©SHUTTERSTOCK.COM/METAMORWORKS

Chenxi Liu and Israel Cohen

Foreign Accent Conversion

Methods, models, and research challenges

This article aims to provide a structured and technically grounded overview of foreign accent conversion (FAC). Our goal is not only to summarize existing methods but also to clarify how FAC systems are designed, evaluated, and compared. To this end, we make three main contributions. First, we provide a clear, systematic classification of FAC methods into data- and model-based approaches, offering a structured view of the field and its development. Second, we present an in-depth review of model-based FAC pipelines, highlighting their core components, recent advances, and the main differences among major model types. Third, we complement this qualitative review with a meta-analysis of recent FAC studies. This meta-analysis plays an important role because FAC articles often use different datasets, model settings, and evaluation protocols, which makes direct comparison difficult. By quantitatively analyzing commonly reported evaluation results, we identify broader performance trends, key tradeoffs, and the main gaps that still limit practical deployment.

Introduction

FAC is a speech-to-speech transformation task that modifies the accent-related characteristics of an input speech while preserving its linguistic content and speaker identity. From a signal processing perspective, the input is a speech waveform produced with a source accent, and the output is a transformed waveform representing the same utterance from the same speaker in a target accent. The objective is therefore to alter accent-related acoustic cues, such as pronunciation, rhythm, stress, prosody, and intonation, without changing the speech content or the speaker's voice identity. This capability has strong potential for cross-cultural communication, language education, and speech-related assistive applications.

FAC is related to several neighboring speech tasks, but it fundamentally differs from each. Unlike voice conversion (VC) [1], [2], which mainly changes speaker-dependent vocal characteristics such as timbre, FAC keeps the speaker identity fixed and instead modifies only accent-related cues such as pronunciation and prosody. FAC also differs from accented text-to-speech (TTS), which generates accented speech from text and therefore assumes text input and paired text-speech supervision. By contrast, FAC operates directly on speech input and performs accent transformation without generating speech from text. Compared with speech translation, FAC changes the accent within the same language. Even when FAC systems use components from automatic speech recognition (ASR), TTS, or VC [2], [3], FAC remains a constrained speech-to-speech transformation task. Its main goal is to produce natural target-accented speech while preserving intelligibility and the source speaker's identity. A more detailed comparison between FAC

Dr. Konstantinos Slavakis acted as the associate editor for this article and recommended it for publication.

Digital Object Identifier 10.1109/MSP.2026.3703111
Date of current version: 2 October 2026

1053-5888 © 2026 IEEE. All rights reserved, including rights for text and data mining, and training of artificial intelligence and similar technologies.

Authorized licensed use limited to: EssilorLuxottica SA. Downloaded on October 06,2026 at 08:48:14 UTC from IEEE Xplore. Restrictions apply.

and neighboring tasks, including VC and accented TTS, is provided in the "Positioning of FAC in Speech Synthesis" section.

Currently, FAC faces several fundamental challenges. First, parallel training data are scarce. An ideal FAC dataset would contain the same utterances spoken by the same speaker with different accents. In practice, such datasets are rare because speakers usually only speak in their native accent [4], [5]. Second, accent is shaped by both phonetic and prosodic factors, including segmental features such as phoneme articulation and suprasegmental features such as rhythm and intonation [6]. These accent-related features are also inherently entangled with speaker identity, making it challenging to change an accent without altering the speaker's original voice. Third, although companies such as Krisp Technologies [7] and Sanas [8] have begun commercializing real-time accent-conversion products, FAC remains an active and evolving research area. Most academic systems are still trained and evaluated under controlled conditions, such as short read speech, limited accent pairs, and offline conversion. Therefore, bridging the gap between research prototypes and robust, scalable systems for everyday communication remains an important open challenge.

Despite these challenges, FAC has significant potential in several real-world application domains. In language learning, it can support pronunciation correction and accent adaptation, helping second-language learners move toward more native-like speech [4], [9]. In speech technology, FAC can improve ASR robustness through accent-diverse data augmentation [10], [11]. FAC can also improve communication quality in customer service settings, such as international call centers, where accent differences can affect both intelligibility and user experience. Similarly, FAC can improve communication in online meetings, such as Zoom, by reducing accent-related barriers during real-time speech interaction. In multimedia applications, FAC can support movie dubbing [9], [11], [12], multiaccent audiobooks, and other cross-regional speech services [4]. It may also serve as an assistive tool for more effective human–computer interaction [12]. Together, these applications highlight the potential of FAC to improve both speech technology and real-world communication.

This article provides a unified technical overview of FAC, categorizing existing methodologies by their data requirements and architectural designs to trace the field's evolution from traditional frameworks to contemporary deep learning-based systems. We examine representative model-based FAC pipelines, focusing on how they disentangle linguistic content, preserve speaker identity, and precisely modify accent-related features. Furthermore, to address the lack of standardization stemming from disparate datasets, accent pairs, and listening-test protocols across the literature, we conduct a comprehensive meta-analysis of commonly reported evaluation metrics. By synthesizing qualitative architectural reviews with quantitative evaluation trends, this article highlights critical performance trade-offs and outlines open challenges to guide future research in FAC.

In the remainder of this article, we first present an overview of how FAC has developed over time in the "Background and Development" section, where we differentiate between traditional and deep learning (DL)-based methods. We then discuss traditional FAC methods in the "Traditional FAC Methods" section. Next, we review the three main categories of DL-based methods: sequence-to-sequence (seq2seq) FAC, TTS-guided FAC, and discrete speech token-based FAC in the "Seq2seq FAC," "TTS-Guided FAC," and "Discrete Speech Token-Based FAC" sections, respectively. This is followed by the "Feature Disentanglement Strategies in FAC" section, which discusses the methods used for separating accent-related features from other speech factors, such as speaker identity and linguistic content. We then present an overview of the datasets and evaluation metrics that are used in FAC. Based on commonly used evaluation metrics, we conduct a meta-analysis of current FAC research in the "Meta-Analysis of Current FAC Research" section to compare the performance of different FAC studies. Finally, the "Discussion" section summarizes current limitations and challenges as well as promising future research directions for FAC.

Background and development

In this section, we provide the essential background for FAC. To help readers gain a better overview of the FAC area, we describe its development over the years with a detailed timeline and its positioning in speech synthesis, including how it differs from neighboring fields such as VC and accented TTS. Finally, we discuss how FAC can be categorized based on its data requirements.

The emerging nature of FAC

Although FAC holds great potential to address diverse real-world problems, it remains a relatively new field, as shown in Figure 1. Traditional FAC methods first emerged in the early 2010s, and more advanced DL-based approaches only emerged in the late 2010s and early 2020s. The literature is sparse, with many related studies focusing on voice cloning [1], [13] or TTS development [14] rather than accent conversion. To date, only two overview articles exist: a brief conference overview by Shrestha et al. [15] and a recent survey of statistical and DL techniques [16]. However, both provide

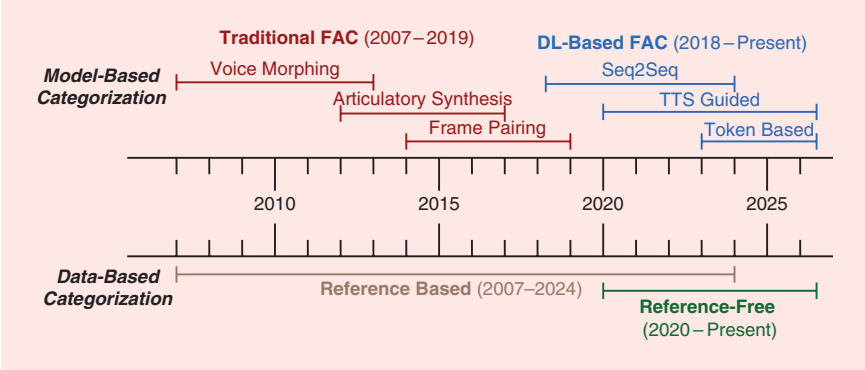

FIGURE 1. Timeline of FAC development and categorization of approaches.

Authorized licensed use limited to: EssilorLuxottica SA. Downloaded on October 06,2026 at 08:48:14 UTC from IEEE Xplore. Restrictions apply.

only limited coverage of FAC methodologies and evaluation practices. Therefore, certain critical questions remain unanswered, such as 1) What are the state-of-the-art methods in FAC? 2) What are the primary limitations of current FAC approaches? 3) What are the key challenges and research gaps? 4) What are the promising future directions for FAC? The lack of comprehensive reviews or meta-analyses highlights the need for systematic studies to clarify the current state of the FAC field and guide its future advancement.

Positioning of FAC in speech synthesis

FAC is a specialized task within the broader field of speech synthesis, which generally involves processes that either generate speech from text or modify the properties of existing speech. FAC is uniquely situated between VC and accented TTS, sharing methodologies with both while maintaining distinct modeling objectives and data assumptions. Table 1 summarizes the key distinctions among FAC, VC, and accented TTS along the dimensions of input modality, conversion target, preservation constraint, and data requirements.

Relation to VC

Although FAC is often considered a subcategory or a special case of VC, the two fields differ fundamentally in their modeling assumptions and objectives: the VC system typically assumes that speech can be factorized into linguistic content and speaker identity. The main goal of a typical VC system is to modify a source speaker's voice to sound as if it were produced by a target speaker, effectively changing the speaker's identity while keeping the linguistic content unchanged [17]. Therefore, most VC systems focus on spectral conversion, which seeks to change speaker timbre and global prosody shaping at the speaker level, typically not distinguishing "accent" from other aspects of speaking style. By contrast, FAC adopts a different factorization: speaker identity and linguistic content are treated as invariants, while "accent" is modeled as a third factor to be modified independently. The core goal of FAC systems is to preserve the source speaker's voice identity and underlying linguistic content while selectively modifying only accent-specific features so that the converted speech sounds as if the same speaker had produced it with a different accent. Specifically, FAC requires finer-grained control of segmental and suprasegmental patterns—such as phoneme realization, prosody, rhythm, stress placement, intonation contour, and duration—to match a target accent. As a result, the two tasks carry opposing constraints on speaker characteristics. Although VC is explicitly designed to alter speaker-dependent features and requires speaker-to-speaker feature mapping, FAC requires accent-to-accent mapping under a speaker-identity constraint.

Consequently, the two tasks also differ in their data requirements: VC typically employs multi-speaker corpora organized by speaker identity (parallel or nonparallel), whereas FAC requires speech data organized by accent category, ideally with matched speaker content across source and target accents to enable learning of accent-specific transformations independently of speaker variation.

Relation to accented TTS

Accented TTS and FAC both model accent-specific phonetic and prosodic features—including phoneme realization, pitch contour, and phoneme duration—but differ fundamentally in input modality and modeling objective. Accented TTS is a *text-driven* generation task that operates in the TTS domain: it synthesizes target accented speech directly from text input, generating both the acoustic signal and its accent style from scratch [18], [19]. By contrast, FAC is a *speech-driven* conversion task that operates entirely in the speech-to-speech domain: it takes an existing second-language (L2) speech utterance as input. It modifies its accent-specific features toward a target accent, without access to the text transcript. This distinction in input modality leads to opposing constraints on speaker identity. In accented TTS, speaker identity is a controllable conditioning factor, and the system can generate new acoustic realizations of the input text. In FAC, speaker identity is treated as an invariant that must be preserved throughout the conversion, making speaker-accent disentanglement a central technical challenge rather than a design choice. Consequently, the two tasks also differ substantially in their data requirements: Accented TTS requires paired text-speech data with accent-specific phonetic lexicons for supervision. FAC requires accent-parallel or accent-labeled speech corpora organized by accent category with matched speaker content across accents. In essence, accented TTS treats accent as a *synthesis target*, whereas FAC treats it as a *transformation target* under a speaker-preservation constraint.

Data-based categorization

After reviewing the literature, we categorize FAC approaches based on their data requirements during training and inference (Table 2). From a data-based perspective, FAC methods fall into two main categories: 1) *reference-based* methods, which require

Table 1. Key distinctions of FAC, VC, and accented TTS.

Property	FAC	VC	Accented TTS
Input	Accented speech	Source speaker speech	Text
Output	Target-accent speech	Target speaker speech	Accented speech
Conversion target	Accent features	Speaker identity	N/A (Synthesis)
Preserved	Linguistic content + speaker identity	Linguistic content	N/A
Data requirement	Accent-parallel/labeled speech	Speaker-paired speech	Accented text–speech pairs

N/A: not applicable.

Authorized licensed use limited to: EssilorLuxottica SA. Downloaded on October 06,2026 at 08:48:14 UTC from IEEE Xplore. Restrictions apply.

parallel data (i.e., recordings from both native first language (L1) and nonnative L2 speakers with the same linguistic content) for both training and inference, and 2) *reference-free* methods, which do not require parallel data during inference. Reference-free methods can be further divided into 1) *autoregressive* approaches, which require parallel data only during training, and 2) *non-autoregressive* approaches, which eliminate the need for parallel data during both training and inference. In subsequent sections, we also introduce FAC methods by their underlying model architectures. This data-based view complements the model-based categorization, offering a clearer picture of how FAC techniques have evolved over the years.

Traditional FAC methods

This section examines early-stage FAC approaches developed before and during the early 2010s. Traditional FAC methods can be categorized into three main categories: voice morphing, articulatory synthesis, and frame-pairing. Each represents a different methodology for achieving accent conversion.

Voice morphing

Key ideas

One of the earliest approaches to FAC was voice morphing [20], [21]. The basic idea is to decompose speech into two spectral components: *spectral slope* and *spectral detail*. The spectral slope represents the broad spectral structure of the speech signal and is closely related to speaker identity, such as timbre. By contrast, the spectral detail captures the finer acoustic variations associated with pronunciation characteristics. Accent conversion is then performed by combining the L2 speaker's spectral slope with a morphed version of a native L1 speaker's spectral detail. In this way, the converted speech is intended to retain the original speaker's voice while incorporating more native-like pronunciation cues.

Main steps

It first decomposes the "teacher" L1 and "learner" L2 speech into spectral slope and spectral detail components through cepstral decomposition and keeps the L2 learner's spectral slope. Then, both speakers' spectral details are converted into pulses (i.e., dot-like representations of peaks and valleys) through pulse density modulation to encode the sparse pulse patterns. These pulses from both speakers are aligned and averaged pairwise. Next, the spectral details are blended by a morphing factor α, with the blended result being $(1-\alpha)\times$ Learner Pulses $+\,\alpha\times$ Teacher Pulses. Increasing α will gradually incorporate more native details and create a smooth accent transformation. Finally, the morphed details are merged back with the L2 learner's spectral slope, and resynthesize the audio using a vocoder.

Articulatory synthesis

Key ideas

Articulatory synthesis methods for FAC operate in the articulatory domain, leveraging measurements of vocal tract movements—such as positions of the tongue, lips, and jaw—typically captured via electromagnetic articulography (EMA). A typical articulatory-based method consists of building an articulatory synthesizer for the L2 speaker that maps contextualized articulatory gestures to acoustic sounds, driving the model with L1 articulatory gestures normalized to the L2 articulatory space. This results in speech that captures the linguistic gestures (e.g., prosody and segmental cues) of the L1 speaker and the L2 speaker's voice quality. Several techniques can be used to build the articulatory synthesizer, including unit-selection synthesis [22], statistical parametric synthesis [23], and deep neural network (DNN) synthesis [24], [25].

Main steps

Articulatory synthesis-based FAC first collects paired EMA articulator positions and acoustic recordings, aligned with phonetic transcriptions. Because speakers have different vocal tract geometries, the L1 articulatory trajectories are normalized into the L2 speaker's articulatory space before conversion. An articulatory-to-acoustic model is then trained to map articulatory gestures to acoustic features, using unit selection, statistical parametric, or DNN-based synthesis, depending on the implementation. Subsequently, the trained model is applied to the normalized L1 gestures to generate more native-like speech. Then, prosodic features such as L1 pitch contours are adapted to the L2 speaker's pitch range to maintain natural intonation. Finally, a vocoder reconstructs the accent-converted speech waveform from the mapped acoustic features.

