

A Survey of Automated Presentation Coaching: Systems, Methods, and Open Challenges

Wen Liang^{1,2}, Li Siyan¹, Zackary Rackauckas³, Julia Hirschberg¹

¹Columbia University, USA ²Red Hat, USA ³RoleGaku, USA

{wl2904, siyan.li, zcr2105}@columbia.edu, julia@cs.columbia.edu

Abstract

Automated coaching for oral presentations sits at the intersection of computer-assisted pronunciation training (CAPT), prosody modeling, and speech synthesis, yet no prior work has systematically surveyed and compared existing systems along these dimensions. This survey reviews and categorizes automated presentation coaching systems, spanning pronunciation tutors, fluency and prosody coaches, multimodal trainers, and conference Q&A practice tools. We introduce a five-dimensional task taxonomy—covering segmental pronunciation, lexical stress, suprasegmental prosody, pacing, and content faithfulness—and explicitly map surveyed systems onto it to reveal coverage gaps. We further review the core technical methods these systems employ: TTS-based exemplar generation and diagnostic methods for pronunciation, prosody, and fluency assessment. Key open challenges include the scarcity of annotated presentation corpora, achieving accent-fair feedback across diverse L1 backgrounds, and delivering low-latency diagnostics for real-time rehearsal.

1 Introduction

Oral presentations in English, including technical talks, research seminars, and product demos, impose demands well beyond everyday conversation. Presenters must articulate specialized terminology accurately, manage pacing relative to slide transitions, and modulate prosody so that discourse structure is clear to listeners. For English as a Second Language (ESL) speakers, these demands are compounded by the tendency to carry over sound patterns from one’s native language into English, accent-specific prosodic patterns, and limited opportunities for realistic rehearsal (Munro and Derwing, 1995; Derwing and Munro, 2005).

The last two decades have seen growing interest in *automated coaching* for oral presentations,

spurred by progress in computer-assisted pronunciation training (CAPT) (Witt and Young, 2000; Cucchiaroni et al., 2009), neural text-to-speech (TTS) (Chen et al., 2024; Ren et al., 2021), and multimodal analysis (Baltrušaitis et al., 2019). Systems such as Rhema (Tanveer et al., 2015), Mirror Mirror (Schneider et al., 2015), and recent LLM-augmented Q&A coaches (Aiba et al., 2024) each target different facets of presentation skill. Yet the field currently lacks a unified survey that (a) catalogs and compares these systems, (b) identifies which presentation dimensions they address, and (c) exposes the problems that remain.

This paper fills that gap. We survey automated coaching systems for oral presentations, focusing on speech-based dimensions: pronunciation, lexical stress, prosody, pacing, and content faithfulness. Our scope spans both L2-specific and general public-speaking tools, since many techniques transfer across these settings. We deliberately set aside visual and gesture coaching (Schneider et al., 2015; Damian et al., 2015) except when they are integrated with speech feedback, and we do not aim to cover the full literature on general-purpose ASR or spoken language assessment.

Contributions. Concretely, this survey: (1) introduces a five-dimensional task taxonomy for presentation coaching and maps surveyed systems to reveal coverage gaps; (2) systematically reviews and categorizes existing automated presentation coaching systems; (3) reviews the core technical methods – TTS-based exemplar generation and diagnostic approaches – that underpin these systems; and (4) discusses open challenges in corpora, accent fairness, real-time deployment, and the gap between research and industry practice.

Literature Search Strategy. To ensure comprehensive coverage, we conducted a systematic search across ACL Anthology, IEEE Xplore, ISCA Archive, Google Scholar, and Semantic Scholar using queries combining terms such as “presenta-

tion coaching,” “pronunciation training,” “CAPT,” “prosody assessment,” “speech fluency,” “TTS coaching,” and “L2 speaking.” We included peer-reviewed publications from 1997 to 2025 that directly address automated coaching or assessment of oral presentation or spoken language skills. We supplemented keyword-based search with citations from key papers (Witt and Young, 2000; Golonka et al., 2014; Aiba et al., 2024), following established survey methodology (Frederick Eneye et al., 2025). Studies focused exclusively on visual or gestural coaching without a speech component were excluded, as were general-purpose ASR or TTS papers. This process yielded 15 representative systems and approximately 50 supporting references spanning CAPT, prosody modeling, TTS, and educational technology.

