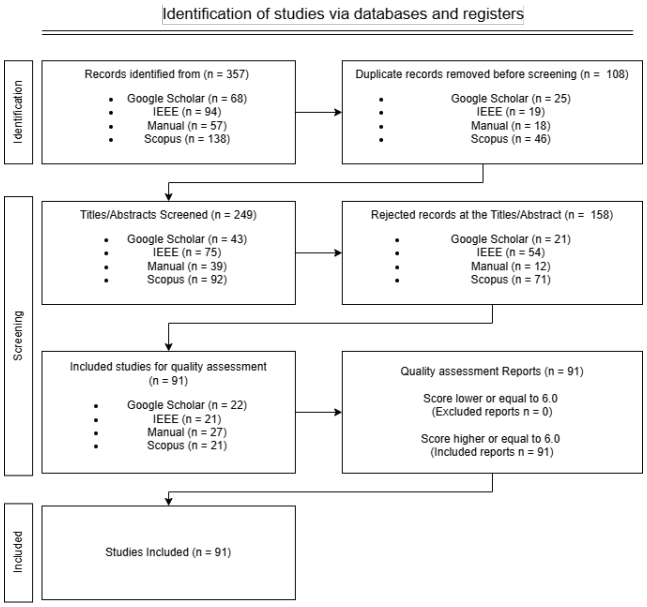

Figure 1. Flow Diagram of the Screening and Inclusion process.

these 91 reports, with zero studies scoring 6.0 or lower. All 91 evaluated reports achieved a score higher than 6.0.

Regarding the distribution of scores for the 91 included studies, 46 studies scored highly between 10.0 and 12.0 (indicating high methodological completeness and reproducibility), 35 studies scored between 8.0 and 9.5, and 10 studies clustered near the lower acceptable bounds between 6.0 and 7.5. It is notable that all 91 selected studies passed the 6.0 cutoff score, resulting in zero exclusions at this stage. The 12 Quality Assessment checklist questions structured within the Parsifal platform according to Kitchenham's guidelines were designed to rigorously evaluate methodological completeness and reproducibility. The cutoff point of 6.0 (50%) was strictly adopted to ensure a baseline of quality while preventing the exclusion of pioneering or early-stage architectural studies that might lack extensive evaluations but introduce valuable novel concepts. Likewise, maintaining this specific threshold was fundamentally necessary due to the relatively limited volume of literature explicitly dedicated to AT; a more aggressive cutoff would risk over-filtering and severely restricting the scope of the state-of-the-art mapping. Although the 100% approval rate weakens the exclusionary function of this specific step — which we explicitly acknowledge as a statistical limitation — this outcome primarily reflects the high restrictiveness of the preceding PICOC-based screening and inclusion criteria (such as EC2 and EC3), which effectively eliminated low-quality or out-of-scope reports before they could reach the full-text quality assessment phase.

3 Results and Discussion

A systematic and manual data extraction process was executed in the 91 previously selected articles. Relevant information was mapped, such as the most frequently used architectures, datasets and languages predominantly explored in the context of accents, as well as adopted objective and subjective evaluation metrics.

3.1 Datasets

The analysis of the studies reveals that the datasets most frequently used for the development and validation of the VC and AC models are mainly consolidated around English language accents, establishing a hierarchy between foundational data for general speech modeling and specialized data for accent transfer.

As detailed in Table 2, the L2-ARCTIC dataset is the most used resource in the literature, appearing 28 times. Rather than reflecting mere adoption bias, its dominance is technically justified because it was built specifically for voice and accent conversion, as well as mispronunciation detection [Zhao *et al*., 2018b]. L2-ARCTIC is frequently evaluated in conjunction with CMU-ARCTIC (19) and ARCTIC (8) because these corpora provide a strict cross-speaker textual parallelism; they are composed of non-native (L2) and native (L1) speakers reading the exact same phonetically balanced utterances [Aryal and Gutierrez-Osuna, 2014a; Kominek and Black, 2004; Zhao *et al*., 2018b]. This parallel symmetry is an indispensable mathematical requirement for frame-level and sequence-level alignments, allowing models to calculate reference-based objective metrics, such as Mel Cepstral Distortion (MCD), to map correct native-like pronunciation while preserving the speaker's vocal timbre [Ding *et al*., 2022]. Additionally, the symmetric demographic structure of L2-ARCTIC — comprising 24 speakers from six distinct L1 backgrounds and balanced by gender — combined with its manually annotated and force-aligned phonemic boundaries, provides a highly robust framework for validating the separation of speaker and accent and preventing timbre leakage. By offering detailed annotations of systematic pronunciation errors such as phonemic substitutions, insertions, and deletions [Radzikowski *et al*., 2021] this corpus provides a level of phonetic granularity that large-scale generic corpora cannot replicate. In contrast, while the Speech Accent Archive (3) is also a relevant dataset that provides a wider variety of accents from over 200 native-language backgrounds [Watanabe and Kameoka, 2024]; its samples often lack the long-term speaker-specific parallel data needed to train and evaluate complex parametric mapping functions.

Table 2. Main Datasets Mapped in the Literature *(* indicates strictly parallel corpora)*

Dataset Name	Qty	Dataset Name	Qty
L2-ARCTIC *	28	NPTEL2020	3
VCTK	24	Glasgow Herald corpus *	3
LibriSpeech	21	Speech Accent Archive	3
CMU-ARCTIC *	19	TIMIT	3
LJSpeech	11	AfriSpeech	2
LibriTTS	11	GigaSpeech	2
Common Voice	9	ATR Japanese database	2
ARCTIC *	8	JVS	2
LibriTTS-R	7	MagnaTagATune	2
VoxCeleb1	5	CENSREC-1-C	2
LibriLight	4	WenetSpeech	2
VoxCeleb2	4	Wisconsin X-ray microbeam	2

However, it is crucial to distinguish between the frequency of use of L2-ARCTIC and its actual suitability for the future of Accent Transfer. Although its historical dominance is justified by the reliance of early latent architectures on strictly parallel data, it is not inherently the most suitable dataset for modern state-of-the-art architectures. Its limited demographic scope (restricted to only 24 speakers across 6 linguistic backgrounds) restricts the evaluation of zero-shot, SSL, and LLM-based models, which fundamentally require massive, diverse, and non-parallel data to generalize effectively across real-world accents. Therefore, its high frequency of use is more indicative of a historical methodological limitation than of absolute superiority.

For the initial training phases, studies use a native dataset such as VCTK (24) and LibriSpeech (21) to finetune models to understand different English dialects [Yamagishi *et al*., 2019; Panayotov *et al*., 2015]. To make sure the final audio sounds professional and clear, LJSpeech is often used to train high-quality vocoders like HiFi-GAN [Kong *et al*., 2020; Li *et al*., 2023].

Additionally, VoxCeleb 1 and 2 are used for teaching models to recognize unique voices, allowing the models to work with new speakers they have never heard before, also known as zero-shot conversion [Nagrani *et al*., 2017; Chung *et al*., 2018]. Common Voice(9) is also frequently used in the literature, particularly due to its large-scale and diverse speech data [Ardila *et al*., 2020]. Specialized datasets such as TIMIT and NPTEL2020, which focus on Indian accents, also appear to address specific phonetic or regional challenges [Zhong *et al*., 2025b].

Studies use LibriTTS (11) and its restored version LibriTTS-R (7) to provide multi-speaker English data that is cleaner than standard audiobooks and commonly used for pre-training conversion models [Koizumi *et al*., 2023; Zen *et al*., 2019]. LibriLight (4), GigaSpeech (2) and WenetSpeech (2) have around 10k hours of audio to train models to generalize across thousands of different speakers [Kahn *et al*., 2020; Chen *et al*., 2021; Zhang *et al*., 2022]. The Glasgow Herald corpus (3) provides phonetically balanced sentences for statistical mapping [Aryal and Gutierrez-Osuna, 2014a], while AfriSpeech (2) is used to tackle Pan-African English accents [Olatunji *et al*., 2023]. For Japanese speech research, the ATR Japanese database (2) and the JVS corpus (2) are the main tools for multi-speaker identity training [Cheng *et al*., 2025]. Finally, ultra-specialized data like the Wisconsin X-ray microbeam (2) allows models to study the physical movements of the tongue and jaw [Siriwardena *et al*., 2024], while CENSREC-1-C (2) tests if models can still work in noisy environments. Even music datasets like MagnaTagATune (2) are sometimes used to verify whether speech-based audio codecs can handle musical complexity [Zeghidour *et al*., 2021].

Despite the dominance of American English, a group of languages with intermediate frequency, such as Hindi (27), Mandarin (25), Spanish (24), Korean (20) and Vietnamese (15), remains abundant in the literature and also shown in Figure 2. This prevalence is almost entirely attributable to their inclusion in the L2-ARCTIC corpus, which has emerged as the standard for training models to process non-native English pronunciation [Zhao *et al*., 2018b; Cheng *et al*., 2025]. In these studies, models are typically trained to map specific L1 patterns back to a native American or British English reference [Bai *et al*., 2025b; Zhou *et al*., 2023b].

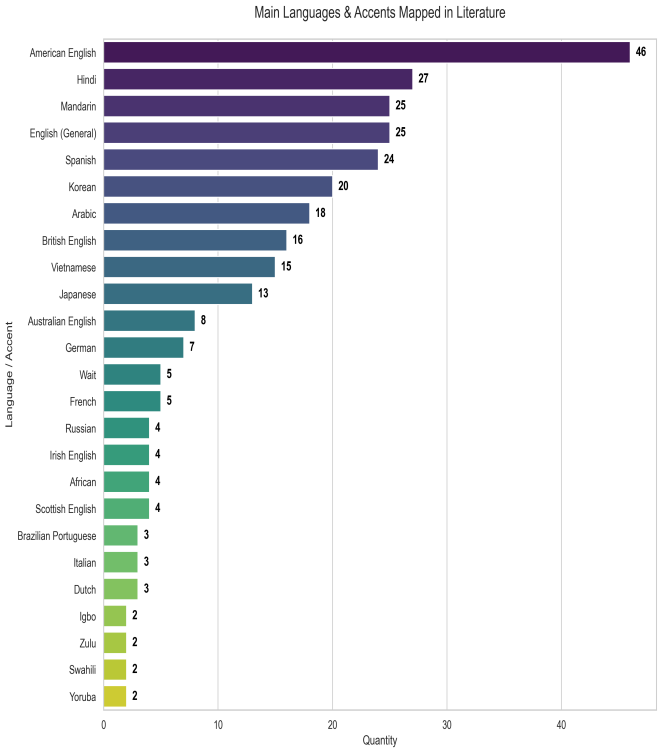

Figure 2. Main Languages & Accents Mapped in the Literature

Some studies explore regional variations such as Australian English (8) and Scottish English (4) for different analytical purposes. Instead of cross-lingual mapping, these dialects are used to evaluate how well model's handles intra-lingual micro-variations, such as subtle vowel shifts and distinct prosodic patterns [Huang and Toda, 2023; Zhou *et al.*, 2023a; Onda *et al.*, 2025]. Some studies test the model's ability to preserve a speaker's specific regional identity [Zhong *et al.*, 2025b,a].