Frame-pairing

Key ideas

Frame-pairing methods for FAC involve segmenting both the source L2 speech and the target L1 speech into short frames. These frames are then paired based on predefined similarity metrics, such as dynamic time warping (DTW) alignment, acoustic similarity, or phonetic similarity, to build a frame-pair lookup table. This table serves as the basis for training a spectral conversion model that maps the spectral features of the nonnative source speaker to those of the native target speaker, effectively adapting the pronunciation while preserving the source speaker's vocal identity. In contrast to traditional VC techniques, which rely on temporal alignment and transfer both the target speaker's voice characteristics and accent, frame-pairing approaches selectively match frames to incorporate only the desired native pronunciation elements. This results in output speech that sounds more native-like without altering the speaker's identity.

Table 2. Parallel data requirements for FAC approaches.

Parallel Data Requirement		Training	Inference
Reference based		✓	✓
Reference-free	Autoregressive	✓	×
	Non-autoregressive	×	×

✓: parallel data are required; ×: no parallel data are needed.

Authorized licensed use limited to: EssilorLuxottica SA. Downloaded on October 06,2026 at 08:48:14 UTC from IEEE Xplore. Restrictions apply.

Main steps

In frame-pairing-based FAC, L1 and L2 utterances are first segmented into short frames and represented as acoustic features, such as mel frequency cepstral coefficients (MFCCs), mel cepstral coefficients (MCEPs), or phonetic posteriorgrams (PPGs). Then, corresponding L1–L2 frame pairs are established using different matching strategies: 1) DTW-based pairing aligns frames according to the temporal structure of parallel utterances. 2) Acoustic similarity-based pairing matches source and target frames based on their acoustic similarity using MFCCs after vocal-tract-length normalization. 3) Phonetic similarity-based pairing uses PPGs and symmetric Kullback–Leibler (KL) divergence to identify frames with similar phonetic content. These paired frames are used to train a Gaussian mixture model [26] that maps the joint distribution of source and target MCEPs. During conversion, maximum-likelihood parameter generation with global variance [27] is applied to generate smooth converted spectral features. Then, the native L1 pitch trajectory is normalized to match the L2 pitch range to preserve the L2 speaker's natural vocal range. Finally, a vocoder is used to reconstruct the accent-converted waveform.

Limitations of traditional FAC methods

Traditional FAC methods have several limitations that restrict their practical use. Voice morphing requires well-aligned parallel recordings from L1 and L2 speakers. It can also reduce speech naturalness when the accent change is large because it modifies spectral details in a relatively direct way. In addition, it uses a simple separation between speaker identity and accent-related information, which cannot fully capture important accent cues such as pitch, rhythm, stress, and intonation.

Articulatory synthesis models accent in a more physically meaningful way by using vocal tract movements, such as tongue, lip, and jaw positions. However, they depend on EMA, which is expensive, time consuming, and difficult to collect at scale. As a result, existing articulatory FAC systems are usually trained on small datasets and a limited set of accent pairs, making them difficult to generalize to many speakers and accents. Recent work showing that self-supervised speech models encode articulatory information suggests a possible path toward more practical articulatory-informed FAC systems without direct EMA measurements [28].

Frame-pairing methods are more data driven but still require parallel or closely matched L1–L2 speech. Their performance also depends heavily on the quality of frame alignment. As conversion is performed mainly frame by frame, these methods have difficulty modeling longer-range accent patterns, such as rhythm, stress placement, and intonation across a full utterance. They also tend to produce lower-quality speech than modern neural approaches.

These limitations motivated the shift toward DL-based FAC methods, which can learn richer speech representations, handle nonparallel data more flexibly, and generate more natural converted speech.

Seq2seq FAC

Seq2seq models

Seq2seq models were first introduced by Sutskever et al. [29] for neural machine translation. A seq2seq model usually has an encoder–decoder architecture. The encoder derives a hidden representation (often a fixed-dimensional vector or a sequence of states) from the input sequence, and the decoder learns to generate the output sequence based on this representation. To capture contextual information and handle length mismatches between input and output sequences, Bahdanau et al. [30] introduced an attention mechanism, commonly integrated between the encoder and the decoder, allowing the decoder to dynamically focus on relevant parts of the input.

In recent years, seq2seq models have gained significant attention for speech-related tasks, extending beyond their origins in machine translation to domains such as TTS synthesis [31], [32], VC [33], [34], and FAC [35], [36], [37], [38]. The use of sequence mapping and contextual modeling in seq2seq models significantly improves performance on these speech-related tasks compared to conventional approaches. Therefore, in this section, we focus on seq2seq-based FAC approaches. We first introduce two representative types of seq2seq FAC models based on their data requirement: reference based and the reference-free. Then, we present a general pipeline that includes the core steps of most seq2seq FAC systems, providing a foundational framework for understanding current research and advancing these models.

Reference-based versus reference-free seq2seq models

Seq2seq FAC models can be broadly divided into two families based on their data requirements during inference. Reference-based models [37], [38] require parallel recordings—paired audio of the same content spoken in both the L2 and L1 accents—during both training and inference. At inference time, these systems rely on explicit accent embeddings extracted from a native reference utterance to guide the conversion. The conversion model then combines three types of input: linguistic content, the speaker identity extracted from the L2 source utterance, and the target accent embedding extracted from the L1 reference. One representative example of this type is Accentron [38]. The key innovation of Accentron is that it enables zero-shot conversion for unseen L2 speakers during training by conditioning a seq2seq model on separately extracted speaker and accent embeddings, rather than training a dedicated model for each L1–L2 speaker pair.

Reference-free models [35], [36] remove the requirement for parallel data at inference time by internalizing the accent transformation within the trained model's learned parameters. A representative approach by Quamer et al. [35] decomposes the conversion into two sequential stages: a translator that maps L2 bottleneck features (BNFs) to their L1 counterparts (trained on parallel data), followed by a synthesizer that reconstructs speech from the converted features conditioning on the source speaker's identity. This design makes inference more practical because only the L2 utterance is needed, while the target-accent characteristic is

Authorized licensed use limited to: EssilorLuxottica SA. Downloaded on October 06,2026 at 08:48:14 UTC from IEEE Xplore. Restrictions apply.

internalized in the trained model. Moreover, it also allows zero-shot conversion of unseen speakers without requiring a native reference at test time.

General seq2seq FAC pipelines

Despite differences among individual systems, most seq2seq FAC models follow the same high-level four-stage pipeline, as shown in Figure 2 and discussed next.

Step 1: Feature extraction and disentanglement

The pipeline first decomposes the input speech into three complementary representations: linguistic content, speaker identity, and accent characteristics so that only the accent-related features can be modified in later stages. Linguistic content represents "what is said" and is typically extracted by a speaker-independent acoustic model as BNFs or PPGs. This is analogous to extracting the musical notes of a melody, independent of the instrument or performance style. Speaker identity represents "who is speaking" and is captured by a speaker encoder as a speaker embedding that preserves vocal traits such as timbre, resonance, and pitch range. This is like the instrument itself, such as a guitar, violin, or flute. Accent characteristics represent "how it is said" and capture pronunciation patterns, rhythm, stress, and intonation. This is similar to the musician's playing style: the same notes can sound different depending on timing, emphasis, phrasing, and melodic contour, just as the same words can sound different across accents.

Step 2: Accent transformation

This stage performs the core accent conversion. The system modifies the accent-related representation while keeping the linguistic content and speaker identity largely unchanged. In reference-based models, the target accent embedding is extracted explicitly from a native-accent reference utterance and used to guide the model toward the desired accent during inference. Therefore, reference-based models require parallel data during both training and inference. By contrast, reference-free models learn the accent transformation during training and apply this learned mapping directly to the source speech at inference. Therefore, they only require parallel data during training.

Step 3: Acoustic synthesis

After accent transformation, the seq2seq model combines the linguistic representation with the source speaker identity and the target-accent embeddings to generate the target mel spectrogram that represents the converted speech. In reference-based systems, the target-accent information may appear as an explicit accent embedding; in reference-free systems, it may already be embedded in the converted BNFs, PPGs, or other intermediate representations.

Step 4: Waveform reconstruction

Finally, a neural vocoder, such as WaveRNN or HiFi-GAN, converts the generated mel spectrogram into the output speech waveform.

TTS-guided FAC

Building upon the foundational seq2seq FAC approaches, which primarily use encoder–decoder architectures to transform accented speech directly into target accents without explicit intermediate representations. Recent research has shifted toward leveraging the powerful representation-learning

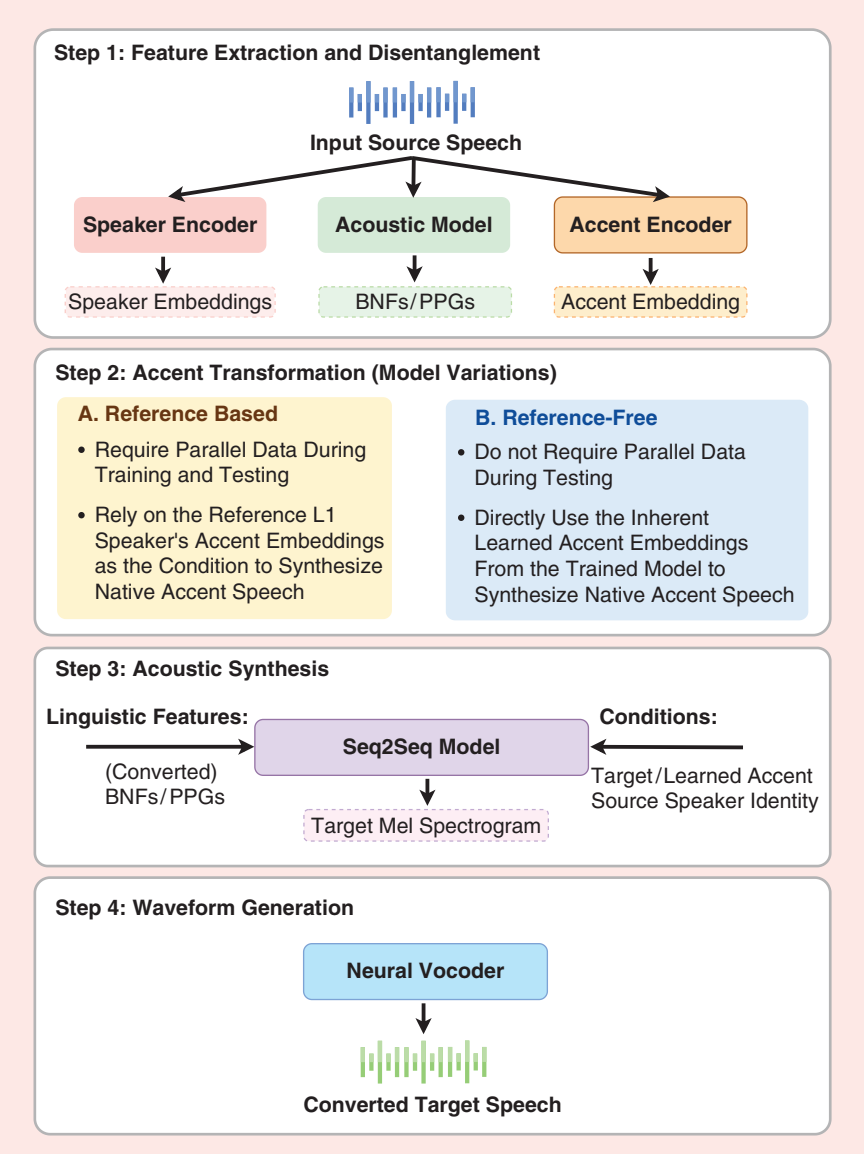

FIGURE 2. General seq2seq FAC pipeline.

Authorized licensed use limited to: EssilorLuxottica SA. Downloaded on October 06,2026 at 08:48:14 UTC from IEEE Xplore. Restrictions apply.

capabilities of TTS systems and integrating TTS models to guide accent conversion. The next section explores two main branches of TTS-guided FAC approaches: TTS-guided end-to-end (E2E) models that use pretrained TTS systems as teachers, and diffusion-based approaches that use generative modeling advances.

E2E TTS-guided FAC

A typical E2E TTS-guided FAC model includes the following key components, and its general pipeline is illustrated in Figure 3.

- *Pretrained TTS*: The TTS system consists of a text encoder and decoder. The text encoder takes text inputs (e.g., transcriptions or phoneme sequences) to generate text embeddings. The decoder generates acoustic features from the text representations, guided by the target native-accented speech by using the mean-square error loss L_{TTS}. Therefore, it is expected that the output speech from the TTS model represents pronunciation patterns of the target accent and is free from other accents. The training process is shown in the "Train 1" stage in Figure 3.

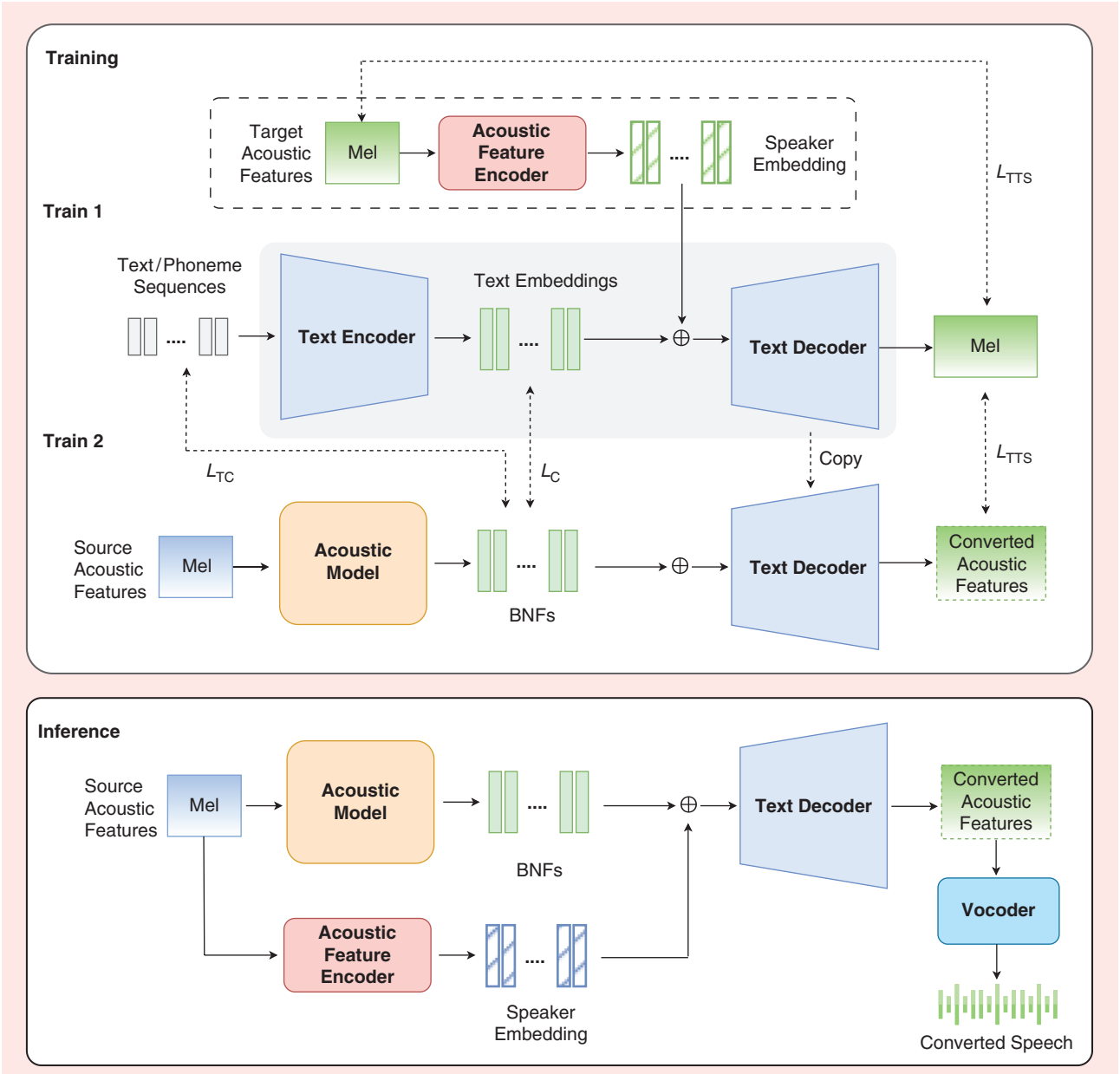

FIGURE 3. A general TTS-guided FAC architecture, based on [3], [9], and [39]. Training: a two-stage process where 1) the TTS model is pretrained on native speech with optional acoustic conditioning (dashed box) in Train 1, and 2) the acoustic model learns accent-agnostic linguistic representations guided by the pretrained TTS through multiple losses (dashed arrows) in Train 2. Inference: Source accented speech is processed by the acoustic model to extract BNFs, which are then concatenated (⊕) with optional acoustic features (e.g., speaker embeddings), and decoded to generate accent-converted speech through the TTS decoder and neural vocoder.