Organization. Section 2 reviews foundational technologies. Section 3 presents the five-dimensional task taxonomy that organizes the remainder of the paper, along with the system inputs, outputs, and operating modes that connect the taxonomy to practical coaching workflows. Section 4 surveys existing systems and maps them onto the taxonomy. Section 5 reviews the core technical methods these systems employ. Section 6 covers datasets and evaluation. Section 7 discusses remaining open problems and future directions, including the gap between research and industry deployment. Section 8 concludes.

2 Background: Foundational Technologies

We briefly review the core technologies that underpin automated presentation coaching: pronunciation assessment, prosody analysis, shadowing pedagogy, and neural TTS. See detailed discussions of diagnostic and synthesis methods in Section 5.

Computer-Assisted Pronunciation Training (CAPT). CAPT systems localize segmental errors using Goodness of Pronunciation (GOP) scores (Witt and Young, 2000) or ASR confidence measures (Cucchiaroni et al., 2009). Earlier approaches relied on HMM forced alignment (Rabiner, 1989; Franco et al., 1997); modern systems leverage CTC (Graves et al., 2006; Cao et al., 2024) and self-supervised representations (Baevski et al., 2020; Hsu et al., 2021) for alignment-free mispronunciation detection (Xu et al., 2021; Gong et al., 2022). These methods form the diagnostic backbone of most pronunciation coaching systems.

Prosody Analysis. Effective presentations utilize prosody (intonation, phrasing, rhythm) to signal structure (Hirschberg, 2004). Listener-impression studies (Shoda et al., 2023) confirm that even small pitch and timing adjustments significantly affect perceived speaker competence. Assessment typically compares log- F_0 contours and duration patterns between learner and reference utterances (Sakoe and Chiba, 1978; Rosenberg, 2010).

Shadowing Pedagogy. Exemplar-guided shadowing (imitation) is well established for improving L2 pronunciation, prosody, and fluency (Hori, 2008; Kadota, 2019; Hamada, 2018). When mimicking short auditory demonstrations with explicit focus on speech rate and prominence, learners experience improvements in timing, stress, and intonation (Hsieh et al., 2013). This pedagogy directly motivates the use of TTS to generate controllable, adjustable exemplars for coaching at scale.

Neural Text-to-Speech (TTS). Recent non-autoregressive flow-matching TTS models (e.g., F5-TTS (Chen et al., 2024), Voicebox (Le et al., 2023), CosyVoice 2 (Du et al., 2024)) synthesize highly natural speech with real-time factors below 1. These systems offer fine-grained control over speaking rate, pause insertion, and emphasis, making them well-suited for generating coaching exemplars. Zero-shot style transfer from short enrollment clips enables personalized references (Jia et al., 2018; Casanova et al., 2022).

3 Presentation Coaching Taxonomy

Building on the foundational technologies reviewed above, we formalize a five-dimensional taxonomy for automated presentation coaching. The taxonomy organizes presentation skills by the nature of the feedback they require and the methods available to assess them, providing a systematic framework for comparing existing systems and identifying coverage gaps. Figure 1 illustrates the structure.

3.1 Taxonomy Dimensions

We define each dimension, its assessment methods, and the kind of feedback it yields; Section 4 maps the full set of surveyed systems onto these dimensions, and Table 1 provides the complete mapping.