When non-English target languages are used, they generally serve to train or test specific capabilities of advanced architectures. For example, Brazilian Portuguese (3) and Italian (3) are predominantly used to evaluate zero-shot and few-shot multilingual model's, testing whether the model can synthesize speech with minimal training data [Casanova *et al.*, 2023; Ueda *et al.*, 2024]. African languages, such as Igbo, Zulu, Swahili and Yoruba are used to test the robustness of Automatic Speech Recognition (ASR) systems against highly diverse phonetic structures [Olatunji *et al.*, 2023; Nechaev and Kosyakov, 2024; Prabhu *et al.*, 2023]. Languages like Dutch (3), Russian (4) and Japanese (13) are also frequently used in cross-lingual studies to assess whether or not a model can successfully disentangle and preserve a speaker's original voice timbre while synthesizing a different language [Casanova *et al.*, 2024; Correa *et al.*, 2023].

In the context of low-resource languages **(RQ6)**, AT faces a major performance bottleneck because existing English-centric models struggle to process the distinct phonetic, phonological, and prosodic variations of the target accents [Halychanskyi *et al.*, 2026b]. Specifically, African languages present highly diverse phonotactic structures, tonal contours, and morphological complexities that test the robustness of neural representations under low-resource and noisy conditions [Nechaev and Kosyakov, 2024]. Similarly, regional variations of Hindi and Arabic represent complex dialectal glides typically misaligned by standard frame-based mappers [Halychanskyi *et al.*, 2026b; Lertpetchpun *et al.*, 2026]. When applied to tonal languages like Mandarin, these models suffer from representation entanglement and "prosody leakage", where English rhythmic structures incorrectly distort lexical tones [Lertpetchpun *et al.*, 2026]. Brazilian Portuguese serves as another notable example due to its unique phonetic features, such as complex rhotic codas and dental stop variations, which are highly prone to representation entanglement during zero-shot synthesis [Matos *et al.*, 2026]. Without modeling these highly divergent regional accents to establish localized baselines, global speech synthesis models remain structurally biased.

3.2 Evaluation Metrics

Measurement of the effectiveness of accent and voice conversion models **(RQ3)** is based on established standard TTS metrics. As detailed in Table 3, rather than relying on a single metric, current research calculates a variety of objective and subjective measurements to evaluate accent. Objective metrics are often applied because they do not require human effort and can be extracted automatically, guaranteeing reproducibility and scalability in experiments. However, subjective metrics, which depend on human perception, are also used to evaluate and reinforce accent transfer performance.

Objective Metric	Qty	Objective Metric	Qty
WER	32	PER	2
MCD	28	MAE	2
CER	14	LFC	2
F0 RMSE	10	NISQA	2
SECS	8	F0 PCC	2
COSSINE SIMILARITY	6	MOSNET	2
UTMOS	3		
SES	2		

Table 3. Frequency of objective metrics mapped in the literature. MCD (Mel Cepstral Distortion); WER, CER, and PER (Word, Character, and Phoneme Error Rate); SES (Speaker Embedding Similarity); UTMOS (UTokyo-SaruLab MOS); NISQA (Non-Intrusive Speech Quality Assessment); F0 RMSE (Fundamental Frequency Root Mean Square Error); PCC (Pearson Correlation Coefficient); SECS (Speaker Embedding Cosine Similarity); MAE (Mean Absolute Error).

Metrics such as WER (32), Character Error Rate (CER) (14) and Phoneme Error Rate (PER) (2) are used to evaluate intelligibility and appear in the majority of the analyzed studies to indicate content preservation. Nevertheless, a high frequency of adoption does not equate to suitability for specifically evaluating accent transfer. Heavily relying on ASR-based metrics like WER is insufficient because standard ASR models are intrinsically biased toward native pronunciations, often interpreting valid non-native phonetic traits as transcription errors. Similarly, while the studies heavily rely on general MOS (29) to evaluate human perception, its widespread use is not entirely suitable for AT, as listeners often conflate acoustic naturalness with accent authenticity. To mitigate this, more suitable and specialized subjective metrics have emerged, such as Similarity MOS (SMOS)

(27) to evaluate target identity retention, alongside Naturalness MOS (NMOS) (19) and Quality MOS (QMOS) (9).

For acoustic quality and prosody, the most frequently calculated metrics are MCD (28), which quantify the spectral distance between synthesized and target speech, Fundamental Frequency Root Mean Square Error (F0 RMSE) (10) as the fundamental frequency (F0) captures the intonation and rhythm that characterize a target accent. Additionally, to validate speaker identity and accent similarity, studies employ embedding based metrics, such as Speaker Embedding Cosine Similarity (SECS) (8) and Cosine Similarity (6). Moreover, there is a noticeable trend towards neural based metrics such as UTokyo-SaruLab MOS (UTMOS) (3), Non-Intrusive Speech Quality Assessment (NISQA) (2) and MOSNET (2), which automatically calculate human perception scores regarding the naturalness and quality of speech.

Furthermore, beyond the predominant main metrics, the data mapped revealed a "long tail" of less frequent objective metrics, each appearing only once in the reviewed literature. To assess fine-grained acoustic quality and prosodic, isolated studies ([Takashima *et al*., 2012; Lumban Tobing *et al*., 2019; Joglekar *et al*., 2025; Kameoka *et al*., 2020; Zhou *et al*., 2025; Melechovsky *et al*., 2024c; Zhong *et al*., 2025a]), calculated metrics such as Mel-Cepstral Distance (MEL-CD), Logarithmic Global Variance Distortion (Log-GV Distortion), Structural Similarity Index Measure (SSIM), Local Duration Ratio (LDR), Frame Distortion (FD), F0 Frame Error (FFE) and Voicing Flag RMSE (VF RMSE). Similarly, for embeddings and automatic verification techniques to better quantify speaker identity and accent similarity, metrics, such as Accent Cosine Similarities (ACCCOS), Speaker Cosine Similarities (SPKCOS), Silhouette Coefficient for Speaker Clusters (SCSC), Automatic Speaker Verification (ASV), Speaker Similarity (SS) and Accent Embedding Similarity (AES), provide a more specific evaluation of accents and are employed in some studies ([Zhong *et al*., 2025b; Bai *et al*., 2024; Inoue *et al*., 2025; Halychanskyi *et al*., 2026a; Huang *et al*., 2024]).

The literature ([Felps and Gutierrez-Osuna, 2010; Takeuchi and Tatekura, 2018; Yixin *et al*., 2025]) also features classical algorithms , such as ITU-T P.563, HTK Scores, Linear Discriminant Analysis (LDA), along with specific intelligibility and alignment measures, such as Longest Common Subsequence Ratio (LCSR).

Evaluating accent through human perception is also important, as objective metrics cannot calculate subtle voice nuances that only a human listener can recognize, as detailed in Table 4. Consequently, subjective metrics remain standard in accent and voice conversion evaluation. The studies relies on MOS (29) to evaluate human perception of accents, followed by SMOS (27), to evaluate how close the generated voice is to the target identity or accent. Furthermore, to ensure that the voices generated maintain human characteristics and do not sound artificial, NMOS (19) and QMOS (9) are significantly applied, along with naturalness evaluations such as Naturalness (NAT) (1).

The literature also demonstrates the need to isolate the accent; for this purpose, metrics focused exclusively on pronunciation stand out, known as Accentedness (11) and Accent MOS (AMOS) (5). In these evaluations, listeners are instructed to ignore the audio quality and focus specifically on scoring the intensity, fidelity and authenticity of the foreign or regional generated accent by the model.

Subjective Metric	Qty	Subjective Metric	Qty
MOS	29	AMOS	5
SMOS	27	MUSHRA	5
NMOS	19	CMOS	4
ACCENTEDNESS	11	PMOS	4
QMOS	9	VSS	2
ABX	8	EGWA	2
XAB	7	NAT	1

Table 4. Frequency of subjective metrics mapped in the literature. MOS(Mean Opinion Score); S/N/Q/A/C/PMOS(Similarity, Naturalness, Quality, Accent, Comparative and Prosody MOS); MUSHRA(Multiple Stimuli with Hidden Reference and Anchor).

In conclusion, a substantial part of accent validation is based on comparative tests. Given the challenge of evaluating an accent in isolation, the literature frequently goes to direct comparison paradigms, such as ABX (8) and XAB (7) metrics, where the evaluator listens to a reference and must decide which of the tested models produces the most faithful accent. This preference based approach is reinforced by the use of the Comparison MOS (CMOS) (4), the Prosody MOS (PMOS) (4) and the Multiple Stimuli with Hidden Reference and Anchor (MUSHRA) (5) metrics. Finally, specialized metrics such as Voice Similarity Score (VSS) (2) and Empirically Grounded, Well-Anchored scale (EGWA) (2) are occasionally used to capture specific vocal characteristics. Ultimately, this diverse range of subjetive metrics highlights that a comprehensive evaluation must balance general speech quality, speaker identity, and direct listener preference.

3.3 Architectures for Accent Transfer

To systematically analyze the state-of-the-art in AT, understand the technological progress of the field and answer the **(RQ1)**, we extracted architecture details and evaluation metrics from recent literature. Based on the core mechanisms used to achieve disentanglement and synthesis, we propose a classification that categorizes recent models into three main criteria:

1. Disentanglement architectures, relying on mechanisms like VAEs and Seq2Seq attention;
2. High-fidelity Generative and Diffusion architectures; and
3. Self-Supervised Learning (SSL) and Discrete Unit architectures.

To map the current scenario of accent transfer, Table 5 illustrates the frequency of the main architectures identified in our review of the literature, highlighting the predominance of architectures such as Tacotron 2 and HuBERT.

For each category, we compiled key objective metrics (e.g., WER, CER and Cosine Similarity) and subjective metrics (e.g., MOS, NMOS) from the best-performing reported by the original authors. It is important to note that direct numerical comparison across different models is limited due to variations in datasets, baselines and evaluator demographics used in each study. Therefore, the compiled metrics are intended to map the current capabilities and comprehend trends

Table 5. Main Architectures Mapped in the Literature Review.

Architectures	Qty	Architectures	Qty
Tacotron 2	12	MaskCycleGAN-VC	2
HuBERT	6	ML-VAE	2
GMM	5	NMF	2
FastPitch	3	Tacotron	2
FastSpeech 2	3	U-Net	2
Transformer	3	VoiceGrad	2
VITS	3	VQ-VAE	2
CVAE	2	wav2vec 2.0	2
FACodec	2	WavLM	2

with each architecture progression rather than serve as an absolute benchmark.

One of the goals of AT is to modify the speech of a non-native speaker (L2) to sound like a native speaker (L1) while preserving the original linguistic content and speaker identity (timbre).

Historically, this task was addressed through the articulatory modeling and concatenative synthesis proposed by Felps *et al*. [2012]; Aryal and Gutierrez-Osuna [2014a]. These pioneering methods attempted to map the characteristics of the physical vocal tract. Although theoretically valid for isolating pronunciation, these techniques failed to generalize complex prosody between different speakers, resulting in high error rates and robotic synthesis. Contemporary studies show that pure articulatory representations only become acoustically viable when combined with deep latent representations [Siriwardena *et al*., 2024].

In summary, regarding **(RQ2)**, accents are primarily represented either via continuous prosodic and content features, such as pitch (F_0) contours, phone-level durations, and bottleneck features (BNFs) or d-vectors extracted from speaker or accent classifiers trained with Generalized End-to-End (GE2E) loss [Watanabe and Kameoka, 2024; Ding *et al*., 2022] or discrete acoustic tokens quantized via Vector Quantization (VQ) from self-supervised learning (SSL) backbones like HuBERT [Cheng *et al*., 2025; Zhong *et al*., 2026]. Regarding training techniques, the field has evolved from parallel frame-level alignment utilizing Vocal Tract Length Normalization (VTLN) to unsupervised, non-parallel, and zero-shot training paradigms [Aryal and Gutierrez-Osuna, 2014b; Ding *et al*., 2022]. These approaches employ information bottlenecks, Multi-Level Variational Autoencoders (ML-VAEs), or factorized speech codecs, for example, FACodec, alongside adversarial losses (GRL) to mathematically decouple content, speaker timbre, and prosodic style, ultimately considering AT as an in-context translation task within a discrete language modeling framework [Huang *et al*., 2024; Halychanskyi *et al*., 2026a].