Authorized licensed use limited to: EssilorLuxottica SA. Downloaded on October 06,2026 at 08:48:14 UTC from IEEE Xplore. Restrictions apply.

- *Acoustic feature encoder*: The acoustic feature encoder extracts conditioning features from the target speech. For example, Liu et al. [3] use a speaker encoder to extract fixed-dimensional speaker embeddings from target speech, enabling the TTS decoder to synthesize speech with the target speaker's identity. However, this component is optional, and its specific implementation varies based on which acoustic characteristics (e.g., speaker identity, accent, and so on) the system aims to capture.
  - *Acoustic model*: To enable text-free inference and only use the accented speech as input, the system uses an acoustic model (e.g., accented ASR model [3] or speech encoder [39]) to extract linguistic representation directly from the source speech. The training leverages the pretrained TTS module as a teacher, using its outputs (text embeddings and predicted acoustic features) to guide the acoustic model's learning. Formally, let $\mathbf{h}_n^s$ and $\mathbf{h}_n^t$ denote the speech and text embeddings, respectively, at phoneme position $n \in \{1, \ldots, N\}$; c_n the ground-truth phoneme label; $\hat{Y}_t$ and Y_t the predicted and target mel spectrogram frames, respectively, at time step $t \in \{1, \ldots, T\}$. The training objective is normally to minimize the combination of multiple complementary losses: $L_{\text{total}} = \lambda_{\text{TTS}} L_{\text{TTS}} + \lambda_{\text{TC}} L_{\text{TC}} + \lambda_{\text{C}} L_{\text{C}}$, where the λ terms are empirical weights tailored to system requirements, specifically:
  - *Reconstruction loss* (L_{TTS}): The mean squared error between the predicted $(\hat{Y})$ and target (Y) acoustic features, computed as $L_{\text{TTS}} = (1/T)\Sigma_{t=1}^{T} \left\| Y_t - \hat{Y}_t \right\|_2^2$.
  - *Text classification loss* (L_{TC}): Ensures phonemic correctness to strip source accents [40], computed via cross-entropy $L_{\text{TC}} = -(1/N)\Sigma_{n=1}^{N} \log P(c_n | \mathbf{h}_n^s)$, with P as the posterior from a linear phoneme classifier.
  - *Contrastive loss* (L_C): Aligns the speech and text embedding spaces [39], [40], using a margin hyperparameter $\tau > 0$: $L_{\text{C}} = (1/N)\Sigma_{n=1}^{N} [\| \mathbf{h}_n^s - \mathbf{h}_n^t \|_2^2 + \Sigma m \neq n \max(0, \tau - \| \mathbf{h}_n^s - \mathbf{h}_m^t \|_2)^2]$. In practice, this formulation can be extended to incorporate additional task-specific losses tailored to specific system requirements. Through these acoustic-linguistic and feature alignments, the acoustic model learns accent-agnostic linguistic representations from accented acoustic features, effectively "seeing through" the accent to capture the underlying linguistic content. During this training phase, only the acoustic model parameters are updated while the TTS system remains frozen (as illustrated in the "Train 2" stage in Figure 3). At inference time, the acoustic model is used to transform the accented input speech into linguistic representations, which are then forwarded to the pretrained TTS decoder to generate the converted speech mel spectrograms ("Inference" stage in Figure 3).
- *Neural vocoder*: Finally, the accent-converted speech mel spectrogram is transformed back to speech waveforms by a neural vocoder.

Diffusion-based E2E TTS-guided FAC

Recent advances in diffusion models have led to their integration into TTS-guided FAC systems, enhancing speech with higher naturalness and better quality. Bai et al. [9] and Deja et al. [41] are two representative diffusion-based E2E TTS-guided FAC approaches.

Similarity with E2E TTS-guided FAC

The diffusion-based FAC systems retain training processes that are similar to E2E TTS-guided FAC approaches. Both [9] and [41] include two training stages: TTS and accent conversion. Similar to the fundamental architecture of E2E TTS-guided FAC systems, [9] comprises a text-encoding module that converts text input into linguistic representations $\mathbf{h}^t$ and a speech-encoding module that transforms input speech features into linguistic representations $\mathbf{h}^s$. The training objective also includes acoustic feature alignment to make sure that the linguistic representations $\mathbf{h}^s$ from the speech encoder reside in the same space as the target linguistic representation $\mathbf{h}^t$ from the text-encoding module, and that $\mathbf{h}^s$ is a suitable input for the speech-decoding module.

Differences

Although traditional E2E TTS-guided FAC systems use recurrent neural network or Transformer decoders that directly predict mel spectrograms in a single step, diffusion-based FAC systems use diffusion-based speech-decoding modules to generate or refine mel spectrograms through multiple iterative denoising steps. In this process, the model learns to recover a clean target-accent speech representation from a noisy or partially degraded acoustic representation while preserving the desired linguistic, speaker, and accent-related conditions. The next two systems illustrate different ways of incorporating diffusion into TTS-guided FAC.

Bai et al. [9] use a diffusion Transformer (DiT) as the speech decoder. During training, the decoder receives a noisy mel spectrogram, a diffusion time step, and conditioning representations that encode linguistic content, speaker characteristics, and prosodic information such as pitch and energy. Rather than directly predicting the final mel spectrogram, the DiT learns a transformation that maps the noisy representation toward the clean target-accent mel spectrogram, using a flow-matching objective. During inference, the speech encoder first converts source accented speech into linguistic representations. The diffusion decoder then predicts the pitch and energy information and iteratively refines the converted mel spectrogram through ordinary differential equation (ODE)-based sampling. In this design, diffusion primarily serves as the acoustic generator, replacing the conventional single-pass TTS decoder.

Deja et al. [41] adopt a score-based denoising diffusion probabilistic model with a U-Net decoder. Their system first uses saliency maps from an externally trained accent classifier to identify regions of the source mel spectrogram that are strongly associated with the source accent. These regions are then suppressed to reduce accent-specific cues before reconstruction. After noise is added to the processed mel spectrogram, the U-Net diffusion decoder denoises it under speaker and target-accent

Authorized licensed use limited to: EssilorLuxottica SA. Downloaded on October 06,2026 at 08:48:14 UTC from IEEE Xplore. Restrictions apply.

conditioning. Thus, in this design, diffusion is used mainly as a denoising and reconstruction module that converts an accent-suppressed acoustic representation into target-accent speech.

Advantages

A general advantage of TTS-guided FAC is the utilization of a pretrained TTS decoder that has already learned a strong mapping from linguistic representations to natural target-accent speech. Therefore, FAC training focuses primarily on ensuring that the speech encoder produces TTS-compatible linguistic features. Compared with direct seq2seq FAC, this design provides a stronger acoustic prior, reduces the burden of learning pronunciation correction and speech generation jointly from limited FAC data, and usually improves training stability, pronunciation accuracy, and intelligibility [3], [39]. In diffusion-based variants, naturalness is further improved because the decoder refines the mel spectrogram through multiple denoising steps rather than predicting it in a single pass, which helps recover fine acoustic details and reduces oversmoothed outputs [9], [41].

Limitations

Despite these advantages, TTS-guided FAC still depends strongly on the quality and accent coverage of the pretrained TTS model, and mismatches between the speech encoder's output and the decoder's input space can limit conversion accuracy. In addition, the disentanglement among accent, speaker identity, and prosody remains imperfect, which may lead to residual source-accent traits or reduced speaker similarity [3], [9], [39]. For diffusion-based methods, a major limitation is slow inference because the mel spectrogram must be generated through multiple iterative denoising steps, making these systems less suitable for real-time FAC applications [9], [41].

Discrete speech token-based FAC

Building upon the foundational advancements of TTS-guided FAC, recent FAC has moved beyond frame-level acoustic feature mapping toward using discrete speech tokens as an intermediate representation [42], [43], [44], [45]. By discretizing speech into intermediate units, these models effectively decouple linguistic content from the speaker's original timbre, enabling more precise manipulation of pronunciation patterns. Moreover, the discrete, compact nature of these tokens makes them compatible with language modeling frameworks, such as large language models (LLMs) [42], [44] and diffusion/flow-matching models [43], [45]. Current research [46] categorizes speech tokens into two types: acoustic and semantic. Acoustic tokens are derived from neural codecs designed to encode speech at low bit rates while preserving rich acoustic detail. Semantic tokens arise from self-supervised learning and exhibit strong correlations with phonemes, effectively preserving linguistic content while discarding paralinguistic variations such as timbre.

Representative token-based FAC models

Several representative models were proposed in the 2024–2025 period. These frameworks, SpeechAccentLLM, TokAN, FAC-FACodec, and MacST, represent distinct yet complementary strategies for leveraging LLMs, diffusion-based and flow-matching synthesis, and data augmentation to achieve high-fidelity accent conversion.

LLM-based unified frameworks

SpeechAccentLLM [42] introduces a unified LLM-based system that treats FAC and TTS as a joint multitask learning problem. At its core is the SpeechCodeVAE, which decomposes nonnative speech into three disentangled components: content tokens, speaker timbre, and prosody features. The conversion process is handled by a transformer-based LLM that maps nonnative tokens to their native counterparts. To address the "stochastic errors" (random glitches or inconsistencies) inherent in LLM-generated token sequences, the system includes a SpeechRestorer module, i.e., a postprocessor that ensures temporal coherence and prosodic continuity before the final waveform synthesis by a flow-based decoder.

Diffusion and flow-matching synthesis

TokAN [43] focuses on the use of self-supervised semantic tokens [derived from Hidden-Unit BERT (HuBERT)] to perform accent normalization. The model operates through a multistage pipeline where source speech is first quantized and deduplicated to isolate phonetic content. An autoregressive model then "translates" these nonnative token sequences into native-normalized versions. A significant innovation in TokAN is the use of a flow-matching duration predictor, which assigns proper timing to these normalized tokens. These expanded features then guide a diffusion-based mel decoder to iteratively generate high-fidelity spectrograms, which are finally converted to audio using a BigVGAN vocoder.

FAC-FACodec [45] addresses the need for controllable accent conversion by utilizing a factorized speech codec to disentangle speech representations. Unlike other models that perform a total conversion, this system employs a diffusion-based conversion strategy on the content representation. It introduces a user-controllable parameter, τ, to control the "strength" of the accent conversion by corrupting content features with Gaussian noise up to τ time step and iteratively denoising toward a native pronunciation prior. The model allows for a flexible balance: users can choose to retain some of the original speaker's characteristic accent or push for full native-like fluency.

Data augmentation

Although previous models focus on architecture, MacST [**44**] addresses the "data scarcity" problem in FAC through innovative data synthesis. The framework first uses LLMs to "transliterate" English text into similar representations in a target language's script (e.g., Hindi or Japanese). These transliterated texts are then fed into a multilingual TTS system to produce accented English speech that is perfectly aligned with the original native text, effectively creating a "synthetic parallel" corpus without requiring multi-accented human speakers. Therefore, MacST can be used to generate the high-quality supervised training sets necessary to advance state-of-the-art FAC research.

Authorized licensed use limited to: EssilorLuxottica SA. Downloaded on October 06,2026 at 08:48:14 UTC from IEEE Xplore. Restrictions apply.

General discrete speech token-based FAC pipeline

Despite variations in the underlying architectures, most modern discrete token-based FAC systems follow a unified four-stage pipeline, as shown in Figure 4:

1) *Decomposition and tokenization*: The L2 waveform is factorized into speaker-agnostic content tokens (semantic tokens in SpeechAccentLLM, TokAN, and MacST; acoustic tokens in FAC-FACodec), speaker timbre embeddings, and prosodic features.
2) *Latent token conversion*: A generative engine—such as an LLM, an autoregressive transformer, or a diffusion-based ODE solver—maps nonnative accent token sequences to native equivalents.
3) *Temporal and prosodic refinement*: Specialized postprocessing refinement modules, such as SpeechRestorer in SpeechAccentLLM and the duration predictor in TokAN, refine speech output by correcting token-level inconsistencies. This effectively enhances prosodic continuity and maintains precise duration alignment by predicting native-like timing.
4) *Reconstruction*: The converted token is recombined with the original speaker's identity and the target prosody to synthesize the target mels using a decoder, and then synthesize the final waveform using a neural vocoder.

Advantages and limitations of token-based FAC

The shift to discrete tokens challenges the current understanding of FAC by offering a stronger solution to the entanglement problem while introducing new computational and phonetic complexities. We analyze its strengths and limitations as follows:

Advantages: The utilization of discrete tokens offers three primary advantages:

1) *Hard information bottleneck*: Discretization forces the model to discard fine-grained acoustic variance, effectively "stripping" the speaker's timbre from the linguistic content.
2) *Data efficiency and scalability*: The utilization of high-quality native TTS models [42], [43] and the text transliteration method [44] significantly reduces reliance on scarce parallel FAC datasets by generating massive "pseudo-parallel" corpora.
3) *LLM compatibility*: The rapid advancement of LLMs has established discrete speech tokens as a foundational paradigm for speech representation. By discretizing speech into compact, text-like units, FAC can be framed as a seq2seq translation task. This allows speech to be seamlessly integrated into text-centric LLM architectures, leveraging their powerful modeling capabilities for cross-modal tasks.

Limitations: Despite these architectural advantages, the shift toward discrete token representations also introduces several inherent limitations:

- *Computational overhead*: The iterative nature of diffusion-based denoising [43], [45] and the inference demands of LLMs [42] typically require longer synthesis time compared to lightweight, non-autoregressive generative adversarial network models. This presents a significant barrier to real-time deployment.
- *The prosody gap*: Many token-based methods prioritize segmental (phonetic) accuracy. Because they often retain or scale original prosody, they may struggle to correct suprasegmental "prosodic accents," such as incorrect syllable stress or tunes.
- *Information granularity*: Relying on quantized tokens can lead to information loss. If the codebook or k-means clusters are not sufficiently granular, subtle phonetic distinctions or emotional nuances may be discarded.
- *Dependency on pretrained priors*: The quality of conversion is heavily dependent on the underlying tokenizer (e.g., Whisper or HuBERT) or TTS model. If these models contain inherent linguistic biases, the FAC system may struggle with unnatural speech patterns or produce "caricatured" accent patterns.