Pronunciation (segmental) covers phone- and word-level correctness, including salient vowel/consonant contrasts and technical terminology. Assessment relies on GOP (Witt and Young, 2000),

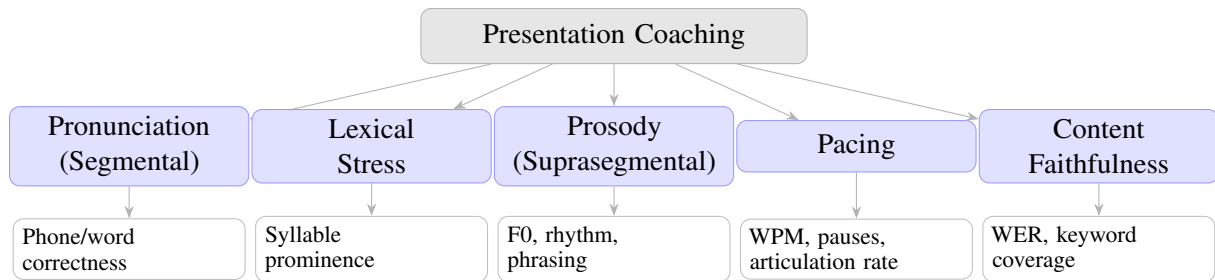


Figure 1: Five-dimensional taxonomy for automated presentation coaching. Existing systems (Table 1) cover different subsets; lexical stress and content faithfulness are the least-addressed dimensions.

CTC-based (Cao et al., 2024), or self-supervised methods (Xu et al., 2021); feedback for this dimension identifies top- k error phones. This is the most widely addressed dimension in the literature.

Lexical stress addresses syllable prominence in multi-syllabic words (e.g., *AL-go-rithm*). Feedback identifies the incorrect stress position and proposes re-stress drills. Despite its documented impact on comprehensibility (Munro and Derwing, 1995), this dimension is almost entirely neglected by existing systems (Korzekwa et al., 2021).

Prosody (suprasegmental) includes intonation (F0), phrasing, rhythm, and intensity contours that signal discourse structure. Feedback often includes F0 RMSE and Pearson r vs. a reference. Several systems address prosody with varying granularity.

Pacing covers words-per-minute (WPM), articulation rate, and pause placement relative to punctuation and slide boundaries. Feedback highlights WPM deviation and pause statistics.

Content faithfulness measures coverage of key content without insertions or omissions, flagging missing keywords per slide. Providing the ASR with a list of expected technical terms from the slides helps it correctly transcribe domain-specific vocabulary that a general-purpose model might otherwise miss. This is the least-addressed dimension in the literature.

3.2 Inputs, Outputs, and Assumptions

Beyond the five coaching dimensions, systems also differ in their input requirements and output granularity, with each input-output pattern tied to specific taxonomy dimensions. On the **input** side, systems targeting the *pronunciation* and *lexical stress* dimensions require, at minimum, a text transcript and a learner recording for forced alignment (Witt and Young, 2000; Neri et al., 2002; Strik et al., 2009; Xu et al., 2021). Systems addressing *prosody* and *pacing* additionally assume slide-aligned scripts

with section boundaries to enable discourse-level analysis (Hincks, 2005; Chen et al., 2014; Schneider et al., 2015). Systems targeting *content faithfulness* require the reference slide content for keyword matching. Systems that support personalized exemplars across any dimension may further accept a brief enrollment clip for voice cloning (Jia et al., 2018; Casanova et al., 2022).

On the **output** side, feedback granularity maps directly to the taxonomy. *Pronunciation* and *lexical stress* systems produce per-word or per-phone error flags (Witt and Young, 2000; Cao et al., 2024; Korzekwa et al., 2021). *Prosody* and *pacing* systems report global scores such as F0 deviation, articulation rate, and pause frequency (Shen et al., 2021; Saito et al., 2023). *Content faithfulness* systems flag missing keywords and WER per slide (Aiba et al., 2024). Multimodal trainers additionally output behavioral feedback on gaze and gesture (Schneider et al., 2015; Damian et al., 2015; Ramanarayanan et al., 2015). Across these systems, text normalization, grapheme-to-phoneme conversion, and forced or CTC-based alignment are standard preprocessing steps, while phonetic labels from the learner are generally not required.