3.4 The Acoustic & Latent Architectures: Seq2Seq & VAEs

Historically, early AT systems relied on traditional VC techniques that demanded strictly parallel data, recordings of the exact same text spoken by both native (L1) and non-native (L2) speakers. However, aligning these parallel recordings frame-by-frame (e.g., via Dynamic Time Warping) was not only highly restrictive for data collection but also inherently flawed, as acoustic alignment inevitably leaked the non-native speaker's original accent into the converted speech [Aryal and Gutierrez-Osuna, 2014b].

The first major methodological advance to overcome this data bottleneck occurred with the introduction of Phonetic Posteriogram (PPG) coupled with Sequence-to-Sequence (S2S) architectures and Variational Autoencoder (VAE). Extracted from speaker-independent ASR models, PPG represent the pure phonetic content of speech, allowing frames to be paired based on linguistic similarity rather than acoustic proximity. Studies such as Zhao *et al*. [2018a]; Zhao and Gutierrez-Osuna [2019] demonstrated that this PPG-based alignment significantly reduces the reliance on strictly parallel data while improving synthesis quality, reporting relative increases in acoustic quality (MOS) ranging from 22% to 20% compared to baseline methods.

Studies such as Huang and Toda [2023], using these methods achieved high levels of accent suppression, reaching an Accentedness score of 4.36 (on a 1-5 scale) . However, these early VAE architectures often sacrificed general audio quality and speaker retention, leading to high WER > 40% and moderate Voice Similarity (57.3%). Later, the Accent-VITS [Ma *et al*., 2023] architecture demonstrated that conditional VAE modifications could improve fidelity, reaching an NMOS of 3.91 and SMOS of 3.95, although it still struggled with zero-shot flexibility.

To further address prosody mismatches and sequence alignment, models incorporated explicit attention mechanisms and non-autoregressive structures. For example, applying Concentrated Attention Zang *et al*. [2022] improved alignment stability between input sequences and mel-spectrograms, yielding an QMOS of 4.06. Similarly, non-autoregressive networks like the Full STS architecture [Nechaev and Kosyakov, 2024] minimized latency while maintaining high voice similarity, MOS of 4.30, and low transcription errors (CER of 3.52%). Additionally, advancements in vector quantization, such as the Phoneme-level RVQ-VAE [Karapiperis *et al*., 2024], successfully disentangled prosody within a discrete latent space, achieving an impressive NMOS of 4.28 and a WER of 3.18%.

3.5 High-Fidelity Generative Architectures: Diffusion & Flow-Matching

Despite the optimizations achieved by advanced attention and discrete quantization, traditional latent architectures fundamentally struggled with acoustic over-smoothing [Sisman *et al*., 2020]. Because standard VAEs and parametric models optimize for average spectral distributions using minimum mean square error or maximum likelihood, they often generated muffled speech lacking high-frequency acoustic and temporal details [Sisman *et al*., 2020].

To eliminate this structural limitation, the field shifted towards high-fidelity Generative and Diffusion Models. Early generative attempts, such as MaskCycleGAN [Kaneko *et al*., 2021], employed adversarial training to learn non-parallel mappings and restore spectral variance. While this approach stabilized intelligibility, reducing WER to between 14.8% and 11.6%, it still yielded a limited NMOS of 2.76 Correa *et al*. [2023], highlighting the need for more stable generative frameworks.

The real breakthrough came with Diffusion Models.

Using stochastic denoising and flow-matching, these models synthesize complex distributions with high precision. Modern latent diffusion models such as LatentVoiceGrad [Kameoka *et al*., 2025], significantly reduced error rates, achieving a CER of 2.99% while maximizing spectral quality (PMOS of 3.93) and SECS of 0.844. This approach allowed for refined control with frameworks such as FAC-FACODEC reaching a native conversion probability of 89.86% [Halychanskyi *et al*., 2026a]. Furthermore, diffusion-based approaches have been extensively adapted for low-resource scenarios. The HQ-SVC framework [Bai *et al*., 2025a] decoupled audio codecs to extract content and speaker embeddings independently, using a diffusion model to refine output and achieving a high NMOS of 4.215 even in challenging zero-shot tasks.

3.6 The State-of-the-Art: SSL & Discrete Tokens

Currently, the state-of-the-art (SOTA) in AT is characterized by a transition from continuous audio formats to SSL and discrete acoustic tokens. To illustrate this progression, Table 6 presents a technological cascade of the architectural evolution and its corresponding performance achievements. Previously, continuous formats like mel-spectrograms limited performance due to acoustic mixing, a scenario where speaker characteristics and spoken words are inherently entangled. Modern architectures solve this issue directly by applying specific methods to isolate the semantic meaning of the speech from stylistic elements such as emotions and accents. As detailed in Table 6, this structural evolution is clearly reflected in the test results, demonstrating a progressive improvement in zero-shot capabilities, intelligibility, and overall preservation of speakers and accent transfering.

Systems such as wav2vec 2.0 and HuBERT establish the basis for this isolation by applying an Information Bottleneck through Vector Quantization (VQ). However, a recent study by Zhong *et al*. [2026] demonstrates that naive codebook size reduction is insufficient for complete disentanglement. Instead, the choice of network layers and the application of ASR supervision are critical to effectively purge accent information from Discrete Speech Representation Tokens (DSRTs). Guided by these principles, newer frameworks treat AT as a straightforward token prediction task.

Given these advancements, the research divides into two primary paths: STS and TTS. STS methods process the audio signal directly, altering voice features without generating text. STS framework such as USD-AC framework [Huang *et al*., 2024] forces a strict mathematical split between content and style using adversarial training and Gradient Reversal Layers (GRL). When facing unseen foreign accents from the Speech Accent Archive (SAA), USD-AC maintained phonetic accuracy with a CER of 3.27% while reaching an Accent Embedding Similarity (AES) of 0.846. Similarly, the SEVC model [Wang *et al*., 2025] uses a graph-based strategy to mathematically limit the overlap between tone and text, achieving a MOS of 4.05 and a WER of 6.68% on the LibriSpeech dataset.

To address limitations in rhythm modeling and controllability, recent frameworks incorporate diffusion and flow-matching mechanisms evaluated under rigorous multi-accent conditions. DLM-AN [Bai *et al*., 2026a] utilizes masked discrete diffusion over WavLM-derived tokens, employing a Common Token Predictor to selectively reuse native-like source tokens. Tested on an extended L2-ARCTIC dataset, covering seven accents (Arabic, Chinese, Hindi, Korean, Spanish, Vietnamese and Native American English), it achieves a WER of 10.64% and a SECS of 0.8521 when generating native targets from scratch. Similarly, Cosy-Accent [Bai *et al*., 2026b] resolves the trade-off between prosodic naturalness and temporal rigidity through a non-autoregressive flow-matching architecture with explicit duration control via position scaling. Evaluated on the same extended L2-ARCTIC corpus, it entirely circumvents the need for real non-native training data through source-synthesis, securing a WER of 13.26% and a SECS of 0.8291

Recent work integrates these approaches, combining the strengths of Large Language Models (LLMs) and Neural Audio Codecs. Factorized codecs like FAC-FACodec [Halychanskyi *et al*., 2026a] establish this division at the core level by operating strictly on quantized content latents to decouple pronunciation from prosody and timbre. When evaluated on the L2-ARCTIC benchmark at maximum conversion strength, it maintains a low WER of 10.0% while achieving a high speaker similarity of 0.91. Building on this, SpeechAccentLLM [Cheng *et al*., 2025] employs a CTC-guided SpeechCodeVAE to ensure token locality. Tested on L2-ARCTIC, 20 non-native speakers across six L1 backgrounds, this joint FAC-TTS generative engine reduces the zero-shot WER from a 14.4% baseline to 9.1% while delivering a fluency CMOS of 4.07. Expanding this unified approach, Vevo2 [Zhang *et al*., 2026] uses a 6.25 Hz chromagram-based prosody tokenizer alongside a 12.5 Hz content-style tokenizer. This flow-matching architecture achieves robust timbre disentanglement across domains, reaching a zero-shot WER of 3.01% and a speaker similarity of 0.755 on the SeedTTS evaluation set.

Earlier architectures required strictly parallel corpora (mapped L1-L2 pairs) to learn the transformation function. In contrast, modern discrete-token generative models extract semantic representations via SSL and condition the generative decoder independently, eliminating the requirement for paired data and shifting AT to an unconstrained language modeling task.

Consequently, a newly emerging frontier bypasses traditional retraining entirely through latent space manipulation. Methods such as Activation Steering [Yang and Hansen, 2026] apply post-hoc vectors to the intermediate layers (e.g., layer 15 of Qwen3-TTS) during inference. When evaluated on Mandarin-accented English from L2-ARCTIC corpus, this training-free approach neutralized accents by dropping the WER to 2.24% while strictly preserving speaker's timbre, achieving a similarity score of 0.76. Likewise, the Accent Vector approach [Lertpetchpun *et al*., 2026] extracts task vectors from XTTS-v2 backbones fine-tuned via LoRA. This enables compositional accent manipulation in multilingual zero-shot scenarios without accented training data, achieving a Speaker Similarity (SSIM) of 0.89 in cross-lingual settings like English-accented Spanish.

The state-of-the-art presents significant architectural

Table 6. Evolution of the State-of-the-Art in Accent Transfer. The "Training Constraint" column clarifies whether the architecture requires strictly paired (parallel) data or operates in a zero-shot/unsupervised (non-parallel) manner.

Model / Reference	Architecture	Training Constraint	Evaluation Dataset(s)	Intelligibility (WER/CER)	Naturalness / Quality	Similarity / Accent
Acoustic & Latent Architectures						
Huang et al. Huang and Toda [2023]	VAE + PPGs	Parallel Required	L2-ARCTIC	WER: > 40.0%	MOS: 3.66	Voice Sim: 57.3%
Accent-VITS Ma *et al*. [2023]	Conditional VAE	Non-Parallel	Multi-Accent	–	NMOS: 3.91	SMOS: 3.95
Full STS Nechaev and Kosyakov [2024]	Non-Autoregressive Seq2Seq	Non-Parallel	L2-ARCTIC, NPTEL	CER: 3.52%	MOS: 4.04	MOS (Sim): 4.30
Karapiperis et al. Karapiperis *et al*. [2024]	Phoneme-level RVQ-VAE	Non-Parallel	VCTK, LibriTTS	WER: 3.18%	NMOS: 4.28	Cosine Sim: 0.949
High-Fidelity Generative Models						
MaskCycleGAN Kaneko *et al*. [2021]	CycleGAN / Adversarial	Non-Parallel	VCC 2018	WER: 11.6% – 14.8%	NMOS: 2.76	–
HQ-SVC Bai *et al*. [2025a]	Decoupled Codec + Diffusion	Zero-Shot	OpenSinger, M4Singer	STOI: 0.799	NMOS: 4.215	Cosine Sim: 0.627
LatentVoiceGrad Kameoka *et al*. [2025]	Latent Diffusion	Non-Parallel	VCTK	CER: 2.99%	pMOS: 3.93	SECS: 0.844
The SOTA - SSL, Discrete Tokens & LLMs						
USD-AC Huang *et al*. [2024]	SSL (HuBERT) + GRL	Non-Parallel	L2-ARCTIC, SAA	CER: 3.27%	–	AES: 0.846
SEVC Wang *et al*. [2025]	SSL + Structural Entropy	Non-Parallel	LibriSpeech	WER: 6.68%	MOS: 4.05	Sim: 2.95 (1-5)
DLM-AN Bai *et al*. [2026a]	WavLM + Discrete Diffusion	Zero-Shot	L2-ARCTIC (7 Accents)	WER: 10.64%	–	SECS: 0.8521
CosyAccent Bai *et al*. [2026b]	Flow-Matching + Source-Synth	Non-Parallel	L2-ARCTIC	WER: 13.26%	–	SECS: 0.8291
FAC-FACodec Halychanskyi *et al*. [2026a]	Factorized Codec + Diffusion	Zero-Shot	L2-ARCTIC	WER: 10.0%	–	Sim: 0.910
SpeechAccentLLM Cheng *et al*. [2025]	LLM + CTC-guided Codebooks	Zero-Shot	L2-ARCTIC	WER: 9.1%	CMOS: 4.07	–
Vevo2 Zhang *et al*. [2026]	Flow-Matching + Multi-Tokenizer	Zero-Shot	SeedTTS	WER: 3.01%	N-CMOS: +1.47	Sim: 0.755
Activation Steering Yang and Hansen [2026]	Post-hoc LLM Latent Control	Training-Free	L2-ARCTIC	WER: 2.24%	–	Sim: 0.760
Accent Vector Lertpetchpun *et al*. [2026]	Task Vectors (XTTS-v2 + LoRA)	Training-Free	Multilingual (Zero-shot)	–	–	SSIM: 0.890