Feature disentanglement strategies in FAC

A fundamental challenge across all DL-based FAC approaches is feature disentanglement—the separation of accent characteristics from speaker identity and linguistic content within

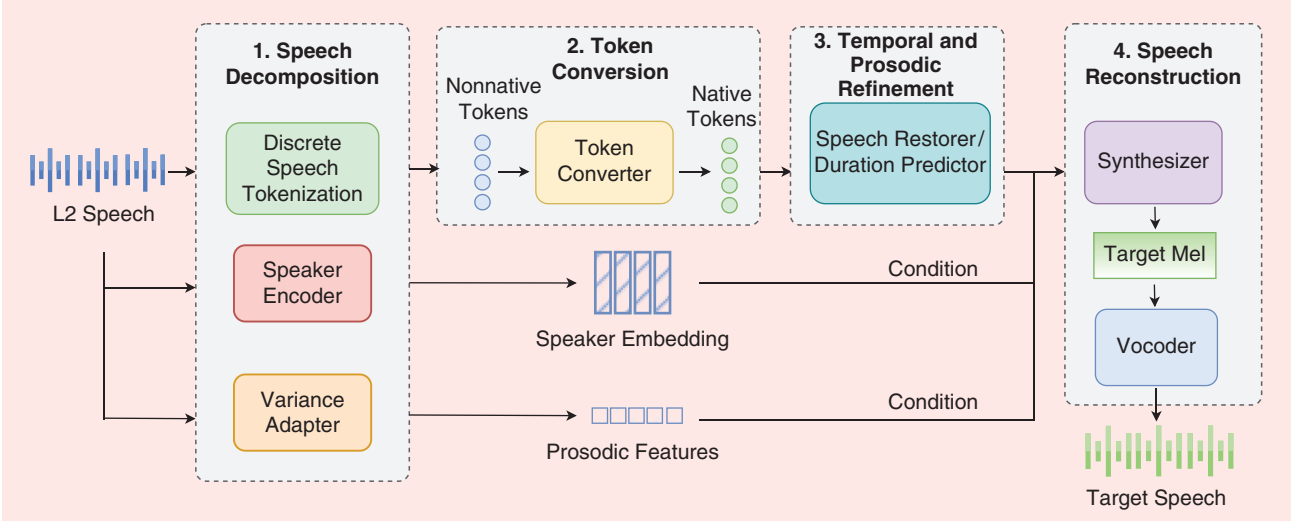

FIGURE 4. General discrete speech token-based FAC pipeline.

Authorized licensed use limited to: EssilorLuxottica SA. Downloaded on October 06,2026 at 08:48:14 UTC from IEEE Xplore. Restrictions apply.

learned representations. In FAC, accent and speaker identity are encoded through overlapping supra-segmental and segmental features—such as F0, pitch contour, and spectral characteristics—making their disentanglement in learned representations inherently difficult. The quality of the feature disentanglement directly determines conversion performance. Therefore, recent work explores several strategies to achieve this goal. This section surveys the principal disentanglement strategies employed in FAC and in closely related accented speech synthesis systems, organized by their underlying mechanisms.

Variational autoencoder-based disentanglement

Standard variational autoencoders (VAEs) struggle to completely separate accent from speaker identity and linguistic content features because they treat observations as independent and identically distributed. To overcome this, recent literature has heavily adopted multilevel VAEs (ML-VAEs) [47], [48], [49]. ML-VAEs leverage group-level supervision by assuming that observations within a specific group (e.g., speech samples of the same accent) share a common factor of variation. By maximizing the group evidence lower bound using minibatches [50], an ML-VAE disentangles the latent space into independent variables: a style variable z_s (speaker identity) and a group-level content variable z_a^g (accent). The total loss for optimizing an ML-VAE incorporates a reconstruction loss between the predicted mel spectrogram $\hat{X}$ and the ground-truth mel spectrogram X: $L_{\text{recon}} = \| \hat{X} - X \|_2$, where $\| \cdot \|_2$ denotes L_2 norm. The total loss also includes a KL divergence loss $\mathcal{L}_{\text{kl}}$ scaled by a coefficient β. $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{recon}} + \beta \mathcal{L}_{\text{kl}}$.

To further enforce disentanglement, recent models such as DART [47] incorporate vector quantization (VQ) into the ML-VAE encoder. By mapping the continuous latent variables to a predefined codebook of discrete vectors, the VQ layer acts as a bottleneck, filtering out irrelevant information. This integration is guided by an additional commitment loss $\mathcal{L}_c$, which effectively reduces the complexity of the latent space and promotes better separation between speaker and accent embeddings.

Adversarial training for feature separation

Adversarial learning is a widely used strategy for explicitly wiping out residual, entangled information from latent representations [48], [51]. Because an ML-VAE might still allow accent information to leak into the speaker embedding (or vice versa), adversarial training introduces an auxiliary classifier (a discriminator) to enforce strict separation [48]. For example, to remove speaker identity from linguistic BNFs, an auxiliary speaker classifier is attached to the content encoder. The classifier is trained via cross-entropy loss to predict the speaker, while the generator (encoder) is optimized through an adversarial loss $\mathcal{L}_{\text{adv}}$ to fool the classifier. This is commonly formulated by forcing the classifier's output probability (p) to approximate a uniform distribution vector (e), thereby confusing the discriminator: $\mathcal{L}_{\text{adv}} = \| p - e \|_2^2$. Conversely, this exact formulation is used to wipe accent information out of speaker embeddings by applying an adversarial accent classifier in [48], ensuring that speaker embeddings do not dominate the accent generation during inference.

Additional disentanglement mechanisms

Beyond VAEs and adversarial training, some other advanced mechanisms are also utilized to isolate accents: 1) mutual information (MI) minimization: some frameworks explicitly minimize the MI between the extracted speaker and accent latent variables [52]. By introducing an MI penalty term ($\mathcal{L}_{\text{MI}}$) into the total loss function, the model is penalized for learning common features across different representations, thereby mathematically enforcing the latent variables to remain statistically independent. This constraint significantly reduces cluster overlap in the latent space, enabling robust disentanglement and precise, independent control over speech features like accent, pitch, and pause duration. 2) Unsupervised speech decomposition: Huang et al. [12] utilize pretrained self-supervised models (like wav2vec 2.0 or HuBERT) to decompose speech into highly specific independent attributes—such as linguistic content, rhythm, pitch energy, and speaker style—without relying on parallel data or transcriptions. By applying an adversarial accent discriminator directly to the isolated speaker attribute, they successfully strip away the accent features from the speaker's style, forcing the accent to be captured by a stand-alone, learnable embedding. 3) Pseudo-Siamese disentanglement network (PSDN): Jia et al. [10] utilize dual streams (a target stream and an auxiliary stream), PSDN leverages structural asymmetry to filter out source accents. Because the auxiliary stream receives additional raw, entangled acoustic features, it acts as an "exit" that easily absorbs the source accent. This forces the target stream's content representation to drop the source accent, allowing the model to generate the new target accent based purely on the clean linguistic content and the target speaker's timbre.

Balancing disentanglement and speaker identity

Differences in speaker identity preservation across FAC systems largely stem from how effectively they disentangle accent and speaker information. In models without explicit separation, residual source-accent cues can leak into the speaker embedding, interfering with target-accent synthesis and reducing voice similarity. Methods such as adversarial accent removal, vector-quantized bottlenecks, and architectural asymmetry improve preservation by isolating speaker timbre from accent-related features. However, this creates an inherent tradeoff: because accent and speaker traits partly overlap in the acoustic space, overly strong disentanglement may remove identity-relevant cues, degrading speaker similarity and naturalness.

Datasets, evaluation metrics, and comparative DL-based FAC system overview

Evaluation metrics

The commonly used subjective and objective evaluation metrics are summarized in Tables 3 and 4, respectively. From Table 3 we can see that recent FAC studies primarily assess the synthesized speech of proposed systems using three subjective evaluation metrics: accentedness, speech naturalness, and voice similarity. These evaluations are typically conducted

Authorized licensed use limited to: EssilorLuxottica SA. Downloaded on October 06,2026 at 08:48:14 UTC from IEEE Xplore. Restrictions apply.

via listening tests on platforms such as Amazon Mechanical Turk (AMT), where participants were recruited to rate different metrics using standardized scales. For instance, acoustic quality is commonly rated using a five-point mean opinion score (MOS) scale. At the same time, voice similarity is evaluated on a 14-point scale, and accentedness is measured with a nine-point Likert scale [38]. Additionally, [53] evaluates prosody transfer by analyzing differences in prosodic features such as duration, average F0, and F0 range between original utterances and the converted speech.

Recent studies have also employed a diverse set of objective evaluation metrics for assessing the proposed FAC systems. For example, [54] used pretrained ASR models to measure word error rate (WER) and character error rate for both original and converted speech. Ding et al. [38] utilized t-distributed stochastic neighbor embedding (t-SNE) to visualize speaker and accent embedding spaces. Quamer et al. [53] applied t-SNE to analyze embeddings for source, target, and converted speech while also examining the effect of codebook size on mel cepstral distortion.

Table 3. Subjective evaluation metrics for FAC.

Subcategory	Methods	Description/Scales	References
Accentedness	9-point MOS (↓)	MOS (1: no accent; 9: strong accent)	[12], [35], [36], [38], [55], and [56]
	XAB/AB preference tests	Listener preference between options X, A, and B	[3], [4], [57], [58], and [59]
	5-point MOS (↓)	MOS (1: no accent; 5: strong accent)	[2], [4], [10], [11], [60]†, [5]†, and [61]†
	Best–worst scaling (BWS)	BWS: $(N_{\text{best}} - N_{\text{worst}})/N_{\text{total}}$	[9] and [49]
Voice similarity	7-point voice similarity score (VSS) (↑)	VSS (1: not confident; 7: very confident)	[36] and [56]
	14-point VSS (↑)	VSS (–7: different speaker; +7: same speaker)	[38], [35], and [59]
	5-point VSS (↑)	VSS (1: not close; 5: very close)	[2], [4], [10], [49], [57], [60], and [61]
	XAB speaker test	Speaker identity preference between options X, A, and B	[3], [4], [11], and [12]
	BWS	BWS: $(N_{\text{best}} - N_{\text{worst}})/N_{\text{total}}$	[9] and [43]
Speech naturalness	5-point MOS (↑)	MOS (1: bad; 5: excellent)	[3], [4], [10], [11], [12], [35], [36], [38], [42], [49], [55], [56], [57], [58], [59], and [62]
	Multiple stimuli with hidden reference and anchor (MUSHRA) scores (↑)	MUSHRA: perceptual evaluation for speech quality	[9] and [43]
Others	XAB general tests	Evaluates overall listener perception	[62]

N: The (best, worst, or total) number in BWS; (↑): higher values are better; (↓): lower values are better; †: the specific study that has the reverse scale direction within its category.

Table 4. Objective evaluation metrics for foreign accent conversion.

Subcategory	Methods	Description/Scales	References
Accentedness	Accent embedding t-distributed stochastic neighbor (SNE)	t-SNE for visualizing accent embedding distributions	[38], [62]
	Longest common subsequence ratio (LCSR)	LCSR: measures semantic token similarity between converted and target sequences	[58]
	Common accent classification	Predict the accent of the synthesized speech	[59]
	Accent embedding similarity (AES) (↑)	AES: measures accent conversion accuracy (1–5)	[12]
	Accent embeddings cosine similarity (AECS)	Utterance-level AECS	[4]
	Accent classifier accuracy (ACC)	ACC	[2]
Voice similarity	Speaker embedding t-SNE	t-SNE for visualizing speaker embedding distributions	[38], [55], [62]
	Speaker embedding similarity (SES) (↑)	SES: evaluates speaker preservation	[12]
	Speaker embedding cosine similarity (SECS) (↑)	SECS (closer to one is better)	[4], [5], [43], [45], [60], [61]
	Equal error rate (EER) (↓)	EER for speaker verification	[49]
Speech naturalness	Mel cepstral distortion (MCD) (↓)	MCD: measures synthesis quality	[36], [49], [55], [62]
	MOSNet (↑)	Predicted MOS	[12]
Others	Word error rate (WER) (↓)	WER: measures intelligibility using an ASR system	[2], [5], [11], [39], [42], [43], [45], [49], [55], [62]
	Character error rate (CER) (↓)	CER	[12]
	Difference in duration (DDUR) (↓)	Absolute DDUR between L1 and L2 ground-truth speeches	[36]
	F0 Pearson's correlation coefficient (PCC) (↑)	PCC: evaluates pitch similarity	[4]
	F0 root-mean-square error (RMSE) (↓)	RMSE: evaluates pitch similarity	[4], [36]

(↑): Higher values are better; (↓): lower values are better.

Authorized licensed use limited to: EssilorLuxottica SA. Downloaded on October 06,2026 at 08:48:14 UTC from IEEE Xplore. Restrictions apply.

Datasets

The ARCTIC [63] and L2-ARCTIC [64] datasets are foundational resources in FAC research, providing native and nonnative English speech corpora, respectively. However, their limitations pose significant challenges to the development of robust FAC systems. These challenges are delineated in the next section, emphasizing their impact on accent conversion tasks.

Limited diversity in nonnative speaker representation

The L2-ARCTIC dataset has expanded to include 24 speakers across six L1 language groups—Hindi, Mandarin, Vietnamese, Korean, Arabic, and Spanish—with each comprising four speakers (two male, two female). Although this represents an improvement over the initial release of 10 speakers in [64], the dataset still captures only a fraction of the global diversity in nonnative English accents. With only four speakers per L2, L2-ARCTIC struggles to capture intra-language variability in accent, which is influenced by factors such as regional dialects, proficiency levels, and individual speaker characteristics. This limited representation hampers the generalizability of FAC models. Therefore, FAC research requires datasets with broader L2 coverage and larger speaker cohorts to capture the full spectrum of nonnative accent variation.

Reliance on controlled read speech

Both ARCTIC and L2-ARCTIC consist of read speech based on 1,132 Carnegie Mellon University (CMU) ARCTIC prompts. This controlled, scripted setting lacks the prosodic and contextual variability of spontaneous speech, which is critical for real-world FAC applications. Models trained on such data may fail to generalize to conversational scenarios, where accent features are influenced by the natural dynamics of speech. The absence of spontaneous speech in L2-ARCTIC limits its utility for developing FAC systems capable of handling diverse speaking styles, necessitating the inclusion of unscripted speech corpora.

Scarcity of parallel data

FAC systems often rely on parallel corpora of native and nonnative speech for training. Although L2-ARCTIC provides aligned orthographic and phonetic transcriptions using the Montreal Forced Aligner, its structure—approximately 1,130 utterances per speaker—limits the number of parallel data. With 26 hours of total recorded speech, L2-ARCTIC offers a larger corpus than its initial release; however, the lack of extensive parallel data still poses a challenge for supervised FAC approaches. This scarcity drives the need for semisupervised or unsupervised methods, which may compromise conversion quality, emphasizing the importance of developing strategies to generate or simulate parallel data.

Predominant focus on English-centric accents

Most FAC research, including studies using ARCTIC and L2-ARCTIC, focuses on English as the target language. This English-centric approach overlooks the need for multilingual, multiaccent datasets to support broader language-learning applications. For instance, language learners studying a second language other than English (e.g., English speakers learning Mandarin) would benefit from FAC systems trained on diverse language pairs and accents. The absence of such datasets limits the applicability of FAC in global language education contexts, where multilanguage accent conversion could enhance pronunciation training across various L1–L2 pairs. Future dataset development should prioritize multilanguage corpora to address this gap and support more inclusive FAC applications.

Alternative benchmarks

To mitigate the limitations of CMU ARCTIC and L2-ARCTIC as primary benchmarks, FAC evaluation can be extended using a hybrid protocol that combines existing multiaccent corpora with synthetic parallel speech generation. Several publicly available datasets provide complementary resources for pretraining and robustness evaluation. For example, the CSTR VCTK corpus [65] contains recordings from 110 English speakers with diverse accents and roughly 400 utterances per speaker, providing broader speaker diversity than L2-ARCTIC. The AESRC2020 dataset [66] includes approximately 180 h of accented English speech across 10 accent types and introduces unseen accents in the evaluation set, enabling cross-accent generalization testing. In addition, AccentDB [67] provides approximately 20 h of speech from multiple Indian–English and native English accents, while the Speech Accent Archive [68] contains more than 3,000 recordings from speakers representing more than 200 L1 backgrounds. These datasets can support accent representation learning and robustness evaluation across diverse accents.