3.3 System Modes

Given these inputs and outputs, coaching systems operate in three complementary modes, each serving different taxonomy dimensions. **Exemplar generation** (Section 5) produces controllable target reads per slide and for glossary items, directly supporting the *pronunciation*, *lexical stress*, and *prosody* dimensions by providing reference audio that models correct segment production, stress patterns, and intonation contours. **Comparative assessment** (Section 5) aligns learner audio with a reference to localize deviations, enabling diagnostics across all five taxonomy dimensions—from phone-level pronunciation errors to slide-level con-

tent omissions. A **mixed-initiative practice loop** turns diagnoses into focused drills targeting the weakest dimensions (e.g., re-stress a syllable for *lexical stress*; maintain a pre-equation pause for *spacing*), allowing the learner to re-record and track progress. This loop is implicit in several prior systems (Shen et al., 2021; Saito et al., 2023; Aiba et al., 2024) but rarely articulated explicitly for slide-based presentations. The next section surveys existing systems and maps them onto the taxonomy dimensions defined above.

4 Survey of Existing Presentation Coaching Systems

With the taxonomy in place, we now survey automated systems that target presentation or spoken English coaching. We organize them into four categories: CAPT-based pronunciation systems, prosody and fluency coaches, multimodal trainers, and Q&A/interaction coaches. Table 1 maps each system onto the five taxonomy dimensions.

4.1 CAPT-Based Pronunciation Systems

Early pronunciation coaching focused on segmental accuracy. Franco et al. (1997) pioneered automatic pronunciation scoring for language instruction using HMM-based forced alignment and log-likelihood scores, establishing the feasibility of system-generated feedback on individual phones. Neri et al. (2002) examined the pedagogy-technology interface in CAPT systems, showing that explicit error localization and corrective feedback improve learner outcomes beyond passive listening. Strik et al. (2009) compared GOP, posterior-based, and ASR-confidence approaches for automatic pronunciation error detection, finding that methods vary considerably across L1 backgrounds, motivating per-cohort calibration. More recently, Xu et al. (2021) demonstrated that wav2vec 2.0 representations substantially outperform HMM-based systems on L2 English mispronunciation detection benchmarks. Korzekwa et al. (2021) specifically addressed lexical stress detection in L2 English, using data augmentation and attention-based models to identify incorrect primary-stress placement in multi-syllabic words.

However, these systems share a common limitation: they operate on isolated words or short, read sentences, not on the semi-scripted, long-form speech characteristic of presentations. Feedback is typically provided post-hoc as a list of error flags,

with no practice loop or exemplar generation.

4.2 Fluency and Prosody Coaching

Moving beyond segmental accuracy, a second line of work targets fluency and suprasegmental quality. Hincks (2005) developed one of the earliest computer-aided systems specifically targeting spoken English for academic presentations, providing visual feedback on speaking rate and pitch variation. Shen et al. (2021) presented an interpretable model predicting L2 fluency from acoustic features including articulation rate, pause frequency, and phonation time, identifying the features that most strongly correlate with human ratings across diverse L1 backgrounds. Saito et al. (2023) conducted a comprehensive program on automated L2 comprehensibility assessment, arguing that comprehensibility, rather than narrow segmental accuracy, is a more robust target for coaching.

On the synthesis side, prior work (Onda et al., 2024, 2025) showed that generative spoken language models can simulate and modulate foreign-accented prosody, enabling the generation of corrected exemplars that retain the learner’s vocal identity while gaining similarity to target prosody patterns.

Zechner et al. (2009) developed automated scoring of non-native spontaneous speech for high-stakes testing (e-rater), covering fluency, pronunciation, and prosody holistically.