trade-offs that dictate their application. LLMs achieve unprecedented prosodic naturalness, but they are inherently prone to autoregressive decoding instability. As Han *et al*. [2024] points out, "sampling based decoding strategies bring astonishing diversity to generation, but also pose robustness issues such as typos, omissions and repetition". This makes them unreliable for strict real-time applications like virtual assistants.

Conversely, non-autoregressive flow-matching models solve this temporal instability and offer low latency. Highlighting this advantage, Nechaev and Kosyakov [2024] states that a non-autoregressive architecture "allows real-time accent conversion, and avoids the error accumulation effect associated with sequential output generation". Furthermore, for precise temporal alignment, these models provide a targeted solution. According to Bai *et al*. [2026b], this architecture "implicitly models rhythm for naturalness while offering explicit control over the total output duration, making it adept for tasks like dubbing". Ultimately, the choice of architecture now depends on balancing the need for low-latency, stable inference against the demand for absolute, LLM-driven generative naturalness.

Ultimately, the capacity to blend high intelligibility with zero-shot, inference-time control solidifies these architectures as the definitive state-of-the-art.

3.7 Challenges

Despite significant advances in architectures for accent transfer, the literature reveals substantial challenges that still need to be overcome. Based on the systematic mapping of the analyzed studies **(RQ4)**, this SLR identified three major gaps: (i) a pronounced linguistic bias in training data, (ii) a lack of standardization in evaluation metrics, and (iii) persistent architectural limitations regarding speaker-accent disentanglement and zero-shot generalization.

The first challenge lies in the literature's excessive and disproportionate focus on English accents and variations. A gap is observed in the modeling of other languages and their respective regional variations. Languages with a great diversity of dialects, such as Portuguese in its Brazilian and European variants, lack resources, clean datasets, and in-depth studies. Similarly, there is a notable scarcity of research focused on other low-resource languages and dialects, extending to African languages (Swahili, Zulu, Yoruba, Igbo), indigenous Latin American languages, the multiplicity of Arabic dialects, regional variations of Hindi, and diverse accents of Mandarin [Halychanskyi *et al*., 2026b]. It is crucial to emphasize that these specific languages are highlighted as examples based on the data collected for this review. Nonetheless, the scope of this scientific gap is not limited to them; all other languages and regional dialects not explicitly cited or found in the current literature are equally important and must be included in future studies to achieve true linguistic inclusivity.

The second challenge concerns the evaluation of the quality and similarity of the generated accents. Currently, there is a profound lack of consensus and standardization in the metrics used by the scientific community. The evaluation of accent similarity in speech synthesis has been underexplored, and traditional methods present flaws when applied to linguistic variations. Metrics frequently inherited from general audio quality or intelligibility evaluations, such as WER, CER and UTMOS, present significant limitations and are not recommended for measuring accent similarity and fidelity [Zhong *et al*., 2025a]. These metrics tend to exhibit a punitive bias against underrepresented accents, penalizing perfectly valid pronunciation characteristics and interpreting a regional accent as a recognition error rather than a legitimate phonetic identity [Zhong *et al*., 2025a]. Consequently, the field urgently needs to transition towards objective pronunciation-based metrics, such as distances between vowel formants or phonetic posteriorgrams, and towards more refined subjective evaluation approaches that truly isolate linguistic cues.

This lack of standardized metrics also leads to significant methodological and evaluative inconsistencies across the literature. When synthesizing the results, we noticed widespread variation in the datasets used—ranging from L2-ARCTIC and VCTK to proprietary or synthetic corpora as well as in the chosen baseline architectures and linguistic pairs. Because the studies are so structurally different, conducting a quantitative meta-analysis, such as directly aggregating absolute WER or MOS scores, is not scientifically viable. Therefore, rather than comparing raw numbers, our review focuses on tracing how the architectures have evolved and pinpointing relative improvements within specific approaches. Ultimately, this fragmented landscape underscores

the need for unified benchmarking frameworks and standard evaluation protocols in AT research.

The third challenge involves the architectural bottlenecks of state-of-the-art models, particularly regarding the persistent entanglement of speech attributes and data dependency. Accent transfer remains imperfect because accent and speaker identity are organically entangled; modifying intonation, durations, and formant frequencies without distorting the original speaker's timbre is extremely complex Halychanskyi *et al*. [2026b]. Historically, early models required strictly parallel corpora (e.g., L2-ARCTIC) to learn the transformation function. While the transition from VAEs and GANs to SSL and LLMs mitigated the need for parallel data, it introduced new representational challenges.

To prevent timbre leakage in the latent space, modern architectures employ these VQ information bottlenecks alongside structural entropy to strictly penalize the retention of speaker identity within content representations [Melechovsky *et al*., 2024c; Wang *et al*., 2025], utilizing contrastive training to explicitly separate speaker identity from environmental and linguistic factors [Matos *et al*., 2026]. However, recent analyses demonstrate that a naive reduction in codebook size is insufficient to perfectly separate accent from phonetics, often degrading intelligibility [Zhong *et al*., 2026]. Furthermore, while LLMs offer high naturalness, they suffer from inference instability [Han *et al*., 2024; Cheng *et al*., 2025], and zero-shot computation reaches its limits when facing Out-of-Distribution (OOD) accents with vastly different prosodic structures.

Consequently, regarding research trends and open gaps **(RQ5)**, the next methodological frontiers for Accent Transfer point not only toward inference-time control and non-autoregressive decoding, but also toward linguistic inclusivity and architectural adaptability. Future architectures must be inherently language-agnostic and easily adaptable to any language without requiring massive, language-specific accented corpora. Currently, models are moving towards Activation Steering and Task Vectors, which allow for post-hoc accent neutralization and cross-lingual generalization without retraining or requiring accented data [Yang and Hansen, 2026; Lertpetchpun *et al*., 2026]. Additionally, integrating flow-matching mechanisms promises explicit control over duration and rhythm, offering a pathway to overcome the temporal rigidity of current models [Bai *et al*., 2026a; Halychanskyi *et al*., 2026a].

4 Conclusion

This systematic review of the literature provided a comprehensive analysis of state-of-the-art techniques for accent transfer and conversion within the TTS and VC architectures. By analyzing 91 articles, it was possible to map the evolution of recent architectures and identify the primary methodological trends in the field.

However, any claims regarding a definitive "state-of-the-art" must be exercised with great caution. The current literature lacks a unified benchmarking framework, meaning that direct numerical comparisons of architectures are often confounded by varying training datasets, differing baseline setups, and subjective evaluation protocols. An architecture that achieves superior metrics on a specific parallel corpus like L2-ARCTIC may not necessarily maintain that supremacy in zero-shot, real-world scenarios. Therefore, while modern SSL and discrete token models represent the current technological frontier, superiority in isolated metrics does not guarantee universal architectural supremacy, highlighting the critical need for standardized evaluation challenges.

Despite the noticeable advancements driven by deep learning models, it is concluded that the research field remains disproportionately centered on the English language, neglecting a wide range of other languages and their rich regional variations. Furthermore, the evaluation of accent similarity is critical and lacks standardization. Reliance on traditional intelligibility metrics, such as WER and CER, has proven inadequate. These metrics tend to penalize pronunciation characteristics that are perfectly valid in underrepresented accents, treating speech variations as recognition errors.

Regarding the formulated research questions, this review establishes that modern AT architectures have shifted entirely toward discrete tokens and SSL **(RQ1)**, utilizing vector quantization and generative codecs to disentangle accent from timbre **(RQ2)**. Evaluation remains heavily reliant on WER and MOS, though pronunciation-specific metrics are urgently needed **(RQ3)**. The primary challenges identified include architectural bottlenecks in timbre leakage and severe linguistic bias toward English **(RQ4, RQ5)**, which directly restricts the performance and scalability of accent transfer in low-resource and non-English languages **(RQ6)**.

Declarations

Acknowledgements

In the face of the challenges inherent in scientific research, one realizes that all we have to decide is what to do with the time that is given us. Thus, I am deeply grateful to those who decided to dedicate a part of their time to contribute to the development of this study and this work would not have been possible without the support by SiDi Intelligence & Innovation Center.

Funding

This research was funded by SiDi.

Authors' Contributions

All authors contributed effectively to the development of this research and the preparation of this manuscript.

Competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Availability of research artifacts

All databases and literature sources used in this review are publicly accessible. The search strategies and selection criteria are fully detailed within the manuscript to ensure the reproducibility of the results.

Use of Artificial Intelligence

Perplexity AI was used exclusively post-hoc, after the complete identification, screening and selection of the 91 included studies, to assist the authors in syntactically organizing the extracted lit-

erature, clustering the already-identified architectural patterns, and formatting the reference metadata into the standardized BibTeX format. Generative AI tools were used to correct English grammar and phrasing, as the authors native language is Portuguese. Additionally, Nano Banana 2 was used exclusively to generate the architectural evolution flowchart. All AI-generated suggestions, edits, and visual content were manually reviewed and verified by the authors.

References

Aggarwal, S., Uttam, S., Garg, S., Garg, S., Jain, K., and Aggarwal, S. (2025). Advancements in end-to-end audio style transformation: A differentiable approach for voice conversion and musical style transfer. *AI*, 6(1). DOI: 10.3390/ai6010016.

Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F. M., and Weber, G. (2020). Common voice: A massively-multilingual speech corpus. DOI: 10.48550/arxiv.1912.06670.

Aryal, S. and Gutierrez-Osuna, R. (2014a). Accent conversion through cross-speaker articulatory synthesis. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7694–7698. DOI: 10.1109/ICASSP.2014.6855097.

Aryal, S. and Gutierrez-Osuna, R. (2014b). Can voice conversion be used to reduce non-native accents? In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7879–7883. DOI: 10.1109/ICASSP.2014.6855134.

Azevedo, C. d. A. *et al*. (2026). Texto ou fala? In *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*, chapter 2. BPLN. Available at:`https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/fala/cap-fala/cap-fala.html`.