Beyond existing corpora, scalable synthetic benchmarks can be constructed using recent advances in LLMs and multilingual TTS systems. Recent work, such as MacST (multiaccent speech synthesis) [44], demonstrates that LLM-based text transliteration, combined with multilingual TTS, can generate parallel accented speech while preserving linguistic content. This approach enables large-scale generation of native and L2-accented speech pairs across many L1 backgrounds. In addition, conversational evaluation scenarios can be simulated by generating multiturn dialogue speech using accented TTS rather than isolated read sentences, allowing FAC systems to be evaluated under more realistic prosodic and discourse conditions. Together, this hybrid protocol—combining real multiaccent datasets with controlled synthetic speech—provides a scalable and reproducible benchmark for FAC research, improving accent diversity, speaker coverage, and evaluation robustness beyond the constraints of L2-ARCTIC.

Comparative overview of DL-based FAC systems

To bridge the gap between conceptual understanding and practical implementation, Table 5 provides a component-level breakdown of several representative deep-learning FAC systems surveyed in this article, organized into three main

Authorized licensed use limited to: EssilorLuxottica SA. Downloaded on October 06,2026 at 08:48:14 UTC from IEEE Xplore. Restrictions apply.

Table 5. Comparative overview of deep learning FAC systems: model components, input–output data flow, training dataset of each model component, vocoder, parallel-data requirement, and released resources, summarized per system.

Category	system	Main Model (Submodels)	Input → Output	Train Dataset (Pretrained Model)	Vocoder	Par. Tr/Inf	Res.
S2S	Accentron [38] (2022)	Speaker encoder	Source speech → speaker emb.	VoxCeleb1	WaveRNN	✓/✓	Demo
		Accent encoder	Source speech → accent emb.	Speech accent archive			
		Acoustic model	Source speech → BNFs	LibriSpeech			
		Seq2seq	BNFs + accent + speaker emb. → mel spec.	L2-ARCTIC + ARCTIC			
	ZS-NoRef [35] (2022)	Speaker encoder	Source speech → speaker emb.	VoxCeleb1/2 + LibriSpeech	WaveRNN	✓/×	1). Demo; 2). Code
		Accent encoder	Source speech → 256-d accent emb.	Speech accent archive			
		Acoustic model	Source speech → BNFs	LibriSpeech			
		Seq2Seq translator	L2 BNF → L1 BNF	L2-ARCTIC + ARCTIC			
		Seq2seq synthesizer	L1 BNF + speaker emb. → mel spec.	L2-ARCTIC + ARCTIC			
TTS	E2E-NoNat. [3] (2020)	Multispeaker TTS (TTS Enc. + Dec.)	Phoneme seq. + speaker emb. → mel spec.	VCTK	WaveRNN	×/×	Demo
		Multitask accented ASR (ASR Enc. + Dec.)	Mel spec. + accent emb. → text emb. + phoneme labels	VCTK			
		Speaker encoder	Mel spec. → speaker emb.	VoxCeleb1/2 + LibriSpeech			
	TTS-guided-NoPar [39] (2023)	Pretrained TTS (Text Enc. + Speech Dec.)	Phoneme seq. → mel spec.	VCTK	WaveGAN	×/×	Demo
		Speech encoder	Mel spec. → BNF speech emb.	L2-ARCTIC			
	Diff.-TTS [9] (2024)	Text encoding (Text Enc. + Dur. Adapter)	Phoneme seq. + speaker emb. → frame-level linguistic repr.	LibriTTS-R	BigVGAN	×/×	Demo
		Speech decoder (Var. Adapter+DiT Dec.)	linguistic repr. + speaker emb. → mel spec.	LibriTTS-R			
		Speech encoding (Whisper Feat.+Speech Enc.)	Source speech → linguistic repr.	L2-ARCTIC			
Token	Speech Accent-LLM [42] (2025)	Content encoder (Whisper+Pre-VQ+VQ)	Perturbed waveform x' → discrete content tokens V	Multilingual base corpus	None (VITS dec.)	×§/×	—
		Speaker encoder	Source waveform x → speaker emb. S	(ECAPA-TDNN)			
		Speech reconstruction (Post-VQ+Var.+Flow+Dec.)	V + S + prosody P → waveform	Multilingual base corpus			
		FAC&TTS model	FAC: L2 tokens/text + Task ID → L1 tokens/speech tokens	FAC&TTS Corpus			
		SpeechRestorer	Predicted tokens → recovered tokens	FAC&TTS Corpus			
	TokAN [43] (2025)	HuBERT tokenizer (HuBERT+K-means+dedup)	Source speech → dedup. HuBERT tokens	(HuBERT-large)	BigVGAN	×§/×	1. Demo, 2. Code
		Speaker encoder	Source speech → speaker emb.	(Resemblyzer)			
		Accent encoder	Source speech → accent emb.	(CommonAccent XLSR-en)			
		AR conversion model	Source tokens + accent emb. → target tokens	L2-ARCTIC + ARCTIC + Chinese-accent ASR			
		NAR synthesizer (Token Enc.+Dur.+Mel Dec.)	Target tokens + speaker emb. +dur. → mel spec.	LibriTTS-R			
	FAC-FACodec [45] (2025)	FACodec encoder	Waveform → repr. h: content $z_c = z_{c1} + z_{c2}$, prosody z_p, acoustic details z_d, and global timbre emb. g.	(NaturalSpeech 3)	None (FACodec Dec.)	×/×	Demo
		Denoiser s_θ	Noisy content z_{c1} → noise ε → denoised content $\hat{z}_{c1}$	LJSpeech			
		Second residual predictor q_ϕ	$[h, \mathrm{detach}(\hat{z}_{c1})]$ → second content residual $\hat{z}_{c2}$	LJSpeech			
		FACodec decoder	$\hat{z}_{c1}, \hat{z}_{c2}, z_p, z_d, g$ → waveform	(NaturalSpeech 3)			

Note that each row is one trainable or pretrained model or its submodels within parentheses. Abbreviations: par.Tr/Inf: parallel data required at training/inference (✓: required, ×: not required); §: no *real* parallel data—synthetic native targets are generated by TTS during training; res: publicly released resources (audio demo/code); emb: embedding; enc.: encoder; dec.: decoder; repr.: representation; mel spec.: mel spectrogram; seq.: sequence; var: variance adapter; dedup.: deduplicated; Dur: duration predictor; dur.: duration values.

Authorized licensed use limited to: EssiorLuxottica SA. Downloaded on October 08,2026 at 08:48:14 UTC from IEEE Xplore. Restrictions apply.

Table 6. Meta-analysis data: evaluation scores and sample sizes.

Study	L1	L2	Conv	WER	N	R	C	Category
Panel A: nine-point accentedness ↓ (1 = native, 9 = heavy accent)								
[36] FAC-NoRef (2021)	1.04 ± 0.04	6.77 ± 0.2	5.33 ± 0.28	—	800	20	United States	S2S
[38] Accentron (2022)	1.06 ± 0.12	7.11 ± 0.21	3.3 ± 0.26	—	360	18	United States	S2S
[35] ZS-NoRef (2022)*	1.1*	6.78*	2.23*	—	800	20	United States	S2S
[39] TTS-guided-NoPar (2023)*	2.03*	7.98*	2.87*	15.8	250	25	United States	TTS
[55] AC-Articulatory (2024)†	8.56 ± 0.14	2.07 ± 0.37	7.32 ± 0.51	17.8	300	20	United States	S2S
Panel B: five-point accentedness ↑ (1 = heavy accent, 5 = native)								
[60] AC-DiscreteUnits (2024)	—	2.12 ± 0.05	4.42 ± 0.11	—	1000	20	United States	Token
[61] Streaming-NAR (2025)‡	—	1.12 ± 0.15	3.78 ± 0.18	14.1	380‡	10	—	TTS
[5] KD-AC (2025)‡	—	1.67 ± 0.05	3.87 ± 0.08	12.4	380‡	10	—	TTS
Panel C: five-point naturalness ↑ (1=unnatural, 5=natural)								
[36] FAC-NoRef (2021)	4.63 ± 0.06	3.7 ± 0.06	3.22 ± 0.1	—	800	20	United States	S2S
[38] Accentron (2022)	4.9 ± 0.1	3.67 ± 0.28	3.43 ± 0.12	—	360	18	United States	S2S
[35] ZS-NoRef (2022)*	4.63*	4.44*	3.03*	—	800	20	United States	S2S
[11] TTS-Transfer (2024)	—	4.43 ± 0.08	3.71 ± 0.08	17.7	400	20	United States	TTS
[55] AC-Articulatory (2024)	3.46 ± 0.18	3.36 ± 0.43	2.45 ± 0.33	17.8	300	20	United States	S2S

L1: native speaker baseline; L2: original L2 speech; conv: converted speech; N: sample size (utterances); R: number of raters; C: reported rater citizenship; category: S2S=seq2seq; TTS: TTS-guided; token: token-based; *: SD imputed via CV method; †: inverted scale (1=heavy accent, 9=native); ‡: sample size imputed (N = 380); ↑: higher value is better; ↓: lower value is better.

categories: seq2seq, TTS guided, and token based. For each system, the table lists every model component, its input and output representations, and the training dataset or pretrained model used. It also specifies the vocoder, whether parallel data are required at training and inference, and links to publicly available demos or code. This level of detail is intended to serve as a practical reference. At a glance, a reader can identify what feature representations each model component expects, which datasets are needed, and how the model components connect to form the full pipeline.

Meta-analysis of current FAC research

We conduct a systematic meta-analysis to synthesize evaluation outcomes across FAC studies quantitatively. From Table 3, we select studies employing the two most prevalent subjective metrics: accentedness and naturalness. To ensure methodological rigor and comparability, we apply the following article inclusion criteria: 1) recent publications from 2021 to 2025, 2) transparent reporting of evaluation scores with standard deviations (SDs) or sufficient information for imputation, 3) documented sample sizes including number of raters, and 4) an evaluation of comparable L1–L2 accent pairs.

Based on these criteria and to quantify accent reduction effectiveness across FAC systems, we conducted a meta-analysis on two perceptual scales: the nine-point accentedness scale (1 = native, 9 = heavy accent), which includes five studies used by seq2seq and TTS-guided FAC methods, and 2) the five-point accentedness scale (1 = heavy accent, 5 = native) which includes three studies used by the token-based FAC method. In both cases, Hedges' g is computed to assess whether the accent of the converted speech improves toward that of native speech and to examine whether accent-reduction findings generalize across measurement paradigms.

The five-point scale (1 = unnatural, 5 = natural) naturalness analysis assessed the extent to which FAC systems preserve speech quality relative to the original L2 baseline. We include the same five studies as in the nine-point accentedness analysis, with one substitution: TTS-Transfer (2024) [11] replaces TTS-Guided-NoPar (2023) [39] due to the availability of a naturalness metric. The complete evaluation data are presented in Table 6. To support reproducibility, the complete meta-analysis code, forest plots, and full evaluation data (including the comprehensive system comparison table) are openly available.[a]

Meta-analysis methodology

Missing SD estimation

Two studies, ZS-NoRef (2022) [35] and TTS-Guided-NoPar (2023) [39], report only mean values without SDs. Direct exclusion of SDs would potentially introduce bias into the results, reducing statistical power by 40% for nine-point accentedness and by 20% for five-point naturalness analyses. Therefore, we use the coefficient of variation method to estimate missing SDs because it provides a principled approach to estimating SDs on bounded scales.

Compute CVs

We first compute the coefficient of variation (CV) values for studies with both SDs and mean values by $\mathrm{CV}_i = (\sigma_i/\mu_i)$, where σ, μ indicates SD and mean value, respectively, for study i.

Estimate missing SDs

We then calculated the median CVs from complete studies with known CVs and applied these to estimate the missing SDs conservatively by $\hat{\sigma}_{\text{missing}} = \mu_{\text{missing}} \times \text{median}(\text{CV}_{\text{known}})$.

[a]https://github.com/ChenxiLiu6/FAC-Meta-Analysis.

Authorized licensed use limited to: EssilorLuxottica SA. Downloaded on October 06,2026 at 08:48:14 UTC from IEEE Xplore. Restrictions apply.

The median CV is robust to outliers and preserves the proportional relationship between mean and variance characteristic of MOS scales.

Standardized effect size calculation

Raw MOS differences cannot be directly compared across studies due to varying baseline levels and sample sizes. Standardization enables meaningful synthesis while accounting for small-sample bias.

Cohen's d (standardized mean difference)

Cohen's d measures the standardized mean difference between two groups by $d = (\mu_{\text{converted}} - \mu_2)/\sigma_{\text{pooled}}$ where $(n_1, n_2), (\sigma_1, \sigma_2)$ are the sample sizes and SDs for the first (converted) and second (L2) groups, respectively. $\sigma_{\text{pooled}} = \sqrt{((n_1-1)\sigma_1^2 + (n_2-1)\sigma_2^2)/(n_1+n_2-2)}$ is the pooled SD, which combines the variability from both groups into a more reliable estimate because a single group's SD could bias the effect size if the groups have different variabilities.

Hedges' g (bias correction)

Cohen's d tends to slightly overestimate the true effect size when sample sizes are small (e.g., studies with sample sizes smaller than 50). Therefore, we use Hedges' g to correct Cohen's d for small-sample bias by $g = d \times J(df)$, where d is Cohen's d, and $J(df) = 1 - 3/(4df - 1)$ is the small-sample correction factor; it approaches one as the sample size increases, converging with Cohen's d for large samples. $df = n_1 + n_2 - 2$ is the degree of freedom.

This standardization step allows meaningful comparison across studies despite varying sample sizes (250–800 utterances). For naturalness, negative g values indicate degradation (e.g., converted <L2), positive ones indicate improvement. This contrasts with accent reduction, where negative values are desirable.

Random-effects meta-analysis

FAC systems employ fundamentally different architectures; random-effects modeling combines results from different studies by assigning greater weight to studies with more precise estimates (i.e., smaller variance). This ensures that more reliable studies have proportionally greater influence on the final estimate.

Specifically, we employed the DerSimonian–Laird random-effects model to pool effect sizes. The weighted average of individual effect sizes is computed by $\bar{g} = \Sigma(w_i \times g_i)/\Sigma w_i$, where $w_i = 1/(\sigma_i^2 + \tau^2)$ is the random-effect weight for study i based on its precision (inverse of variance). $\sigma_i^2 = (n_1+n_2)/(n_1 \times n_2) + g^2/[2(n_1+n_2)]$ is the within-study variance. τ^2 is the between-study variance [calculated in the next step (4-c)]. The incorporation of both within- and between-study variance provides optimal precision in the presence of heterogeneity. The 95% confidence interval (CI) for the pooled effect is $\text{CI} = \bar{g} \pm 1.96 \times \text{SE}(\bar{g})$, where the standard error is $\text{SE}(\bar{g}) = \sqrt{1/\Sigma w_i}$.

Heterogeneity analysis

Heterogeneity analysis quantifies whether FAC systems are truly different. If there are real differences (high heterogeneity), how different are they?

Cochran's Q statistic (total heterogeneity)

$Q = \Sigma[w_i \times (g_i - \bar{g})^2]$. Q measures total variation. It asks, "Is there a real difference between FAC systems?"