4.3 Multimodal Presentation Trainers

A third wave of systems extends speech coaching to the full presentation context, incorporating visual and behavioral signals together with audio. Schneider et al. (2015) developed *Mirror Mirror*, a multimodal presentation coaching system that provides automated feedback on body language, eye contact, and filler words in addition to speech rate. Damian et al. (2015) presented an intelligent tutoring system that analyzes posture, gesture, and vocal quality to generate structured coaching advice for novice presenters, grounded in public speaking pedagogy (Lucas, 2014). Ramanarayanan et al. (2015) assessed multimodal communication skills—including speech, gaze, and gesture—using behavioral signals extracted automatically, demonstrating feasibility for scalable deployment.

While the above systems combine multiple modalities, other work targets individual speech dimensions more directly. Chen et al. (2014) evaluated presentation skills in a tutoring context using

speech recognition and audience feedback. These systems demonstrate the feasibility of presentation-specific feedback loops, though they typically address only one or two of the dimensions in our defined taxonomy at a time.

4.4 Q&A and Interaction-Based Coaching

The most recent systems leverage large language models to extend coaching to the interactive components of presentations. Aiba et al. (2024) proposed a multimodal system for conference Q&A practice, combining ASR with ChatGPT-based question generation and TTS to simulate domain-relevant follow-up questions. Learners deliver a short talk; the system transcribes it, generates content-contingent questions, and reads them aloud, enabling repeated practice of spontaneous responses. Their user study reported reducing presentation anxiety and improving perceived preparedness among first-time conference participants.

Most recently, Chen et al. (2025) introduced PresentCoach, a dual-agent system in which one agent generates benchmark presentation videos from slides using personalized voice synthesis, while another evaluates the learner’s recording against these exemplars and delivers structured feedback. It is the first system to integrate slide-aware exemplar generation with multi-dimensional assessment in a single LLM-driven pipeline, covering four of the five taxonomy dimensions (pronunciation, prosody, pacing, and content faithfulness). Neural phonetic posteriorgrams (PPGs) (Churchwell et al., 2024) offer a complementary phone-level diagnostic suitable for such interactive loops.

4.5 Comparison and Discussion

Together, these four system categories cover a wide arc of presentation skills, yet no single system spans all. Table 1 compares the surveyed systems along the five coaching dimensions from our taxonomy (Section 3), whether the system provides real-time feedback, and whether it targets L2 speakers specifically.

Several patterns emerge from Table 1. First, **lexical stress is almost entirely neglected**: only Korzekwa et al. (2021) directly targets stress detection, despite its well-documented importance for L2 comprehensibility (Munro and Derwing, 1995). Second, **real-time feedback remains rare**: only two surveyed systems provide in-session feedback (Schneider et al., 2015; Aiba et al., 2024); most systems assess post-hoc. Third, **content faith-**

fulness and pacing are often decoupled: systems either check what was said (ASR-based) or how fast (WPM), but seldom do both with slide-aligned structure. Fourth, **no existing system covers all five dimensions simultaneously**: the most recent system, PresentCoach (Chen et al., 2025), covers four (pronunciation, prosody, pacing, and content) but still omits lexical stress, illustrating that integrated, multi-dimensional coaching remains an open challenge. These gaps are compounded by limitations in available data: most CAPT corpora (Zhao et al., 2018; Zhang et al., 2021; Garofolo et al., 1993) focus on isolated words or short sentences, lacking slide structure, domain terminology, and discourse-level features, while large-scale corpora such as LibriSpeech (Panayotov et al., 2015), Common Voice (Ardila et al., 2020), and TED-LIUM (Hernandez et al., 2018) are not designed for L2 coaching. In summary, existing systems address individual dimensions but do not combine exemplar-based practice with fine-grained diagnostics in a unified, presentation-specific workflow (Golonka et al., 2014).

5 Coaching Systems Methods

The systems surveyed in Section 4 draw on two complementary families of methods: **TTS-based exemplar generation**, which produces reference targets for learner practice, and **diagnostic techniques**, which assess how far a learner deviates from those targets. This section reviews both families of approaches employed by prior systems.