Badlani, R., Łancucki, A., Shih, K. J., Valle, R., Ping, W., and Catanzaro, B. (2021). One tts alignment to rule them all. DOI: 10.1109/icassp43922.2022.9747707.

Bai, B., Geng, Y., Wang, F., Wang, C., Guo, P., Gao, Y., and Li, Y. (2025a). Hq-svc: Towards high-quality zero-shot singing voice conversion in low-resource scenarios. DOI: 10.1609/aaai.v40i36.40249.

Bai, Q., Du, Y., Ko, T., Wang, S., Wang, Y., and Li, H. (2026a). Controllable accent normalization via discrete diffusion. DOI: 10.21437/interspeech.2026-1056.

Bai, Q., Inoue, S., Wang, S., Jiang, Z., Wang, Y., and Li, H. (2025b). Accent normalization using self-supervised discrete tokens with non-parallel data. DOI: 10.21437/interspeech.2025-2520.

Bai, Q., Shi, S., Wang, S., Ju, Y., Wang, Y., and Li, H. (2026b). Cosyaccent: Duration-controllable accent normalization using source-synthesis training data. DOI: 10.1109/icassp55912.2026.11460921.

Bai, Q., Wang, S., Liu, Z., Zhang, M., Rao, W., Wang, Y., and Li, H. (2024). Diffusion-based method with tts guidance for foreign accent conversion. In *2024 IEEE 14th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pages 284–288. DOI: 10.1109/ISCSLP63861.2024.10800623.

Casanova, E., Davis, K., Gölge, E., Göknar, G., Gulea, I., Hart, L., Aljafari, A., Meyer, J., Morais, R., Olayemi, S., and Weber, J. (2024). Xtts: a massively multilingual zero-shot text-to-speech model. DOI: 10.21437/interspeech.2024-2016.

Casanova, E., Weber, J., Shulby, C., Junior, A. C., Gölge, E., and Ponti, M. A. (2023). Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone. DOI: 10.48550/arxiv.2112.02418.

Chen, G., Chai, S., Wang, G.-B., Du, J., Zhang, W.-Q., Weng, C., Su, D., Povey, D., Trmal, J., Zhang, J., Jin, M., Khudanpur, S., Watanabe, S., Zhao, S., Zou, W., Li, X., Yao, X., Wang, Y., You, Z., and Yan, Z. (2021). GigaSpeech: An evolving, multi-domain ASR corpus with 10,000 hours of transcribed audio. In *Interspeech 2021*, pages 3670–3674. ISCA. DOI: 10.21437/Interspeech.2021-1965.

Chen, L.-W., Watanabe, S., and Rudnicky, A. (2022). A unified one-shot prosody and speaker conversion system with self-supervised discrete speech units. DOI: 10.1109/icassp49357.2023.10095565.

Chen, S., Liu, S., Zhou, L., Liu, Y., Tan, X., Li, J., Zhao, S., Qian, Y., and Wei, F. (2024a). Vall-e 2: Neural codec language models are human parity zero-shot text to speech synthesizers. DOI: 10.48550/arxiv.2406.05370.

Chen, X., Pei, J., Xue, L., and Zhang, M. (2024b). Transfer the linguistic representations from tts to accent conversion with non-parallel data. DOI: 10.1109/icassp48485.2024.10447205.

Cheng, Z., Zhang, G., Tu, Z., Song, Y., Mao, S., Jiao, X., Li, J., Guo, Y., and Wu, J. (2025). Speechaccentllm: A unified framework for foreign accent conversion and text to speech. DOI: 10.48550/arxiv.2507.01348.

Cheripally, S. (2024). A unified model for voice and accent conversion in speech and singing using self-supervised learning and feature extraction. DOI: 10.48550/arxiv.2412.08312.

Chung, J. S., Nagrani, A., and Zisserman, A. (2018). Voxceleb2: Deep speaker recognition. In *Interspeech 2018*, pages 1086–1090. DOI: 10.21437/Interspeech.2018-2812.

Churchwell, C., Morrison, M., and Pardo, B. (2024). High-fidelity neural phonetic posteriorgrams. DOI: 10.1109/icasspw62465.2024.10669905.

Collet, V. A. (2020). Variação linguistica ao longo da vida em situação de contato dialetal: a variação do rótico em coda e das oclusivas dentais em ataque. Master's thesis, PUCRS.

Correa, I. L., Ueno, S., and Lee, A. (2023). Accent-preserving voice conversion between native-nonnative speakers for second language learning. In *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1179–1186. DOI: 10.1109/APSIPAASC58517.2023.10317245.

Das, S., Williams, J., and Lai, C. (2022). Analysis of voice conversion and code-switching synthesis using vq-vae. DOI: 10.48550/arxiv.2203.14640.

Ding, S., Zhao, G., and Gutierrez-Osuna, R. (2022). Accentron: Foreign accent conversion to arbitrary non-native

speakers using zero-shot learning. *Comput. Speech Lang.*, 72(C):17. DOI: 10.1016/j.csl.2021.101302.

Felps, D., Geng, C., and Gutierrez-Osuna, R. (2012). Foreign accent conversion through concatenative synthesis in the articulatory domain. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(8):2301–2312. DOI: 10.1109/TASL.2012.2201474.

Felps, D. and Gutierrez-Osuna, R. (2010). Developing objective measures of foreign-accent conversion. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(5):1030–1040. DOI: 10.1109/TASL.2009.2038818.

Halychanskyi, Y., Churchwell, C., Wen, Y., and Kindratenko, V. (2026a). Fac-facodec: Controllable zero-shot foreign accent conversion with factorized speech codec. DOI: 10.1109/icassp55912.2026.11461516.

Halychanskyi, Y., Guo, J. S., and Kindratenko, V. (2026b). Accent conversion: A problem-driven survey of sociolinguistic and technical constraints. DOI: 10.48550/arxiv.2604.27281.

Han, B., Zhou, L., Liu, S., Chen, S., Meng, L., Qian, Y., Liu, Y., Zhao, S., Li, J., and Wei, F. (2024). Vall-e r: Robust and efficient zero-shot text-to-speech synthesis via monotonic alignment. DOI: 10.48550/arxiv.2406.07855.

Huang, J.-H., Lee, W.-T., and Wu, C.-H. (2024). Usd-ac: Unsupervised speech disentanglement for accent conversion. In *Interspeech 2024*, pages 4388–4392. DOI: 10.21437/Interspeech.2024-664.

Huang, W.-C. and Toda, T. (2023). Evaluating methods for ground-truth-free foreign accent conversion. DOI: 10.1109/apsipaasc58517.2023.10317592.

Inoue, S., Wang, S., Wang, W., Zhu, P., Bi, M., and Li, H. (2025). Macst: Multi-accent speech synthesis via text transliteration for accent conversion. DOI: 10.1109/icassp49660.2025.10888195.

Jiang, Z., Ren, Y., Ye, Z., Liu, J., Zhang, C., Yang, Q., Ji, S., Huang, R., Wang, C., Yin, X., Ma, Z., and Zhao, Z. (2023). Mega-tts: Zero-shot text-to-speech at scale with intrinsic inductive bias. DOI: 10.48550/arxiv.2306.03509.

Joglekar, A., Singh, D., Bhatia, R. R., and Umesh, S. (2025). Ez-vc: Easy zero-shot any-to-any voice conversion. In Christodoulopoulos, C., Chakraborty, T., Rose, C., and Peng, V., editors, *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 19768–19774, Suzhou, China. Association for Computational Linguistics. DOI: 10.18653/v1/2025.findings-emnlp.1077.

Kahn, J., Riviere, M., Zheng, W., Kharitonov, E., Xu, Q., Mazare, P., Karadayi, J., Liptchinsky, V., Collobert, R., Fuegen, C., Likhomanenko, T., Synnaeve, G., Joulin, A., Mohamed, A., and Dupoux, E. (2020). Libri-light: A benchmark for asr with limited or no supervision. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7669–7673. IEEE. DOI: 10.1109/icassp40776.2020.9052942.

Kalita, A. K. and Patra, R. (2023). Prosody-vc: Voice style transfer with speech prosody. In *2023 IEEE Guwahati Subsection Conference (GCON)*, pages 01–05. DOI: 10.1109/GCON58516.2023.10183559.

Kameoka, H., Kaneko, T., Tanaka, K., Hojo, N., and Seki, S. (2024). Voicegrad: Non-parallel any-to-many voice conversion with annealed langevin dynamics. DOI: 10.1109/taslp.2024.3379901.

Kameoka, H., Kaneko, T., Tanaka, K., and Kondo, Y. (2025). Latentvoicegrad: Nonparallel voice conversion with latent diffusion/flow-matching models. DOI: 10.1109/taslpro.2025.3613926.

Kameoka, H., Tanaka, K., Kwasny, D., Kaneko, T., and Hojo, N. (2020). Convs2s-vc: Fully convolutional sequence-to-sequence voice conversion. DOI: 10.1109/taslp.2020.3001456.

Kaneko, T., Kameoka, H., Tanaka, K., and Hojo, N. (2021). Maskcyclegan-vc: Learning non-parallel voice conversion with filling in frames. DOI: 10.1109/icassp39728.2021.9414851.

Karapiperis, S., Ellinas, N., Vioni, A., Oh, J., Jho, G., Hwang, I., and Raptis, S. (2024). Investigating disentanglement in a phoneme-level speech codec for prosody modeling. DOI: 10.1109/slt61566.2024.10832258.

Kim, J., Kim, S., Kong, J., and Yoon, S. (2020). Glow-tts: A generative flow for text-to-speech via monotonic alignment search. DOI: 10.48550/arxiv.2005.11129.

Kitchenham, B. and Charters, S. (2007). Guidelines for performing systematic literature reviews in software engineering. 2.

Kobayashi, K. and Toda, T. (EasyChair, 2018). sprocket: Open-source voice conversion software. DOI: 10.29007/s4t1.

Koizumi, Y., Zen, H., Karita, S., Ding, Y., Yatabe, K., Morioka, N., Bacchiani, M., Zhang, Y., Han, W., and Bapna, A. (2023). Libritts-r: A restored multi-speaker text-to-speech corpus. DOI: 10.21437/interspeech.2023-1584.

Kominek, J. and Black, A. W. (2004). The CMU Arctic speech databases. In *5th ISCA Workshop on Speech Synthesis (SSW 5)*, pages 223–224. Available at:`https://www.isca-archive.org/ssw_2004/kominek04b_ssw.html`.

Kong, J., Kim, J., and Bae, J. (2020). Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. DOI: 10.48550/arxiv.2010.05646.

Lertpetchpun, T., Trachu, T., Lee, J., Feng, T., Byrd, D., and Narayanan, S. (2026). Accent vector: Controllable accent manipulation for multilingual tts without accented data. DOI: 10.48550/arxiv.2603.07534.

Li, W., Tang, B., Yin, X., Zhao, Y., Li, W., Wang, K., Huang, H., Wang, Y., and Ma, Z. (2020). Improving accent conversion with reference encoder and end-to-end text-to-speech. DOI: 10.48550/arxiv.2005.09271.

Li, Y. A., Han, C., Raghavan, V. S., Mischler, G., and Mesgarani, N. (2023). Styletts 2: Towards human-level text-to-speech through style diffusion and adversarial training with large speech language models. DOI: 10.52202/075280-0860.

Liu, S., Wang, D., Cao, Y., Sun, L., Wu, X., Kang, S., Wu, Z., Liu, X., Su, D., Yu, D., and Meng, H. (2020). End-to-end accent conversion without using native utterances. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages

6289–6293. DOI: 10.1109/ICASSP40776.2020.9053797.