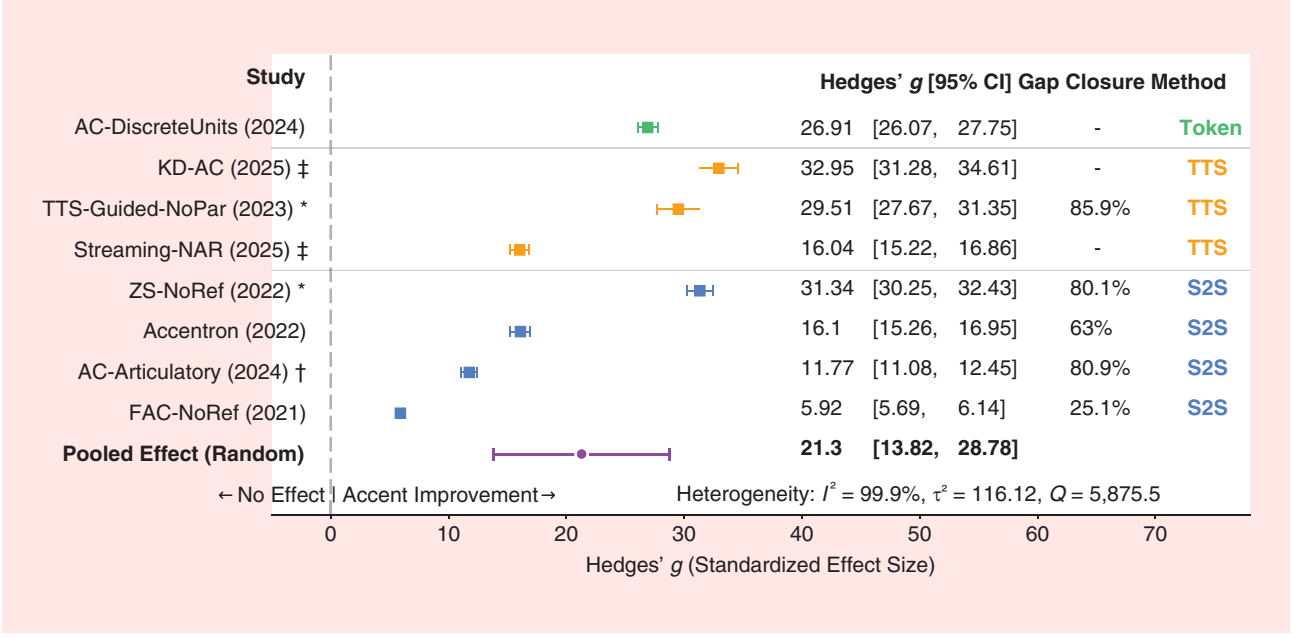

FIGURE 5. Random-effect forest plots for cross-scale accentedness analysis (k = 8). Positive effect = accent improvement toward native speech. *: SD imputed via CV method; †: inverted scale (1 = heavy, 9 = native); ‡: sample size imputed ($N = 380$).

Authorized licensed use limited to: EssilorLuxottica SA. Downloaded on October 06,2026 at 08:48:14 UTC from IEEE Xplore. Restrictions apply.

I^2 statistic (percentage due to real differences)

$I^2 = \max[0, (Q - df)/Q \times 100\%]$, where $df = k - 1$, and k is the total number of studies. I^2 measures what percentage of the differences between studies is real versus just sampling error. I^2 heterogeneity: low (<25%), moderate (25%–50%), high (50%–75%), and very high (> 75%).

Between-study variance (τ^2)

$\tau^2 = \max(0, (Q - df)/C)$ where $C = \Sigma w_i - (\Sigma w_i^2/\Sigma w_i)$, and τ^2 is added to each study's variance to account for between-study differences.

Specific analysis steps

Accent reduction

Gap Closure (%) $= (\mu_2 - \mu_{\text{converted}})/(\mu_2 - \mu_1) \times 100$. It evaluates the relative improvement in accent by comparing the current gap between the converted and L2 speeches to the original gap between L2 and L1 speeches.

Naturalness preservation

Preservation $= \mu_{\text{converted}}/\mu_2 \times 100$. It evaluates the preservation of naturalness in the converted speech compared to the original L2 speech.

Accent-naturalness tradeoff

For studies that report both accentedness and naturalness results, we computed Pearson r and Spearman ρ to assess the direction and strength of any accent-naturalness tradeoff.

Computational cost analysis

We use a real-time factor (RTF) = Total processing time/Total audio duration to examine the FAC system's computational cost.

Meta-analysis results and discussion

Accentedness reduction

Across the nine-point accentedness scale, individual random-effect sizes (g) ranging from 5.92 (FAC-NoRef) to 31.34 (ZS-NoRef). Accent gap closure, defined as *the proportion of the L2–L1 accentedness distance eliminated by conversion*, averaged 67% across five studies. Within this group, four of the five systems exceeded 60% closure, with the TTS-Guided-NoPar method achieving the highest performance, 85.9%. Three studies that used the five-point accentedness scale yielded comparable effect sizes (g = 16.04–32.95), all of which represent substantial standardized improvements. However, accent gap closure is not calculated for this group because L1 accentedness mean values are missing.

Given that Hedges' g is scale-invariant—representing effects in pooled SD units—we combined all eight studies into a single cross-scale random-effects analysis. The resulting pooled estimate of g = 21.3 (95% CI [13.82, 28.78]; Figure 5) indicates a substantial accent reduction across all FAC architectures. Furthermore, no statistically significant difference was found between the nine-point-scale ($g = 18.93$) and five-point-scale ($g = 25.3$) groups. This consistency supports the convergent validity of these independently developed evaluation methods.

Heterogeneity was extremely high across all analyses ($I^2 = 99.9\%, \tau^2 = 116.12, Q = 5{,}875.5$ for the cross-scale analysis), indicating that more than 99% of the observed variance in effect sizes reflects genuine differences between FAC systems rather than sampling error. It also suggests that system-level design decisions, such as conversion strategy, training paradigm, and the choice of speech representation, have a large impact on accent conversion outcomes.

Naturalness preservation

The five-point naturalness meta-analysis ($k = 5$) assessed the extent to which FAC systems preserve speech quality relative to the original L2 baseline. The random-effects pooled effect was $g = -5.37$ (95% CI [−8.27, −2.48]; Figure 6), confirming that all systems degrade naturalness to some degree after accent conversion. Naturalness preservation, computed as $\mu_{\text{converted}}/\mu_2 \times 100$, ranged from 68.2% (ZS-NoRef) to 93.5% (Accentron), with a pooled mean of 81.1%. Notably, none of the studies [5], [39], [60], [61] reported naturalness scores, precluding any assessment of their naturalness preservation.

The magnitude of degradation varied considerably across systems. Accentron exhibited mild degradation (g = −1.11, L2 Pres = 93.5%), while TTS-Transfer (g = −8.99, L2 Pres = 83.7%) and ZS-NoRef (g = −8.59, L2 Pres = 68.2%) showed substantially larger negative effects. This variability was reflected in the heterogeneity statistics ($I^2 = 99.9\%, \tau^2 = 10.88, Q = 2{,}967.2$), confirming that naturalness outcomes are highly system dependent.

To examine whether accent reduction comes at the cost of naturalness, we computed Pearson's r and Spearman's ρ between gap closure and naturalness preservation for the four studies

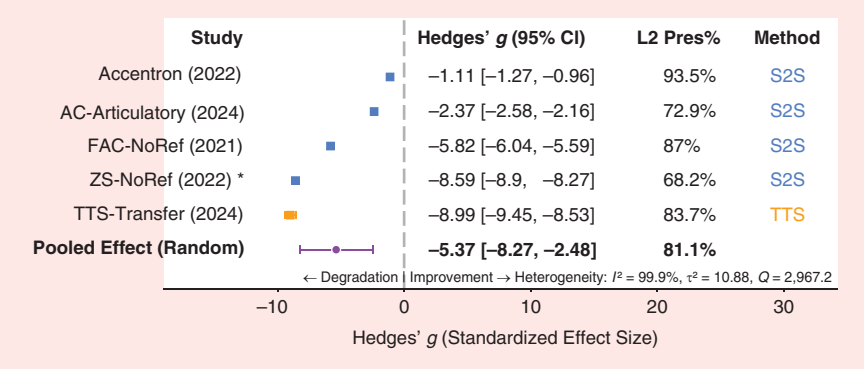

FIGURE 6. Random-effect forest plots for five-point naturalness ↑ meta analysis (k = 5). Negative effect = naturalness degradation from L2 baseline. ∗: SD imputed via CV method; L2 Pres%: naturalness preserved relative to native L2 speech.

Authorized licensed use limited to: EssilorLuxottica SA. Downloaded on October 06,2026 at 08:48:14 UTC from IEEE Xplore. Restrictions apply.

[35], [36], [38], [55] that report both metrics. A moderate negative correlation was observed ($r = -0.64$, $\rho = -0.6$), suggesting that systems achieving greater accent reduction tend to sacrifice more naturalness. Pearson's r captures the linear strength of this relationship, while Spearman's ρ, which operates on ranks rather than raw values, provides a distribution-free robustness check. The agreement between the two strengthens confidence in the trend. Neither reached statistical significance ($p = 0.359$ and $p = 0.4$) due to the limited sample ($k = 4$), but the consistent direction points toward a genuine quality-accent tradeoff.

Sensitivity analysis for CV imputation

Two of the five nine-point studies (ZS-NoRef and TTS-Guided-NoPar) required SD imputation via the CV method. To assess the robustness of the pooled estimate to this imputation, we conducted a sensitivity analysis under three scenarios. 1) The baseline scenario used the median CV from complete studies (L2 CV = 0.03, Conv CV = 0.07), yielding a pooled effect of $g = 16.49$ (95% CI [8.79, 24.18]). 2) A conservative scenario using the maximum observed CV (inflating imputed SDs) reduced the pooled estimate to $g = 8.76$ [6.51, 11.01], while 3) a liberal scenario using the minimum CV increased it to $g = 17.44$ [9.3, 25.58]. Across all three scenarios, the pooled-effect size remained large and statistically significant (all lower CI bounds greater than zero), and heterogeneity remained extremely high ($I^2 > 99\%$). We further performed a leave-one-out analysis by excluding each study in turn. Excluding ZS-NoRef (imputed) yielded $g = 14.26$ [7.75, 20.76], and excluding TTS-Guided-NoPar (imputed) yielded $g = 14.78$ [6.52, 23.05]. Excluding any of the three complete studies produced pooled estimates ranging from 16.59 to 19.14. The conclusion remained consistent: significant accent-reduction effects showed extreme variability. This stability across imputation strategies and study subsets confirms that the CV-based imputation does not introduce meaningful bias into the meta-analytic results.

Convergent evidence from MOS and WER

To provide complementary evidence that MOS-based accentedness ratings capture meaningful speech intelligibility differences, we examined the relationship between accent reduction effect sizes (Hedges' g) and WER for the four studies that report both metrics (Table 6). Because higher Hedges' g indicates greater accent reduction (desirable) while lower WER indicates higher intelligibility (also desirable), a concordant relationship between the two metrics manifests as a negative Pearson correlation. Indeed, we observed $r = -0.78$ and Spearman $\rho = -0.8$): systems that achieved greater perceptual accent reduction (higher g) also produced more intelligible speech (lower WER). For instance, KD-AC achieved the highest accent effect size ($g = 32.95$) and the lowest WER (12.4%), whereas AC-Articulatory had the lowest effect size ($g = 11.77$) and the highest WER (17.8%).

Although neither correlation reached statistical significance ($p = 0.225$ and $p = 0.2$, respectively) due to the limited overlap ($k = 4$), the strong and consistent direction across both correlation measures supports the validity of MOS-based accentedness ratings as a perceptual evaluation tool: the subjective judgments of human raters align with objective speech recognition performance. Future FAC evaluations should routinely report WER alongside MOS to strengthen cross-study comparability and quantify the practical impact of accent conversion on downstream intelligibility.

Computational cost analysis

In terms of computational efficiency, only limited information is available in the current FAC literature. For example, Streaming-NAR [61] reports an RTF of approximately 0.25 on a GTX 1060 GPU, with an E2E latency of roughly 0.8 s. To provide an additional reference point, we evaluated a publicly available token-based FAC system, TokAN [43], under controlled conditions. Using 560 L2-ARCTIC speech files (approximately 2,058 s of audio), we measured a mean RTF of 11.435 ± 1.712 during CPU inference on an Intel Core i7-9700K processor, and a mean RTF of 0.771 ± 0.229 during single-GPU inference on a Nvidia A40 GPU.

Key insights and implications

The accent–naturalness tradeoff is real but not fixed

Among the four studies that report both nine-point scale accentedness and five-point naturalness results, we observed a moderate negative correlation between accent gap closure and naturalness preservation ($r = -0.64$, Spearman $\rho = -0.6$). Although this did not reach statistical significance ($p = 0.359$), given the limited overlap ($k = 4$), it confirms the existence of a quality–accent tradeoff, and the direction is consistent: systems that more aggressively reduced accent tended to sacrifice more naturalness. ZS-NoRef achieved 80.1% gap closure but retained only 68.2% of L2 naturalness, whereas Accentron achieved 63% gap closure while preserving 93.5%. This indicates that the tradeoff boundary is likely shaped by architectural choices—specifically, how speech content, speaker identity, and prosodic features are separated during conversion—rather than by an inherent ceiling on joint performance. Future systems should be optimized and evaluated on both dimensions simultaneously, and reporting only one metric provides an incomplete picture of system capability.

Architectural diversity as opportunity

The extreme heterogeneity ($I^2 > 99\%$ for both dimensions) represents an opportunity rather than a limitation. Unlike mature fields that show incremental improvements toward optimal solutions, FAC exhibits wide performance variations across different systems. This indicates that 1) the field has not converged on optimal architectures, 2) breakthrough improvements remain possible through innovative design, and 3) current approaches may be exploring fundamentally different solution spaces (e.g., seq2seq, TTS guided, and token based) without a clear understanding of which will ultimately succeed.

Does accent conversion ceiling exist?

The highest observed gap closure (85.9%, TTS-guided-NoPar) demonstrates that near-native accent conversion is

Authorized licensed use limited to: EssilorLuxottica SA. Downloaded on October 06,2026 at 08:48:14 UTC from IEEE Xplore. Restrictions apply.

achievable. Yet no system reached 100%, pointing to a practical ceiling of multifactorial origin. At the perceptual level, native listeners do not evaluate accent based on a single cue; they simultaneously integrate segmental details, coarticulation, rhythm, and fine-grained prosodic patterns. Many current systems appear to capture these accent cues only partially. As a result, even small residual deviations in timing or spectral structure may remain perceptually noticeable when compared with an idealized native baseline. At the architectural level, incomplete disentanglement of linguistic content, speaker identity, and prosody, together with reliance on imperfect TTS teachers or synthetic training targets, may introduce systematic residual error that limits how closely converted speech can approach native norms. At the same time, the wide empirical range in gap closure (25.1%–85.9%) and the extreme heterogeneity ($I^2 > 99\%$) suggest that the main limitation at present is more likely architectural than fundamental or perceptual. In other words, current systems do not yet appear to have reached a theoretical perceptual limit. This interpretation is further supported by the fact that both TTS-guided and token-based methods achieve effect sizes of $g > 23$, substantially outperforming earlier seq2seq systems. It indicates that continued methodological advances may further increase gap closure. More specifically, improvements in representation granularity, explicit modeling in suprasegmental structure and prosody, and joint optimization of accent reduction and naturalness preservation are likely to be the most promising directions for surpassing the current 85.9% boundary.

Lack of computational cost reporting

Our computational cost analysis indicates that computational performance can vary substantially across FAC architectures and hardware platforms. However, systematic reporting of computational metrics is largely absent in the existing FAC literature. Most studies primarily evaluate perceptual quality or accent reduction, while runtime efficiency and deployment cost are rarely reported. This lack of standardized measurements makes it difficult to compare system efficiency or assess their practicality for real-world applications. These findings highlight the need for standardized computational benchmarking in future FAC research, including consistent reporting of RTF, latency, and hardware configurations.

Practitioner guidance: Method selection by application and constraint

The meta-analysis findings translate directly into deployment guidance. Table 7 organizes method selection across four representative application scenarios, mapping the system characteristics observed in this analysis—gap closure, naturalness preservation, WER, and computational class—to practical requirements.

Discussion

This article provides a comprehensive overview of FAC by systematically categorizing its methodologies, outlining standard FAC pipelines, and analyzing results from both subjective and objective evaluation metrics across different studies. By integrating these findings, the article provides a clear perspective on the current landscape of FAC research. In this work, we highlight significant advancements in FAC research, including seq2seq, TTS-guided, and token-based FAC approaches. However, critical challenges persist, limiting its full potential. The next sections highlight current limitations and challenges in FAC and outline promising directions for future research, providing insights for ongoing and future work.