5.1 TTS-Based Exemplar Generation

Several surveyed systems use synthesized or recorded model speech as a reference for learner practice (Hincks, 2005; Schneider et al., 2015; Aiba et al., 2024). We review the TTS capabilities these systems exploit and the workflows they follow.

5.1.1 TTS Capabilities for Coaching

Modern neural TTS offers three capabilities that the surveyed systems leverage: **controllability** over rate, pauses, and emphasis; **code-switching** support for technical terms; and **zero-shot voice transfer** from short enrollment clips. Because recent flow-matching models (Chen et al., 2024; Le et al., 2023) and streaming architectures such as CosyVoice 2 (Du et al., 2024) achieve sub-second latency, exemplars can now be rendered on-the-fly during rehearsal. Two controls are most relevant:

| System | Pron. | Stress | Prosody | Pacing | Content | Real-time | L2-specific |
|-----------------------------|-------|--------|---------|--------|---------|-----------|-------------|
| Franco et al. (1997) | ✓ | | | | | | ✓ |
| Neri et al. (2002) | ✓ | | | | | | ✓ |
| Hincks (2005) | | | ✓ | ✓ | | | ✓ |
| Strik et al. (2009) | ✓ | | | | | | ✓ |
| Zechner et al. (2009) | ✓ | | ✓ | ✓ | | | ✓ |
| Chen et al. (2014) | ✓ | | ✓ | ✓ | ✓ | | |
| Damian et al. (2015) | | | ✓ | ✓ | | | |
| Schneider et al. (2015) | | | | ✓ | | ✓ | |
| Ramanarayanan et al. (2015) | ✓ | | ✓ | | | | |
| Shen et al. (2021) | ✓ | | ✓ | ✓ | | | ✓ |
| Korzekwa et al. (2021) | | ✓ | | | | | ✓ |
| Xu et al. (2021) | ✓ | | | | | | ✓ |
| Saito et al. (2023) | ✓ | | ✓ | | | | ✓ |
| Aiba et al. (2024) | | | | | ✓ | ✓ | ✓ |
| Chen et al. (2025) | ✓ | | ✓ | ✓ | ✓ | | |

Table 1: Comparison of existing presentation and spoken-language coaching systems across five taxonomy dimensions: “Pron.” = segmental pronunciation; “Stress” = lexical stress; “Content” = faithfulness/content coverage; “Real-time” = feedback provided during or immediately after recording. Checkmarks indicate that the system explicitly addresses the dimension.

tempo bands that constrain WPM, and *emphasis marks* that elicit pitch accents on key terms.

5.1.2 Workflow and Pedagogical Strategies

Generalizing patterns from prior systems (Hincks, 2005; Schneider et al., 2015) and grounded in shadowing pedagogy (Kadota, 2019; Hamada, 2018), a typical TTS-based coaching pipeline proceeds as follows: (1) render an *anchor* exemplar at a conservative tempo (120–140 WPM) and an optional *target* at a faster tempo (150–170 WPM); (2) record the learner in short chunks (5–12 s per section); (3) align learner audio with the reference using CTC (Graves et al., 2006) or DTW (Sakoe and Chiba, 1978) and compute deviations; (4) surface focused drills based on the highest-error dimensions. Across these systems, effective exemplars tend to be short (<12 s) to prevent cognitive overload and offer multiple speed bands and vocabulary difficulty gradations to accommodate different proficiency levels (Schneider et al., 2015; Golonka et al., 2014). Known limitations include over-constraining style when exemplars are treated as a single correct read, and synthetic emphasis that may not match domain conventions.

5.2 Pronunciation and Fluency Diagnostics

While TTS provides the coaching *targets*, diagnostic methods provide the *assessments* (Witt, 2012; Jurafsky and Martin, 2026). The surveyed systems employ three families of diagnostic techniques,

each addressing different taxonomy dimensions.