Lorandi, A. *et al*. (2011). AquisiÇÃo da linguagem. *Revista Verba Volant*, 2:44–66. Available at:https://revistaseletronicas.pucrs.br/letronica/article/view/4212.

Lumban Tobing, P., Wu, Y.-C., Hayashi, T., Kobayashi, K., and Toda, T. (2019). Voice conversion with cyclernn-based spectral mapping and finely tuned wavenet vocoder. *IEEE Access*, 7:171114–171125. DOI: 10.1109/ACCESS.2019.2955978.

Luo, Y.-J. and Dixon, S. (2024). Posterior variance-parameterised gaussian dropout: Improving disentangled sequential autoencoders for zero-shot voice conversion. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11676–11680. DOI: 10.1109/ICASSP48485.2024.10447835.

Ma, L., Zhang, Y., Zhu, X., Lei, Y., Ning, Z., Zhu, P., and Xie, L. (2023). Accent-vits:accent transfer for end-to-end tts. DOI: 10.1007/978-981-97-0601-$3_1$7.

Matos, A. N., Junior, A. C., and Ponti, M. A. (2026). Contrastive and adversarial disentanglement for speaker representations in Brazilian Portuguese. In Souza, M., de Dios-Flores, I., Santos, D., Freitas, L., Souza, J. W. d. C., and Ribeiro, E., editors, *Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026) - Vol. 1*, pages 591–600, Salvador, Brazil. Association for Computational Linguistics. Available at:https://aclanthology.org/2026.propor-1.58/.

Melechovsky, J., Mehrish, A., Sisman, B., and Herremans, D. (2024a). Accent conversion in text-to-speech using multi-level vae and adversarial training. DOI: 10.1109/tencon61640.2024.10902878.

Melechovsky, J., Mehrish, A., Sisman, B., and Herremans, D. (2024b). Accented text-to-speech synthesis with a conditional variational autoencoder. DOI: 10.1109/tencon61640.2024.10902981.

Melechovsky, J., Mehrish, A., Sisman, B., and Herremans, D. (2024c). Dart: Disentanglement of accent and speaker representation in multispeaker text-to-speech. DOI: 10.48550/arxiv.2410.13342.

Min, D., Lee, D. B., Yang, E., and Hwang, S. J. (2021). Meta-stylespeech : Multi-speaker adaptive text-to-speech generation. DOI: 10.48550/arxiv.2106.03153.

Nagrani, A., Chung, J. S., and Zisserman, A. (2017). Voxceleb: A large-scale speaker identification dataset. In *Interspeech 2017*, pages 2616–2620. ISCA. DOI: 10.21437/Interspeech.2017-950.

Nakatani, H., Tobing, P. L., Takeda, K., and Toda, T. (2020). Cross-lingual voice conversion using a cyclic variational auto-encoder and a wavenet vocoder. pages 520–526. Available at:https://ieeexplore.ieee.org/document/9306335.

Nechaev, V. and Kosyakov, S. (2024). Non-autoregressive real-time accent conversion model with voice cloning. DOI: 10.48550/arxiv.2405.13162.

Nechaev, V. A. and Kosyakov, S. V. (2025). Accent conversion method with real-time voice cloning based on a nonautoregressive neural network model. *Russian Technological Journal*. DOI: 10.32362/2500-316x-2025-13-3-7-20.

Nguyen, T. N., Pham, N.-Q., and Waibel, A. (2022). Accent Conversion using Pre-trained Model and Synthesized Data from Voice Conversion. In *Interspeech 2022*, pages 2583–2587. DOI: 10.21437/Interspeech.2022-10729.

Nguyen, T. N., Pham, N. Q., and Waibel, A. (2024). Accent conversion using discrete units with parallel data synthesized from controllable accented tts. DOI: 10.21437/syndata4genai.2024-11.

Olatunji, T., Afonja, T., Yadavalli, A., Emezue, C. C., Singh, S., Dossou, B. F. P., Osuchukwu, J., Osei, S., Tonja, A. L., Etori, N., and Mbataku, C. (2023). Afrispeech-200: Pan-african accented speech dataset for clinical and general domain asr. DOI: 10.1162/tacl$_{a}$0627.

Onda, K., Imoto, K., Fukayama, S., Saito, D., and Minematsu, N. (2025). Prosodically enhanced foreign accent simulation by discrete token-based resynthesis only with native speech corpora. DOI: 10.21437/interspeech.2025-590.

Oushiro, L. (2021). A importância de estudos de avaliação e percepções sociolinguisticas. *Revista de Letras*, 1. DOI: 10.36517/revletras.40.1.1.

Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. (2015). Librispeech: An asr corpus based on public domain audio books. *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210. DOI: 10.1109/icassp.2015.7178964.

Prabhu, D., Jyothi, P., Ganapathy, S., and Unni, V. (2023). Accented speech recognition with accent-specific codebooks. In Bouamor, H., Pino, J., and Bali, K., editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7175–7188, Singapore. Association for Computational Linguistics. DOI: 10.18653/v1/2023.emnlp-main.444.

Qin, Z., Zhao, W., Yu, X., and Sun, X. (2024). Openvoice: Versatile instant voice cloning. DOI: 10.48550/arxiv.2312.01479.

Qinyu, X., Guanming, L., Jingjie, Y., Haibo, L., and Xiao, C. (2017). Speaker conversion using kernel non-negative matrix factorization. *The Journal of China Universities of Posts and Telecommunications*, 24:60–67. DOI: 10.1016/s1005-8885(17)60234-6.

Quamer, W. and Gutierrez-Osuna, R. (2024). Disentangling segmental and prosodic factors to non-native speech comprehensibility. DOI: 10.1109/taslpro.2025.3592320.

Radzikowski, K., Wang, L., and Nowak, R. (2021). Accent modification for speech recognition of non-native speakers using neural style transfer. *EURASIP Journal on Audio, Speech, and Music Processing*, 2021. DOI: 10.1186/s13636-021-00199-3.

Ren, Y., Hu, C., Tan, X., Qin, T., Zhao, S., Zhao, Z., and Liu, T.-Y. (2022). Fastspeech 2: Fast and high-quality end-to-end text to speech. DOI: 10.48550/arxiv.2006.04558.

Shen, J., Jia, Y., Chrzanowski, M., Zhang, Y., Elias, I., Zen, H., and Wu, Y. (2021). Non-attentive tacotron: Robust and controllable neural tts synthesis including unsupervised duration modeling. DOI: 10.48550/arxiv.2010.04301.

Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerry-Ryan, R., Saurous, R. A., Agiomyrgiannakis, Y., and Wu, Y. (2018). Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. DOI: 10.1109/icassp.2018.8461368.

Shrestha, A., Ghorshi, S., and Panahi, I. (2025). Overview of accent conversion techniques based on deep learning. In *2025 IEEE 6th International Conference on Image Processing, Applications and Systems (IPAS)*, volume CFP2540Z-ART, pages 1–6. DOI: 10.1109/IPAS63548.2025.10924484.

Siriwardena, Y. M., Swedlow, N., Howard, A., Gitterman, E., Darcy, D., Espy-Wilson, C., and Fanelli, A. (2024). Accent conversion with articulatory representations. DOI: 10.21437/interspeech.2024-1416.

Sisman, B., Yamagishi, J., King, S., and Li, H. (2020). An overview of voice conversion and its challenges: From statistical modeling to deep learning. DOI: 10.1109/taslp.2020.3038524.

Takashima, R., Aihara, R., Takiguchi, T., and Ariki, Y. (2013). Noise-robust voice conversion based on spectral mapping on sparse space. pages 71–75. Available at:`https://www.isca-archive.org/ssw_2013/takashima13_ssw.pdf`.

Takashima, R., Takiguchi, T., and Ariki, Y. (2012). Exemplar-based voice conversion in noisy environment. pages 313–317. DOI: 10.1109/SLT.2012.6424242.

Takeuchi, T. and Tatekura, Y. (2018). Speech intelligibility enhancement in noisy environments via voice conversion with glimpse proportion measure. In *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1713–1717. DOI: 10.23919/APSIPA.2018.8659456.

Ueda, L. H., de M. M. Marques, L. B., Simões, F. O., Neto, M. U., Runstein, F., Bó, B. D., and Costa, P. D. P. (2024). Exploring synthetic data for cross-speaker style transfer in style representation based tts. DOI: 10.21437/syndata4genai.2024-16.

Valin, J.-M. and Skoglund, J. (2019). Lpcnet: Improving neural speech synthesis through linear prediction. DOI: 10.1109/icassp.2019.8682804.

van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., and Kavukcuoglu, K. (2016). Wavenet: A generative model for raw audio. DOI: 10.48550/arxiv.1609.03499.

van den Oord, A., Li, Y., Babuschkin, I., Simonyan, K., Vinyals, O., Kavukcuoglu, K., van den Driessche, G., Lockhart, E., Cobo, L. C., Stimberg, F., Casagrande, N., Grewe, D., Noury, S., Dieleman, S., Elsen, E., Kalchbrenner, N., Zen, H., Graves, A., King, H., Walters, T., Belov, D., and Hassabis, D. (2017). Parallel wavenet: Fast high-fidelity speech synthesis. DOI: 10.48550/arxiv.1711.10433.

Wang, L., Yu, Z., Gao, S., Mao, C., Dong, L., and Huang, Y. (2025). Voice conversion via structural entropy. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. DOI: 10.1109/ICASSP49660.2025.10889652.

Watanabe, C. and Kameoka, H. (2024). Ge2e-ac: Generalized end-to-end loss training for accent classification. DOI: 10.1109/apsipaasc63619.2025.10848863.

Wu, D.-Y., Chen, Y.-H., and Lee, H.-Y. (2020). Vqvc+: One-shot voice conversion by vector quantization and u-net architecture. DOI: 10.21437/interspeech.2020-1443.

Yamagishi, J., Veaux, C., and MacDonald, K. (2019). Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92). DOI: 10.7488/ds/2645.

Yamashita, H., Okamoto, T., Takashima, R., Ohtani, Y., Takiguchi, T., Toda, T., and Kawai, H. (2025). Sequence-to-sequence voice conversion with weighted guided attention. *IEEE Access*, 13:216583–216595. DOI: 10.1109/ACCESS.2025.3647153.

Yang, M. and Hansen, J. H. L. (2026). Activation steering for accent-neutralized zero-shot text-to-speech. DOI: 10.48550/arxiv.2603.05977.

Yixin, L., Ning, C., Yuhang, X., and Yangyang, X. (2025). Zero-shot accent conversion model based on the knn regression of content features. *Journal of East China University of Science and Technology*, 51(4):497–504. DOI: 10.14135/j.cnki.1006-3080.20241009001.

Yook, D., Han, G., Chang, H.-P., and Yoo, I.-C. (2024). Cyclediffusion: Voice conversion using cycle-consistent diffusion models. *Applied Sciences*, 14(20). DOI: 10.3390/app14209595.

Zang, X., Xie, F., and Weng, F. (2022). Foreign accent conversion using concentrated attention. In *2022 IEEE International Conference on Knowledge Graph (ICKG)*, pages 386–391. DOI: 10.1109/ICKG55886.2022.00056.

Zeghidour, N., Luebs, A., Omran, A., Skoglund, J., and Tagliasacchi, M. (2021). Soundstream: An end-to-end neural audio codec. DOI: 10.1109/taslp.2021.3129994.