Limitations and challenges in FAC research

Tradeoff among accent conversion, speaker identity, and speech naturalness

A central challenge in FAC lies in balancing effective accent conversion with the preservation of speaker identity and speech naturalness. Many existing systems degrade audio quality, resulting in lower MOSs than those of the original recordings. Some systems exhibit instability when synthesizing long utterances, and incomplete accent conversion is also observed in some shorter utterances. At the same time, efforts to enhance accent disentanglement often come at the cost of reducing speaker similarity. This tradeoff is especially pronounced for unseen speakers, for whom retaining voice identity while modifying accent characteristics remains a persistent challenge.

Table 7. Practitioner guidance: method selection by application and constraint.

Application	Primary Constraint	Recommended Method	Representative System	Empirical Basis
Real-time teleconferencing/ communication	Intelligibility, low latency	Token-based/ streaming TTS	KD-AC, Streaming-NAR	Lowest WER (12.4–14.1%); $g > 16$
Media dubbing/broadcast	Maximum naturalness preservation	S2S or diffusion-based	Accentron, †Diffusion-TTS	Highest L2 pres. (93.5%), mild degradation ($g = -1.11$); †diffusion-TTS optimized for high-fidelity speech synthesis
Zero-shot conversion	No parallel/reference data	TTS-guided	TTS-guided-NoPar, ZS-NoRef	Highest gap closure (80.1%–85.9%) without reference speech
Research benchmark	Maximum accent reduction	TTS-guided or token-based	KD-AC, TTS-guided-NoPar	Highest effect sizes (g = 29.51–32.95); best gap closure in pooled analysis

†: Diffusion-TTS excluded from formal meta-analysis due to its different evaluation protocol (best-worst scaling).

Authorized licensed use limited to: EssilorLuxottica SA. Downloaded on October 06,2026 at 08:48:14 UTC from IEEE Xplore. Restrictions apply.

Limited generalization to unseen accents

Another major limitation is poor generalization to unseen accents, with most FAC systems performing well only on accents included in their training data, such as those in the L2-ARCTIC dataset [64], which has only six accents. Zero-shot capabilities for unseen or underrepresented accents remain underexplored. This is exacerbated by narrow accent scopes, with many studies focusing on unidirectional conversions between limited accent pairs (e.g., Hindi to American English) and requiring retraining for each new pair, thereby reducing scalability and generalizability. The scarcity of large-scale, diverse accent corpora also significantly intensifies this challenge.

Computational overhead and slow inference

The long synthesis time for target accent speech remains a challenge as well, including high computational overhead from multistage training, complex architectures, and diffusion-based synthesis, resulting in slow inference that prevents real-time deployment. These delays arise from computationally intensive processes such as mel spectrogram generation and vocoder steps, creating a tradeoff among speed, model fidelity, and accuracy that is essential for interactive applications. Additionally, heavy reliance on pretrained models (e.g., HuBERT, Whisper, or speaker encoders) can introduce biases if their training data lack accent diversity, further undermining the model's overall robustness.

Lack of standardized evaluation metrics

Current FAC studies predominantly rely on subjective evaluation metrics, which are prone to human bias and can lead to less reliable and robust results. The lack of standardized scales across studies makes it difficult to consistently compare results as different scales for subjective evaluation are often used. Additionally, subjective evaluations also require extensive resources, including time, cost, and human effort, particularly when relying on participant recruitment through platforms like AMT. Existing objective evaluation methods often focus on isolated aspects of synthesized speech, without providing a comprehensive assessment of the entire utterance. For example, the visualization of speaker embeddings or accent embeddings from both original and synthesized utterances as well as the transcription accuracy (e.g., WER, character error rate, and so on). This highlights the need for more standardized, objective, and holistic evaluation methods that can accurately measure and compare both original and synthesized speech in FAC systems.

Potential ethical risks

Although FAC can improve intelligibility and support communication, its use without proper safeguards raises important ethical and societal concerns. FAC not only changes pronunciation, it can also affect how identity, trust, and fairness are perceived in spoken interaction. In effect, FAC sits at the intersection of VC and speech synthesis, and therefore inherits many of the same risks documented in the broader literature on synthetic speech and voice technologies. [69], [70].

Misuse for deception, fraud, and impersonation

A first concern is misuse for deception. Because FAC can make speech sound more native-like while preserving much of the original speaker's characteristics, it may make fraudulent speech more convincing in scams such as voice phishing. More broadly, work on speech generators and voice cloning shows that synthetic or modified speech can support impersonation, social engineering, and other forms of harmful deception when deployed without strong safeguards [69], [71]. In FAC, this risk is especially important because accent strongly shapes listeners' judgments of familiarity, credibility, and social belonging.

Privacy and unauthorized appropriation of voice identity

A second concern is privacy. Speech carries not only linguistic content but also biometric and socially sensitive information, including cues about identity, background, and social status. As a result, FAC may enable the unauthorized use of a person's voice characteristics or speaking style in ways that go beyond ordinary data misuse. Prior work on privacy in speech technology warns that speech systems can expose sensitive personal information and reduce users' control over how their voice is represented [72]. In this setting, FAC may lead to identity misuse, reputational harm, and loss of agency over one's vocal self-presentation.

Digital trust and transparency

A third concern is the effect of FAC on trust in voice-based systems. VC is already recognized as an important threat model for automatic speaker verification and antispoofing research because converted or synthetic speech can be used to attack speaker authentication systems [73]. If FAC becomes widely deployed in real-time communication without clear disclosure, it may also blur the line between communication support and deception. This can weaken trust not only in voice biometrics but also in the authenticity of everyday digital speech interactions.

Erosion of linguistic diversity

A fourth concern is the broader social meaning of accent modification. Accents are closely linked to personal, regional, and cultural identity, and they often reflect lived history and a sense of community belonging. Literature on accent discrimination shows that nonstandard or foreign accents are frequently judged unfairly in social and professional settings [74], [75], while recent work on artificial intelligence voice systems warns that speech technologies may reproduce linguistic privilege and digital exclusion across accent groups [70]. In this context, widespread use of FAC may unintentionally reinforce the idea that some accents are more acceptable than others. Over time, this may increase pressure on speakers to conform to dominant speech norms, weaken linguistic diversity in digital spaces, and shift the burden of adaptation onto speakers rather than institutions and listeners. Taken together, these concerns suggest that FAC should not be evaluated solely by intelligibility or accent-conversion quality.

Authorized licensed use limited to: EssilorLuxottica SA. Downloaded on October 06,2026 at 08:48:14 UTC from IEEE Xplore. Restrictions apply.

Responsible FAC research and deployment should also consider consent, disclosure, misuse prevention, privacy protection, robustness against spoofing, and the social impact of promoting accent normalization as a technical goal [69], [72].

Future trends and directions

FAC has advanced significantly, from traditional reference-based FAC methods to recent TTS-guided and token-based approaches. However, as noted in the previous section, some key challenges remain to enhance its robustness, efficiency, and adaptability. Summarizing the findings from the recent studies analyzed in this work, we identify several promising directions for future research to address these challenges.

Robust synthesis of long utterances

Current FAC systems often fail when synthesizing long, unseen utterances, limiting their utility in conversational applications [35], [38]. Future research should explore generative models, such as diffusion-based TTS [9], [41], which excel at modeling complex temporal dependencies for coherent speech generation. Alternative attention mechanisms, such as Gaussian mixture attention [35], could further enhance robustness by improving alignment across extended sequences. This direction is critical for enabling natural, sustained speech in real-world scenarios, such as virtual assistants or audiobooks [4], where long utterances are common.

Balancing accent conversion and speaker identity

Optimizing the tradeoff between accurate accent conversion and speaker identity preservation remains a core challenge, particularly for unseen speakers or limited datasets. Advanced feature disentanglement techniques, such as VAEs or adversarial training, can separate accent and speaker timbre more effectively [4], [12], [55], [62]. Diffusion-based TTS models, which have shown promise in high-fidelity speech synthesis, could further improve speaker similarity by modeling fine-grained voice characteristics [9]. This direction aligns with the growing demand for personalized speech interfaces, where users expect authentic accent rendering without compromising their unique vocal identity, as seen in applications such as language learning or voice cloning.

Real-time accent conversion with lightweight models

Real-time FAC is essential for interactive applications, such as live translation, online meetings, or a telephone service center. Reducing synthesis time requires lightweight models and nonautoregressive architectures, such as transformer-based or flow-based TTS systems, which help minimize computational overhead [11], [35]. Refinement of training stages or feature choices will also improve efficiency and performance [11]. Recent trends in LLMs that optimize inference through techniques such as model pruning and quantization offer a blueprint for developing efficient FAC systems. These advancements would enable low-latency conversion on resource-constrained devices, addressing real-world needs for seamless, on-the-fly accent adaptation in global communication platforms.

Developing standardized and objective perceptual evaluation metrics

Current FAC evaluation relies heavily on subjective human judgments of speech accentedness, naturalness, and voice identity, which are costly and inconsistent. Future work should develop automated, objective metrics using acoustic models or LLM-based embeddings to quantify accent accuracy, speech quality, and voice similarity [57]. For example, to provide scalable and standardized evaluation frameworks, pretrained LLMs could be fine-tuned to assess phonetic alignment or perceptual similarity [76]. This direction is vital for accelerating FAC development and ensuring consistent performance across diverse applications, from education to entertainment.

Generalization to diverse and unseen accents

Extending FAC frameworks to underrepresented and low-resource accents is crucial for inclusivity, particularly in multilingual societies. Techniques such as few-shot learning or cross-lingual transfer, inspired by LLM generalization, can adapt models to low-resource accents with minimal data [2], [11], [35]. Diffusion-based TTS models, which excel at capturing diverse data distributions, could further enhance generalization by modeling a wide range of accent variations. This direction enables applications such as accent coaching and computer-assisted language learning tools for nonnative speakers in underrepresented linguistic communities.

Multilingual accent conversion

Most current FAC research has concentrated on English, overlooking the vast diversity of other widely spoken languages such as Chinese, Spanish, and so on. Each language has its own dialects and accent variations. Extending FAC beyond English holds significant promise, not only for supporting language learners in achieving more native-like pronunciation across different languages but also for smoother global communication. Combined with ongoing advances in speech-to-speech and TTS translation, multilingual FAC could play an important role in breaking down linguistic barriers and enabling more natural cross-cultural interaction.

In conclusion, FAC has made substantial progress and continues to offer promising directions for both future research and real-world applications. In particular, generative models and LLM-inspired techniques appear promising for developing more robust, efficient, and inclusive FAC systems. Based on the meta-analysis, these future directions can be prioritized by both impact and feasibility. The most immediate priority is to better balance accent reduction with speech naturalness and speaker identity preservation as this remains the most common limitation across current systems and directly affects perceptibility. A second priority is the development of real-time FAC systems, especially the efficient, nonautoregressive, reference-free frameworks that are better suited to interactive applications but still require improvements in quality and robustness. A third priority is to make FAC more reliable with unseen accents and longer sentences as many current systems still rely on limited accent coverage and often degrade when speaking

Authorized licensed use limited to: EssilorLuxottica SA. Downloaded on October 06,2026 at 08:48:14 UTC from IEEE Xplore. Restrictions apply.

style or utterance length differs from training conditions. Progress along these directions would strengthen the practical value of FAC and support its wider use in entertainment, personalized language learning, and real-time communication.

Acknowledgment

This research was supported by the Israel Science Foundation (Grant 1449/23) and the Pazy Research Foundation. This article contains supplementary material available at https://doi.org/10.1109/MSP.2026.3703111.

Authors

Chenxi Liu (chenxi.liu@campus.technion.ac.il) received her B.Sc. degree in digital media technology from Dalian University of Technology, Dalian, China, in 2020, and her M.Sc. degree from the Andrew and Erna Viterbi Faculty of Electrical and Computer Engineering from the Technion—Israel Institute of Technology, Haifa, Israel, in 2023. She is currently a Ph.D. student in the Andrew and Erna Viterbi Faculty of Electrical and Computer Engineering at the Technion. Her research focuses on deep learning-based foreign accent conversion, with broader interests in deep learning, speech synthesis, and speech processing. Liu actively leverages her expertise in machine learning to explore innovative solutions for accent conversion and speech synthesis.

Israel Cohen (icohen@ee.technion.ac.il) received his B.Sc. (Summa Cum Laude), M.Sc., and Ph.D. degrees in electrical engineering from the Technion—Israel Institute of Technology, 3200003 Haifa, Israel, where he is the Louis and Samuel Seiden Professor in Electrical and Computer Engineering. He was an associate editor of *IEEE Transactions on Audio, Speech, and Language Processing* and *IEEE Signal Processing Letters*, a member of the IEEE Audio and Acoustic Signal Processing Technical Committee and the IEEE Speech and Language Processing Technical Committee, and a Distinguished Lecturer of the IEEE Signal Processing Society (SPS). He received the SPS Signal Processing Letters Best Paper Award and the Norman Seiden Prize for Academic Excellence. His research interests include speech enhancement, array signal processing, and deep learning for speech and audio applications. He is a Fellow of IEEE.

References

[1] Z. Qin, W. Zhao, X. Yu, and X. Sun, "OpenVoice: Versatile instant voice cloning," 2024, *arXiv:2312.01479*.

[2] T. N. Nguyen, N.-Q. Pham, and A. Waibel, "Accent conversion using pre-trained model and synthesized data from voice conversion," in *Proc. Interspeech*, 2022, pp. 2583–2587, doi: 10.21437/Interspeech.2022-10729.

[3] S. Liu, D. Wang, Y. Cao, L. Sun, X. Wu, S. Kang, et al., "End-to-end accent conversion without using native utterances," in *Proc. ICASSP 2020 - 2020 IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2020, pp. 6289–6293, doi: 10.1109/ICASSP40776.2020.9053797.

[4] X. Zhou, M. Zhang, Y. Zhou, Z. Wu, and H. Li, "Multi-scale accent modeling and disentangling for multi-speaker multi-accent TTS synthesis," 2024, *arXiv:2406.10844*.

[5] T. N. Nguyen, S. Akti, N. Q. Pham, and A. Waibel, "Improving pronunciation and accent conversion through knowledge distillation and synthetic ground-truth from native TTS," in *Proc. ICASSP 2025-2025 IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2025, pp. 1–5, doi: 10.1109/ICASSP49660.2025.10890229.

[6] J. Vaissière and P. B. De Mareüil, "Identifying a language or an accent: From segments to prosody," in *Proc. Workshop MIDL*, 2004, pp. 1–4.

[7] Krisp Technologies, Inc., "Krisp AI accent conversion." Krisp. 2026. Accessed: Mar. 16, 2026. [Online]. Available: https://krisp.ai/ai-accent-conversion/

[8] "Sanas: Real-time speech AI platform." 2026. sanas.ai. Accessed: May 19, 2026. [Online]. Available: https://www.sanas.ai/

[9] Q. Bai, S. Wang, Z. Liu, M. Zhang, W. Rao, Y. Wang, et al., "Diffusion-based method with TTS guidance for foreign accent conversion," in *Proc. IEEE 14th Int. Symp. Chin. Spoken Lang. Process. (ISCSLP)*, 2024, pp. 284–288.

[10] D. Jia, Q. Tian, K. Peng, J. Li, Y. Chen, M. Ma, et al., "Zero-shot accent conversion using pseudo Siamese disentanglement network," in *Proc. Interspeech*, Aug. 2023, pp. 5476–5480, doi: 10.21437/Interspeech.2023-612.

[11] X. Chen, J. Pei, L. Xue, and M. Zhang, "Transfer the linguistic representations from tts to accent conversion with non-parallel data," in *Proc. ICASSP 2024-2024 IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2024, pp. 12,501–12,505.

[12] J.-H. Huang, W.-T. Lee, and C.-H. Wu, "USD-AC: Unsupervised speech disentanglement for accent conversion," in *Proc. Interspeech*, 2024, pp. 4388–4392, doi: 10.21437/Interspeech.2024-664.

[13] T. Walczyna and Z. Piotrowski, "Overview of voice conversion methods based on deep learning," *Appl. Sci.*, vol. 13, no. 5, 2023, Art. no. 3100, doi: 10.3390/app13053100.