5.2.1 GOP and CTC-Based Assessment

Goodness of Pronunciation (GOP) (Witt and Young, 2000), used by several surveyed systems (Franco et al., 1997; Zechner et al., 2009; Strik et al., 2009), scores each phone by comparing how strongly the acoustic evidence supports the intended sound vs. the most likely alternative, averaged across the phone’s duration. A high GOP score indicates correct pronunciation; a low score flags a likely mispronunciation. Modern “segmentation-free” variants (Cao et al., 2024) derive boundaries from CTC posteriors (Graves et al., 2006) rather than forced alignment, improving robustness to disfluent L2 speech. Self-supervised models (wav2vec 2.0 (Baevski et al., 2020), HUBERT (Hsu et al., 2021), WavLM (Chen et al., 2022)) further advance this line (Xu et al., 2021; Gong et al., 2022), and neural PPGs (Churchwell et al., 2024) provide an alternative backbone for fine-grained phone-level scoring. Thresholds are calibrated per L1 on small rated sets, optimizing equal error rate or UAR against human annotations.

5.2.2 Clone-and-Compare (Personalized Reference)

To control for timbre and speaker style, several systems (Onda et al., 2024, 2025) synthesize an idealized rendition in the learner’s voice via voice conversion (Qian et al., 2019) or zero-shot TTS (Jia

| Corpus | Hrs | L2 | Long | Accent | Prosody |
|--|-------|----|------|--------|---------|
| TIMIT (Garofolo et al., 1993) | 5 | | | | |
| L2-ARCTIC (Zhao et al., 2018) | 26 | ✓ | | ✓ | |
| SpeechOcean762 (Zhang et al., 2021) | 70 | ✓ | | | ✓ |
| Speech Accent Archive (Weinberger, 2015) | - | ✓ | | ✓ | |
| LibriSpeech (Panayotov et al., 2015) | 960 | | | | |
| Common Voice (Ardila et al., 2020) | 1400+ | | | ✓ | |
| TED-LIUM 3 (Hernandez et al., 2018) | 452 | | ✓ | | ✓ |
| GigaSpeech (Chen et al., 2021) | 10k | | ✓ | | |
| EpaDB (Vidal et al., 2019) | 3 | ✓ | | | ✓ |
| Speak & Improve (Knill et al., 2024) | 340 | ✓ | | ✓ | |

Table 2: Major corpora used in presentation coaching research. “Long” = contains long-form or discourse-level speech; “Accent” = includes L1 or accent labels; “Prosody” = includes prosodic annotations. No existing corpus satisfies all four criteria simultaneously.

et al., 2018; Casanova et al., 2022), then compute MFCC/SSL (Davis and Mermelstein, 1980; Baevski et al., 2020) distance curves aligned with monotonic DTW (Sakoe and Chiba, 1978). Distance peaks flag likely mispronunciations while controlling for speaker-specific timbre. End-to-end ASR (Radford et al., 2023; Gulati et al., 2020) also yields word durations and confidence posteriors as complementary signals.

5.2.3 Prosody, Pacing, and Faithfulness

Prosodic quality is computed on $\log-F_0$ after voicing decisions (YAAPT (Zahorian and Hu, 2008)): per-slide RMSE and Pearson r capture intonation match, while word- or phrase-level duration RMSE captures rhythmic alignment. These suprasegmental metrics are used by several surveyed systems (Hincks, 2005; Shen et al., 2021; Saito et al., 2023). Pacing metrics (WPM deviation, articulation rate, pause precision/recall) appear in systems targeting real-time feedback (Schneider et al., 2015). For content faithfulness, constrained-LM ASR with glossary bonuses reduces false omissions on domain terms (Aiba et al., 2024); WER and missing-keyword flags per slide are the primary metrics. The integration of these diagnostic methods into a complete coaching pipeline is detailed in Appendix A.

6 Datasets and Evaluation

The diagnostic methods described above require annotated data for training and evaluation. Table 2 summarizes major corpora and assesses their suitability for presentation coaching according to whether the following are present in the corpus: whether the learners are L2 speakers, long-form speech characteristic of presentations, accent or L1 annotations, and prosody labels (e.g. pitch accent).