Zen, H., Dang, V., Clark, R., Zhang, Y., Weiss, R. J., Jia, Y., Chen, Z., and Wu, Y. (2019). Libritts: A corpus derived from librispeech for text-to-speech. DOI: 10.21437/interspeech.2019-2441.

Zhang, B., Lv, H., Guo, P., Shao, Q., Yang, C., Xie, L., Xu, X., Bu, H., Chen, X., Zeng, C., Wu, D., and Peng, Z. (2022). Wenetspeech: A 10000+ hours multi-domain mandarin corpus for speech recognition. DOI: 10.1109/icassp43922.2022.9746682.

Zhang, X., Zhang, J., Wang, Y., Wang, C., Chen, Y., Jia, D., Chen, Z., and Wu, Z. (2026). Vevo2: A unified and controllable framework for speech and singing voice generation. DOI: 10.1109/taslpro.2026.3671970.

Zhao, G., Ding, S., and Gutierrez-Osuna, R. (2019). Foreign Accent Conversion by Synthesizing Speech from Phonetic Posteriorgrams. In *Interspeech 2019*, pages 2843–2847. DOI: 10.21437/Interspeech.2019-1778.

Zhao, G. and Gutierrez-Osuna, R. (2019). Using phonetic posteriorgram based frame pairing for segmental accent conversion. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(10):1649–1660. DOI: 10.1109/TASLP.2019.2926754.

Zhao, G., Sonsaat, S., Levis, J., Chukharev-Hudilainen, E., and Gutierrez-Osuna, R. (2018a). Accent conversion using phonetic posteriorgrams. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5314–5318. DOI:

10.1109/ICASSP.2018.8462258.

Zhao, G., Sonsaat, S., Silpachai, A., Lucic, I., Chukharev-Hudilainen, E., Levis, J., and Gutierrez-Osuna, R. (2018b). L2-ARCTIC: A Non-native English Speech Corpus. In *Interspeech 2018*, pages 2783–2787. DOI: 10.21437/Interspeech.2018-1110.

Zhong, J., Liu, S., Wells, D., and Richmond, K. (2025a). Pairwise evaluation of accent similarity in speech synthesis. In *Interspeech 2025*, pages 2290–2294. ISCA. DOI: 10.21437/Interspeech.2025-2283.

Zhong, J., Richmond, K., Su, Z., and Sun, S. (2025b). Accentbox: Towards high-fidelity zero-shot accent generation. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE. DOI: 10.1109/icassp49660.2025.10888332.

Zhong, J., Wang, Y., Richmond, K., and Bell, P. (2026). Rethinking discrete speech representation tokens for accent generation. DOI: 10.48550/arxiv.2601.19786.

Zhou, X., Zhang, M., Zhou, Y., Wu, Z., and Li, H. (2023a). Accented text-to-speech synthesis with limited data. DOI: 10.1109/taslp.2024.3363414.

Zhou, X., Zhang, M., Zhou, Y., Wu, Z., and Li, H. (2025). Multi-scale accent modeling and disentangling for multi-speaker multi-accent text-to-speech synthesis. DOI: 10.1109/taslpro.2026.3657306.

Zhou, Y., Wu, Z., Zhang, M., Tian, X., and Li, H. (2023b). Tts-guided training for accent conversion without parallel data. *IEEE Signal Processing Letters*, 30:533–537. DOI: 10.1109/lsp.2023.3270079.

A Appendix: Supplementary Tables

Table 7. Metrics & Datasets - Accent Conversion(AC)

Paper	Obj. Metrics	Subj. Metrics	Dataset Norm.	Languages / Accents	Architecture	Data Dep.	Vocoder
Melechovsky *et al.* [2024a]	MCD, WER	MOS, XAB	L2-ARCTIC, CMU-ARCTIC	Arabic, Mandarin, Hindi, Korean, Spanish, Vietnamese, US English	ML-VAE	Non-parallel	HiFi-GAN
Nechaev and Kosyakov [2025]	WER, CER	NMOS, SMOS, Accentedness	CMU-ARCTIC, L2-ARCTIC, Common Voice, NPTEL2020, Afrispeech	US English, British English, Mandarin, Hindi, Korean, Vietnamese, Spanish, Arabic, Dutch, German, Australian English, Russian, French, Yoruba, Swahili, Igbo, Zulu, Tswana, Idoma, Afrikaans	FastPitch	Non-parallel	HiFi-GAN
Aryal and Gutierrez-Osuna [2014a]	-	-	Glasgow Herald corpus	US English, Spanish	GMM	Parallel	STRAIGHT
Nguyen *et al*. [2024]	-	MOS, SMOS, Accentedness	LJSpeech, LibriTTS-R, L2-ARCTIC, CommonVoice	Hindi, Mandarin, Vietnamese, Korean, Spanish, Arabic	Transformer	Zero-shot	-
Zhao *et al*. [2018a]	-	-	Librispeech, CMU-ARCTIC	US English, Hindi, Korean, Arabic	p-norm DNN	Parallel	STRAIGHT
Siriwardena *et al*. [2024]	MCD, WER	MOS, VSS	LibriSpeech, ARCTIC, L2-ARCTIC, Wisconsin X-ray microbeam	US English, Spanish, Mandarin, Korean, Arabic	BiLSTM, HuBERT	Zero-shot	HiFi-GAN
Radzikowski *et al*. [2021]	WER, CER, MSS	-	UME-ERJ, LibriSpeech	Japanese, British English, US English	CNN-RNN	Non-parallel, Parallel	-
Bai *et al*. [2025b]	WER, SECS	NAT, MUSHRA, Accentedness	L2-ARCTIC, ARCTIC, LibriTTS-R, Internal Dataset	Arabic, Mandarin, Hindi, Korean, Spanish, Vietnamese, US English	HuBERT	Non-parallel	BigVGAN
Zhong *et al*. [2025b]	AccCos, spkCos, SCSC	XAB, ABX	Common Voice, LibriTTS-R, VCTK	US English, Irish	YourTTS, HiFi-GAN	Zero-shot	HiFi-GAN
Prabhu *et al*. [2023]	WER	-	Common Voice, L2-ARCTIC	Australian English, Canadian English, British English, Scottish, US English, African, Hindi, Irish, Arabic, Korean, Mandarin, Spanish, Vietnamese	MaskCycleGAN-VC	Zero-shot	HiFi-GAN
Melechovsky *et al*. [2024b]	WER	MOS, SMOS, AMOS, NMOS	L2-ARCTIC	Arabic, Mandarin, Hindi, Korean, Spanish, Vietnamese	CVAE	Non-parallel	HiFi-GAN
Zhou *et al*. [2023a]	PER, WER	XAB	VCTK, CMU-ARCTIC, Google Internal Dataset, LibriTTS, Unisyn Lexicon	US English, Irish, British English, Scottish, Australian English	Tacotron 2	-	Parallel WaveGAN
Ding *et al*. [2022]	-	Accentedness, MOS, VSS	Librispeech, VoxCeleb1, Speech Accent Archive, L2-ARCTIC, ARCTIC	English	Tacotron 2	Zero-shot	WaveRNN
Ma *et al*. [2023]	MAE, Cosine Sim	SMOS, NMOS, AMOS	DB14, Wenet-Speech, Internal Dataset	Sichuan, Dongbei, Henan, Shanghai	CVAE	-	HiFi-GAN
Shrestha *et al*. [2025]	-	-	-	-	-	-	-
Melechovsky *et al*. [2024c]	MCD, CS, FFE, WER	MOS	LibriTTS, L2-ARCTIC, Voxceleb1, Voxceleb2	English	ML-VAE, FastSpeech 2	-	-
Felps and Gutierrez-Osuna [2010]	P.563, HTK Score, LDA	MOS, EGWA	IDEA, CMU-ARCTIC, WSJ	Mandarin, Italian, Japanese, Russian, Spanish, German, Hindi, British English, Australian English	-	Parallel	-

Table 7 – Metrics & Datasets - Accent Conversion(AC)

Paper	Obj. Metrics	Subj. Metrics	Dataset Norm.	Languages / Accents	Architecture	Data Dep.	Vocoder
Bai *et al.* [2024]	ASV, WER	MOS	LJSpeech, IndicTTS	English, Hindi	Diffusion	Non-parallel	-
Liu *et al.* [2020]	-	MOS, XAB, Accentedness	LibriSpeech, VoxCeleb1, VoxCeleb2, VCTK, TIMIT	US English, Hindi	Tacotron 2	Non-parallel	WaveRNN
Huang and Toda [2023]	WER, CER	NMOS, SMOS, Accentedness	L2-ARCTIC, ARCTIC, LJSpeech, LibriSpeech, LibriLight	US English, Mandarin	Tacotron 2	Parallel	-
Halychanskyi *et al.* [2026a]	WER, SS	MUSHRA	LJSpeech, L2-ARCTIC	US English, Hindi, Mandarin, Vietnamese, Korean, Spanish, Arabic	FACodec	Zero-shot	FACodec
Zhao *et al.* [2019]	-	QMOS, NMOS	Librispeech, L2-ARCTIC, ARCTIC	US English, Korean, Arabic	Tacotron 2	-	WaveGlow
Felps *et al.* [2012]	WS	EGWA	Glasgow Herald corpus, MOCHA, Wisconsin X-ray	US English, Spanish	Concatenative Synthesis	Parallel	STRAIGHT
Zang *et al.* [2022]	-	MOS	LibriSpeech, L2-ARCTIC, ARCTIC	US English, Arabic, Vietnamese, Mandarin, Spanish, Korean, Hindi	Tacotron 2	-	WaveGlow
Watanabe and Kameoka [2024]	-	-	VCTK	US English, British English, Scottish, Irish, Canadian English	GE2E-AC	Non-parallel	-
Li *et al.* [2020]	-	MOS	L2-ARCTIC, ARCTIC, Internal Dataset	US English, Hindi, Korean	Tacotron 2	Non-parallel	WaveRNN
Inoue *et al.* [2025]	AECS, WER, SECS	MUSHRA, Accentedness	L2-ARCTIC, CMU-ARCTIC, LibriTTS-R, VCTK	Hindi, Korean, Japanese, Mandarin, US English	GPT-4o, HuBERT	Synthetic Parallel	HiFi-GAN
Zhou *et al.* [2025]	MCD, F0 RMSE, F0 PCC, FD	NMOS, SMOS, AMOS	L2-ARCTIC	Arabic, Mandarin, Hindi, Korean, Spanish, Vietnamese	Tacotron 2	-	Parallel WaveGAN
Zhong *et al.* [2025a]	VF RMSE, Cosine Sim, WER, CER, MCD, UTMOS, F0 RMSE	XAB, ABX	VCTK, LJSpeech	US English, Scottish	XTTS	-	-
Onda *et al.* [2025]	-	-	JSUT, JVS, JKAC, JMAC, ERJ	Japanese, US English	HuBERT, Tacotron 2	-	-
Cheng *et al.* [2025]	WER	CMOS, Accentedness	AISHELL-1, LibriSpeech, JVS, L2-ARCTIC	Mandarin, US English, Japanese, Arabic, Hindi, Korean, Spanish, Vietnamese	SpeechCodeVAE	-	VITS
Chen *et al.* [2024b]	WER	MOS, Accentedness	LibriTTS-R, L2-ARCTIC	US English, Hindi	FastSpeech 2	Non-parallel	HiFi-GAN
Zhou *et al.* [2023b]	WER, PER	MUSHRA, Accentedness	L2-ARCTIC, VCTK, CMU-ARCTIC	Mandarin, Hindi, US English, British English, Canadian English	Tacotron 2	Non-parallel	Parallel WaveGAN
Huang *et al.* [2024]	CER, AES, SES, MOSNet	MOS, Accentedness	CMU-ARCTIC, L2-ARCTIC, VCTK, SAA	Arabic, Mandarin, Hindi, Korean, Spanish, Vietnamese	wav2vec 2.0, UUVC	Non-parallel	HiFi-GAN
Zhao and Gutierrez-Osuna [2019]	-	MOS	Librispeech, CMU-ARCTIC, L2-ARCTIC	US English, Hindi, Korean, Arabic	GMM	Parallel, Non-parallel	TANDEM-STRAIGHT
Yixin *et al.* [2025]	NISQA, SPK, LCSR	MOS	L1 ARCTIC, L2-ARCTIC, VCTK	US English, Hindi, African, Mandarin	kNN	Non-parallel, Zero-shot	-