[14] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, et al., "FastSpeech 2: Fast and high-quality end-to-end text to speech," 2021, *arXiv:2006.04558*.

[15] A. Shrestha, S. Ghorshi, and I. Panahi, "Overview of accent conversion techniques based on deep learning," in *Proc. IEEE 6th Int. Conf. Image Process., Appl. Syst. (IPAS)*, 2025, pp. 1–6, doi: 10.1109/IPAS63548.2025.10924484.

[16] A. Shrestha, S. Ghorshi, and I. Panahi, "An overview of accent conversion techniques from statistical and deep learning: Its advantages and limitations," *Circuits Syst Signal Process*, vol. 45, pp. 3619–3674, 2026, doi: 10.1007/s00034-025-03346-9.

[17] B. Sisman, J. Yamagishi, S. King, and H. Li, "An overview of voice conversion and its challenges: From statistical modeling to deep learning," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 132–157, 2021, doi: 10.1109/TASLP.2020.3038524.

[18] R. Liu, B. Sisman, G. Gao, and H. Li, "Controllable accented text-to-speech synthesis with fine and coarse-grained intensity rendering," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 32, pp. 2188–2201, 2024, doi: 10.1109/TASLP.2024.3378110.

[19] X. Zhou, M. Zhang, Y. Zhou, Z. Wu, and H. Li, "Accented text-to-speech synthesis with limited data," 2023, *arXiv:2305.04816*.

[20] S. Aryal, D. Felps, and R. Gutierrez-Osuna, "Foreign accent conversion through voice morphing," in *Proc. Interspeech*, 2013, pp. 3077–3081, doi: 10.21437/Interspeech.2013-671.

[21] M. Huckvale and K. Yanagisawa, "Spoken language conversion with accent morphing," in *Proc. 6th ISCA Speech Synthesis Workshop*, 2007, pp. 64–70.

[22] D. Felps, C. Geng, and R. Gutierrez-Osuna, "Foreign accent conversion through concatenative synthesis in the articulatory domain," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 20, no. 8, pp. 2301–2312, Oct. 2012, doi: 10.1109/TASL.2012.2201474.

[23] S. Aryal and R. Gutierrez-Osuna, "Reduction of non-native accents through statistical parametric articulatory synthesis," *J. Acoust. Soc. Am.*, vol. 137, no. 1, pp. 433–446, Jan. 2015, doi: 10.1121/1.4904701.

[24] S. Aryal and R. Gutierrez-Osuna, "Data driven articulatory synthesis with deep neural networks," *Comput. Speech Lang.*, vol. 36, pp. 260–273, Mar. 2016, doi: 10.1016/j.csl.2015.02.003.

[25] S. Aryal and R. Gutierrez-Osuna, "Articulatory-based conversion of foreign accents with deep neural networks," in *Proc. Interspeech*, 2015, pp. 3385–3389, doi: 10.21437/Interspeech.2015-145.

[26] S. Aryal and R. Gutierrez-Osuna, "Can voice conversion be used to reduce non-native accents?" in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2014, pp. 7879–7883, doi: 10.1109/ICASSP.2014.6855134.

[27] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 8, pp. 2222–2235, Nov. 2007, doi: 10.1109/TASL.2007.907344.

[28] C. J. Cho, P. Wu, A. Mohamed, and G. K. Anumanchipalli, "Evidence of vocal tract articulation in self-supervised learning of speech," in *Proc. ICASSP 2023 - 2023 IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2023, pp. 1–5, doi: 10.1109/ICASSP49357.2023.10094711.

[29] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proc. 28th Int. Conf. Neural Inf. Process. Syst.*, vol. 2, 2014, pp. 3104–3112.

[30] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2014.

Authorized licensed use limited to: EssilorLuxottica SA. Downloaded on October 06,2026 at 08:48:14 UTC from IEEE Xplore. Restrictions apply.

[31] Y. Wang, R. J. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, et al., "Tacotron: Towards end-to-end speech synthesis," in *Proc. Interspeech*, 2017, pp. 4006–4010, doi: 10.21437/Interspeech.2017-1452.

[32] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, et al., "Natural TTS synthesis by conditioning wavenet on MEL spectrogram predictions," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2018, pp. 4779–4783, doi: 10.1109/ICASSP.2018.8461368.

[33] H. Miyoshi, Y. Saito, S. Takamichi, and H. Saruwatari, "Voice conversion using sequence-to-sequence learning of context posterior probabilities," in *Proc. Interspeech*, 2017, pp. 1268–1272, doi: 10.21437/Interspeech.2017-247.

[34] T. Kaneko, H. Kameoka, K. Hiramatsu, and K. Kashino, "Sequence-to-sequence voice conversion with similarity metric learned using generative adversarial networks," in *Proc. Interspeech*, 2017, pp. 1283–1287, doi: 10.21437/Interspeech.2017-970.

[35] W. Quamer, A. Das, J. Levis, E. Chukharev-Hudilainen, and R. Gutierrez-Osuna, "Zero-shot foreign accent conversion without a native reference," in *Proc. Interspeech*, 2022, pp. 4920–4924, doi: 10.21437/Interspeech.2022-10664.

[36] G. Zhao, S. Ding, and R. Gutierrez-Osuna, "Converting foreign accent speech without a reference," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 2367–2381, 2021, doi: 10.1109/TASLP.2021.3060813.

[37] G. Zhao and R. Gutierrez-Osuna, "Using phonetic posteriorgram based frame pairing for segmental accent conversion," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 27, no. 10, pp. 1649–1660, Oct. 2019, doi: 10.1109/TASLP.2019.2926754.

[38] S. Ding, G. Zhao, and R. Gutierrez-Osuna, "Accentron: Foreign accent conversion to arbitrary non-native speakers using zero-shot learning," *Comput. Speech Lang.*, vol. 72, Mar. 2022, Art. no. 101302, doi: 10.1016/j.csl.2021.101302.

[39] Y. Zhou, Z. Wu, M. Zhang, X. Tian, and H. Li, "TTS-guided training for accent conversion without parallel data," *IEEE Signal Process. Lett.*, vol. 30, pp. 533–537, 2023, doi: 10.1109/lsp.2023.3270079.

[40] J.-X. Zhang, Z.-H. Ling, and L.-R. Dai, "Non-parallel sequence-to-sequence voice conversion with disentangled linguistic and speaker representations," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 540–552, 2020, doi: 10.1109/TASLP.2019.2960721.

[41] K. Deja, G. Tinchev, M. Czarnowska, M. Cotescu, and J. Droppo, "Diffusion-based accent modelling in speech synthesis," in *Proc. Interspeech*, 2023, pp. 5516–5520, doi: 10.21437/Interspeech.2023-154.

[42] Z. Cheng, G. Zhang, Z. Tu, Y. Song, S. Mao, X. Jiao, et al., "SpeechAccentLLM: A unified framework for foreign accent conversion and text to speech," 2025, *arXiv:2507.01348*.

[43] Q. Bai, S. Inoue, S. Wang, Z. Jiang, Y. Wang, and H. Li, "Accent normalization using self-supervised discrete tokens with non-parallel data," 2025, *arXiv:2507.17735*.

[44] S. Inoue, S. Wang, W. Wang, P. Zhu, M. Bi, and H. Li, "MacST: Multi-accent speech synthesis via text transliteration for accent conversion," in *Proc. ICASSP 2025 - 2025 IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2025, pp. 1–5, doi: 10.1109/ICASSP49660.2025.10888195.

[45] Y. Halychanskyi, C. Churchwell, Y. Wen, and V. Kindratenko, "FAC-FACodec: Controllable zero-shot foreign accent conversion with factorized speech codec," 2025, *arXiv:2510.10785*.

[46] Y. Guo, Z. Li, H. Wang, B. Li, C. Shao, H. Zhang, et al., "Recent advances in discrete speech tokens: A review," 2025, *arXiv:2502.06490*.

[47] J. Melechovsky, A. Mehrish, B. Sisman, and D. Herremans, "DART: Disentanglement of accent and speaker representation in multispeaker text-to-speech," 2024, *arXiv:2410.13342*.

[48] J. Melechovsky, A. Mehrish, B. Sisman, and D. Herremans, "Accent conversion in text-to-speech using multi-level VAE and adversarial training," in *Proc. TENCON 2024-2024 IEEE Region 10 Conf. (TENCON)*, 2024, pp. 473–476.

[49] J. Melechovsky, A. Mehrish, D. Herremans, and B. Sisman, "Learning accent representation with multi-level VAE towards controllable speech synthesis," in *Proc. IEEE Spoken Lang. Technol. Workshop (SLT)*, 2023, pp. 928–935, doi: 10.1109/SLT54892.2023.10023072.

[50] D. Bouchacourt, R. Tomioka, and S. Nowozin, "Multi-level variational autoencoder: Learning disentangled representations from grouped observations," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 2095–2102, doi: 10.1609/aaai.v32i1.11867.

[51] Z. Wang, W. Ge, X. Wang, S. Yang, W. Gan, H. Chen, et al., "Accent and speaker disentanglement in many-to-many voice conversion," in *Proc. 12th Int. Symp. Chin. Spoken Lang. Process. (ISCSLP)*, 2021, pp. 1–5, doi: 10.1109/ISCSLP49672.2021.9362120.

[52] S. Kumar, J. Pradeep, and H. Zaidi, "Learning robust latent representations for controllable speech synthesis," 2021, *arXiv:2105.04458*.

[53] W. Quamer, A. Das, and R. Gutierrez-Osuna, "Decoupling segmental and prosodic cues of nonnative speech through vector quantization," in *Proc. Interspeech*, 2023, pp. 2083–2087, doi: 10.21437/Interspeech.2023-2202.

[54] V. Nechaev and S. Kosyakov, "Non-autoregressive real-time accent conversion model with voice cloning," 2024, *arXiv:2405.13162*.

[55] Y. M. Siriwardena, N. Swedlow, A. Howard, E. Gitterman, D. Darcy, C. Espy-Wilson, et al., "Accent conversion with articulatory representations," in *Proc. Interspeech*, 2024, pp. 4383–4387, doi: 10.21437/Interspeech.2024-1416.

[56] G. Zhao, S. Ding, and R. Gutierrez-Osuna, "Foreign accent conversion by synthesizing speech from phonetic posteriorgrams," in *Proc. Interspeech*, 2019, pp. 2843–2847, doi: 10.21437/Interspeech.2019-1778.

[57] M. Jin, P. Serai, J. Wu, A. Tjandra, V. Manohar, and Q. He, "Voice-preserving zero-shot multiple accent conversion," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2023, pp. 1–5, doi: 10.1109/ICASSP49357.2023.10094737.

[58] Z. Jia, H. Xue, X. Peng, and Y. Lu, "Convert and speak: Zero-shot accent conversion with minimum supervision," in *Proc. 32nd ACM Int. Conf. Multimedia*, ACM, 2024, pp. 4446–4454, doi: 10.1145/3664647.3681539.

[59] G. Zhao, S. Sonsaat, J. Levis, E. Chukharev-Hudilainen, and R. Gutierrez-Osuna, "Accent conversion using phonetic posteriorgrams," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2018, pp. 5314–5318, doi: 10.1109/ICASSP.2018.8462258.

[60] T. N. Nguyen, N. Q. Pham, and A. Waibel, "Accent conversion using discrete units with parallel data synthesized from controllable accented TTS," 2024, *arXiv:2410.03734*.

[61] T.-N. Nguyen, N.-Q. Pham, S. Akti, and A. Waibel, "Streaming non-autoregressive model for accent conversion and pronunciation improvement," 2025, *arXiv:2506.16580*.

[62] J. Melechovsky, A. Mehrish, B. Sisman, and D. Herremans, "Accented text-to-speech synthesis with a conditional variational autoencoder," in *Proc. 2024 IEEE Region 10 Conf. (TENCON)*, 2024, pp. 343–346, doi: 10.1109/TENCON61640.2024.10902981.

[63] J. Kominek and A. W. Black, "The CMU ARCTIC speech databases," in *Proc. 5th ISCA Workshop Speech Synthesis (SSW 5)*, 2004, pp. 223–224.

[64] G. Zhao, S. Sonsaat, A. Silpachai, I. Lucic, E. Chukharev-Hudilainen, J. Levis, et al., "L2-ARCTIC: A non-native English speech corpus," in *Proc. Interspeech*, 2018, pp. 2783–2787, doi: 10.21437/Interspeech.2018-1110.

[65] J. Yamagishi, C. Veaux, and K. MacDonald, *CSTR VCTK Corpus: English Multi-Speaker Corpus for CSTR Voice Cloning Toolkit (Version 0.92)*. Edinburgh, UK: Univ. of Edinburgh, The Centre for Speech Technol. Res. (CSTR), 2019.

[66] X. Shi, F. Yu, Y. Lu, Y. Liang, Q. Feng, D. Wang, et al., "The accented English speech recognition challenge 2020: Open datasets, tracks, baselines, results and methods," in *Proc. ICASSP 2021 - 2021 IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2021, pp. 6918–6922, doi: 10.1109/ICASSP39728.2021.9413386.

[67] A. Ahamad, A. Anand, and P. Bhargava, "AccentDB: A database of non-native English accents to assist neural speech recognition," in *Proc. Twelfth Lang. Resour. Eval. Conf., Eur. Lang. Resour. Assoc.*, 2020, pp. 5351–5358. [Online]. Available: https://aclanthology.org/2020.lrec-1.659/

[68] S. Weinberger, *Speech Accent Archive*. Fairfax, VA, USA: George Mason Univ., 2015. [Online]. Available: http://accent.gmu.edu

[69] W. Hutiri, O. Papakyriakopoulos, and A. Xiang, "Not my voice! a taxonomy of ethical and safety harms of speech generators," in *Proc. ACM Conf. Fairness, Accountability, Transparency (FAccT)*, Rio de Janeiro, Brazil: ACM, 2024, doi: 10.1145/3630106.3658911.

[70] S. Michel, S. Kaur, S. E. Gillespie, J. Gleason, C. Wilson, and A. Ghosh, "'it's not a representation of me': Examining accent bias and digital exclusion in synthetic AI voice services," in *Proc. ACM Conf. Fairness, Accountability, Transparency, (FAccT)*, ACM, 2025, doi: 10.1145/3715275.3732018.

[71] J. Barnett, "The ethical implications of generative audio models: A systematic literature review," in *Proc. AAAI/ACM Conf. AI, Ethics, Soc., (AIES)*, Montreal, QC, Canada: ACM, 2023, doi: 10.1145/3600211.3604686.

[72] A. Leschanowsky and S. Das, "Examining the interplay between privacy and fairness for speech processing: A review and perspective," in *Proc. 4th Symp. Secur. Privacy Speech Commun.*, 2024, pp. 1–11, doi: 10.21437/SPSC.2024-1.

[73] K. Kamel, K. Sood, H. S. Dutta, and S. Aryal, "A survey of threats against voice authentication and anti-spoofing systems," 2025, *arXiv:2508.16843*.

[74] T. Uzun, "Foreign accent, identity and accent discrimination: A literature review," *Int. J. Lang. Acad.*, vol. 11, no. 2, pp. 252–266, 2023, doi: 10.29228/ijla.69110.

[75] I. Hideg, W. Shen, and C. Zhou Koval, "Hear, hear! a review of accent discrimination at work," *Current Opinion Psychol.*, vol. 60, Dec. 2024, Art. no. 101906, doi: 10.1016/j.copsyc.2024.101906. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352250X24001192

[76] B. R. Chernyak, A. R. Bradlow, J. Keshet, and M. Goldrick, "A perceptual similarity space for speech based on self-supervised speech representations," *J. Acoust. Soc. Am.*, vol. 155, no. 6, pp. 3915–3929, Jun. 2024, doi: 10.1121/10.0026358.

Authorized licensed use limited to: EssilorLuxottica SA. Downloaded on October 06,2026 at 08:48:14 UTC from IEEE Xplore. Restrictions apply.