Table 8. Metrics & Datasets - Accent & Voice Conversion

Paper	Obj. Metrics	Subj. Metrics	Dataset Norm.	Languages / Accents	Architecture	Data Dep.	Vocoder
Cheripally [2024]	-	-	Speech Accent Archive	US English, Australian English, Jamaican English, British English	HuBERT	Parallel, Non-parallel	HiFi-GAN
Chen *et al*. [2022]	CER, SES	NMOS, SMOS	VCTK, LibriTTS, ESD	English	HuBERT	One-shot	HiFi-GAN
Nguyen *et al*. [2022]	WER	AMOS, SMOS	CMU-ARCTIC, L2-ARCTIC	US English, Hindi	wav2vec 2.0, Tacotron	Parallel	-
Correa *et al*. [2023]	MCD, WER, CER, F0 RMSE	NMOS, SMOS	CMU-ARCTIC, L2-ARCTIC, UME-JRF	US English, Mandarin, Vietnamese, Japanese	MaskCycleGAN-VC	Non-parallel	HiFi-GAN
Aryal and Gutierrez-Osuna [2014b]	MCD	ABX	Glasgow Herald corpus	US English, Spanish	GMM	Parallel	STRAIGHT
Quamer and Gutierrez-Osuna [2024]	MCD, Cosine Sim	MOS, ABX	L2-ARCTIC, ARCTIC, Librispeech, VoxCeleb1, VoxCeleb2	US English, Vietnamese, Spanish	Tacotron 2	Non-parallel	HiFi-GAN
Ueda *et al*. [2024]	UTMOS, CER, SECS	MOS	LJSpeech, Blizzard2021	Brazilian Portuguese, English, Spanish	FastPitch	Low-resource	Parallel WaveGAN
Churchwell *et al*. [2024]	WER	MUSHRA	Common Voice, CMU-ARCTIC, TIMIT, VCTK	English	VITS	-	VITS
Nechaev and Kosyakov [2024]	WER, CER	MOS, SMOS, QMOS, AMOS	CMU-ARCTIC, L2-ARCTIC, Speech Accent Archive, Common Voice, VoxCeleb1, VoxCeleb2, LibriSpeech, NPTEL2020, VCTK, GigaSpeech, LibriTTS-R, LJ Speech, AfriSpeech	US English, British English, Hindi, Mandarin, Korean, Vietnamese, Spanish, Arabic, Dutch, German, Russian, Australian English, African, Yoruba, Swahili, Igbo, Zulu	FastPitch	Non-parallel	HiFi-GAN
Qin *et al*. [2024]	-	-	MSML	US English, British English, Mandarin, Japanese	VITS	Zero-Shot	HiFi-GAN

Table 9. Metrics & Datasets - Voice Conversion(VC)

Paper	Obj. Metrics	Subj. Metrics	Dataset Norm.	Languages / Accents	Architecture	Data Dep.	Vocoder
Aggarwal *et al*. [2025]	-	MOS, NMOS	CMU-ARCTIC, L2-ARCTIC, IRMAS	US English, Hindi, Korean, Mandarin, Spanish, Arabic	Encoder-Decoder	Non-parallel	-
Sisman *et al*. [2020]	MCD, F0 RMSE, WER	MOS, NMOS, QMOS, ABX, XAB	-	-	GMM, CycleGAN, StarGAN	-	WaveNet
Das *et al*. [2022]	MCD, F0 RMSE, WER	MOS, SMOS	SIWIS	German, French, English, Italian	VQ-VAE	Non-parallel	WaveRNN
Kameoka *et al*. [2020]	MCD, LFC, LDR	QMOS, SMOS, MOS	CMU-ARCTIC	US English	ConvS2S-VC	Parallel	WORLD
Nakatani *et al*. [2020]	MCD	MOS	Internal Dataset	Canadian English, British English, US English, Spanish, French, German	RNN Parallel, CycleVAE	Non-parallel	WaveNet
Yook *et al*. [2024]	Cosine Sim, MCD, MOSNet	SMOS	VCTK	English	CycleDiffusion	Non-parallel	HiFi-GAN
Takashima *et al*. [2012]	Mel-CD	ABX, XAB	ATR Japanese, CENSREC-1-C	Japanese	NMF	Parallel	STRAIGHT
Joglekar *et al*. [2025]	SSim, UTMOS	NMOS, SMOS	Librispeech, Vox Populi, NPTEL2020, Vaani, Common Voice, Shrutilipi, IIIT-H, IITM	US English, British English, Bengali, Hindi, Tamil, Telugu, Kannada, German, Spanish	CFM	Zero-shot	BigVGAN
Bai *et al*. [2025a]	F0 RMSE, NISQA	NMOS, SMOS	Opensinger, M4Singer, NHSS-Song, LibriTTS-P, NHSS-Speech	Mandarin, English	FACodec, DDSP	Zero-shot	NSF-HiFiGAN

Table 9 – Metrics & Datasets - Voice Conversion(VC)

Paper	Obj. Metrics	Subj. Metrics	Dataset Norm.	Languages / Accents	Architecture	Data Dep.	Vocoder
Karapiperis *et al*. [2024]	MCD, F0 RMSE, F0 PCC, WER, Cosine Sim	NMOS, SMOS	Internal Dataset	US English, British English, Canadian English	RVQ-VAE	-	HiFi-GAN
Kameoka *et al*. [2025]	CER, SECS, RTF	QMOS, SMOS	VCTK, LibriLight, LibriSpeech, LibriTTS	English	VoiceGrad, LatentVoiceGrad	Non-parallel, Zero-shot	HiFi-GAN
Takashima *et al*. [2013]	MCD	ABX	ATR Japanese, CENSREC-1-C	Japanese	NMF	Parallel	STRAIGHT
Luo and Dixon [2024]	MCD, WER, Cosine Sim	NMOS, SMOS	VCTK, VCC2020	British English	DSAE	Zero-shot	MelGAN
Kalita and Patra [2023]	MCD, F0 RMSE, SECS, CER, WER	NMOS, SMOS, PMOS	VCC 2018, LibriSpeech, VCTK	English	LSTM	Non-parallel	WaveNet
Yamashita *et al*. [2025]	CER, MCD	MOS	CMU-ARCTIC, L2-ARCTIC, Hi-Fi-CAPTAIN, LibriSpeech, CSJ	US English, Spanish, Vietnamese, Japanese	WSAS-VC	Parallel	MS-FC HiFi-GAN
Qinyu *et al*. [2017]	MCD	QMOS, ABX	CMU-ARCTIC	US English	GKNMF	Parallel	STRAIGHT
Takeuchi and Tatekura [2018]	-	MOS	JNAS	Japanese	GMM	-	WORLD
Kobayashi and Toda [2018]	MCD	MOS	VCC 2016, VCC 2018	English	GMM	Parallel, Non-parallel	WORLD
Wang *et al*. [2025]	MCD, F0 RMSE	NMOS, SMOS	LibriSpeech	English	WavLM	-	HiFi-GAN
Lumban Tobing *et al*. [2019]	MCD, Log-GV Distortion	NMOS, SMOS	VCC 2018, CMU-ARCTIC	English	CycleRNN	Parallel	WaveNet
Kameoka *et al*. [2024]	MCD, CER, LFC	PMOS	CMU-ARCTIC, LibriLight, LibriSpeech	US English, Hindi	VoiceGrad, U-Net	Non-parallel	HiFi-GAN
Wu *et al*. [2020]	MCD, WER, SECS	NMOS, SMOS	VCTK	English	U-Net	One-shot	MelGAN

Table 10. Metrics & Datasets - Architectures

Paper	Obj. Metrics	Subj. Metrics	Dataset Norm.	Languages / Accents	Architecture	Data Dep.	Vocoder
Ren *et al*. [2022]	MAE	MOS, CMOS	LJSpeech	English	FastSpeech 2	Non-parallel	Parallel WaveGAN
Kim *et al*. [2020]	-	-	LJSpeech, LibriTTS	US English	Glow-TTS	Non-parallel	WaveGlow
Valin and Skoglund [2019]	-	-	NTT	-	LPCNet	-	-
Jiang *et al*. [2023]	WER, SECS	QMOS, POS, SMOS	GigaSpeech, WenetSpeech, VCTK, LibriSpeech, AISHELL-3	English, Mandarin	Mega-TTS, P-LLM	Zero-shot	HiFi-GAMN
Min *et al*. [2021]	MCD, WER, CER	MOS, PMOS, QMOS, SMOS	LibriTTS, VCTK	US English, British English, Hindi, African, Australian English	StyleSpeech	-	MelGAN
Shen *et al*. [2018]	-	-	Internal US English Dataset	US English	Tacotron 2	-	WaveNet
Shen *et al*. [2021]	-	-	LibriTTS	US English, British English, Australian English, Nigerian	Tacotron	-	WaveRNN
Badlani *et al*. [2021]	-	-	LJ Speech, LibriTTS	English	RAD-TTS	-	WaveGlow, HiFi-GAN
van den Oord *et al*. [2017]	-	-	Internal Dataset	English, Japanese	IAF	Non-parallel	Parallel WaveNet
Zeghidour *et al*. [2021]	-	-	LibriTTS, Freesound, MagnaTagATune, VCTK	English	SoundStream	-	-

Table 10 – Metrics & Datasets - Architectures

Paper	Obj. Metrics	Subj. Metrics	Dataset Norm.	Languages / Accents	Architecture	Data Dep.	Vocoder
Li *et al.* [2023]	MCD, WER, F0 RMSE	CMOS, NMOS, SMOS	LJSpeech, VCTK, LibriTTS	English	StyleTTS 2, WavLM	Zero-shot	HiFi-GAN, iSTFTNet
Chen *et al.* [2024a]	WER, SSIM	SMOS, CMOS, DNSMOS	Libriheavy, LibriSpeech, VCTK	English	Transformer	Zero-shot	Vocos
Han *et al.* [2024]	MCD, WER	QMOS, SMOS, CMOS	LibriSpeech	English	Transformer	Zero-shot	Vocos
van den Oord *et al.* [2016]	-	-	VCTK, Google Internal Dataset, TIMIT, MagnaTagATune	US English, Mandarin	WaveNet	Parallel	-
Casanova *et al.* [2024]	-	-	LibriTTS-R, LibriLight, Common Voice, FLORES+, DAPS	English, Spanish, French, German, Italian, Brazilian Portuguese, Polish, Turkish, Russian, Dutch, Czech, Arabic, Mandarin, Hungarian, Korean, Japanese	VQ-VAE, GPT-2	Zero-shot	HiFi-GAN
Casanova *et al.* [2023]	SECS	MOS, SMOS	VCTK, LibriTTS, TTS-Portuguese Corpus, M-AILABS, MLS, Common Voice	English, Brazilian Portuguese, French	VITS	Zero-shot	HiFi-GAN