As Table 2 shows, no existing corpus simultaneously provides L2 speech, long-form presentation structure, accent labels, and prosodic annotations. TIMIT and L2-ARCTIC focus on isolated sentences; TED-LIUM 3 and GigaSpeech (Chen et al., 2021) are long-form but not L2-specific; Common Voice provides accent labels but consists of short read sentences. Newer resources partially close the gap: EpaDB (Vidal et al., 2019) offers detailed phone-level pronunciation annotations for L2 Spanish speakers of English, and the recently released Speak & Improve Corpus (Knill et al., 2024) provides approximately 340 hours of spontaneous L2 English speech with CEFR proficiency scores and diverse L1 backgrounds. However, neither includes slide structure or presentation-specific annotations. This gap directly limits the ability to train and evaluate systems on the dimensions most critical for presentation coaching, particularly discourse prosody and slide-aligned pacing. We recommend creating small, controlled *presentation mini-sets*, consisting of 5–10 minute talks per speaker with slide markers, emphasis annotations, and diverse accent coverage, as a near-term community benchmark.

To evaluate coaching systems using such data, we consolidate metrics across five categories (Table 3), each aligned with one or more taxonomy dimensions. Importantly, the individual metrics we recommend are not novel; they are established measures drawn from prior work in prosody analysis, fluency assessment, and speech quality evaluation. Our contribution lies in their systematic alignment with the five taxonomy dimensions to form a coherent evaluation framework for presentation coaching. This alignment ensures that each taxonomy dimension has clearly defined, measurable targets, enabling researchers to evaluate coaching systems comprehensively rather than on isolated aspects.

Segmental metrics assess phone- and word-level pronunciation accuracy. Phone/word F1 measures the precision and recall of correctly produced segments, while unweighted average recall (UAR) ensures that performance on rare error types is not masked by frequent correct phones. Both are derived from GOP (Witt and Young, 2000) or CTC-based (Cao et al., 2024) diagnostic outputs.

Prosody metrics capture how well a learner’s intonation matches the reference. $\log-F_0$ RMSE quantifies average pitch deviation, and Pearson r measures the correlation of pitch contours, reflecting whether the learner’s intonation follows the same rises and falls as the target (Rosenberg, 2010;

| Category | Key Metrics | Diagnostic Method |
|--------------|------------------------------|---------------------------------|
| Segmental | Phone/Word F1, UAR | GOP (Witt and Young, 2000), CTC |
| Prosody | Log- F_0 RMSE, Pearson r | Pitch track + DTW |
| Pacing | WPM deviation, Pause rate | Alignment boundaries |
| Faithfulness | WER, Glossary hit-rate | Constrained ASR |
| Perceptual | MOS, PESQ | Human / Neural pred. |

Table 3: Evaluation metrics for presentation coaching, organized by taxonomy dimension. We adopt individual metrics from prior work; we consolidate them into a unified framework aligned with the five coaching dimensions.

Zahorian and Hu, 2008).

Pacing metrics evaluate temporal control. WPM deviation measures how far the learner’s speaking rate falls from the target band, and pause rate captures whether pauses occur at appropriate boundaries (commas, slide breaks) rather than mid-phrase (Shen et al., 2021; Hincks, 2005).

Faithfulness metrics assess content coverage. WER measures overall transcription accuracy against the script (Radford et al., 2023), while glossary hit-rate tracks whether domain-critical terms are produced correctly, using constrained ASR with glossary bonuses (Aiba et al., 2024).

Perceptual metrics provide a holistic quality check. Mean Opinion Score (MOS) captures human judgments of overall speech quality, and PESQ (Rix et al., 2001) offers an automated proxy. Neural MOS predictors (Lo et al., 2019) can further scale these assessments without requiring human listeners for every evaluation.

Metrics alone are insufficient without evidence that they translate into actual learner improvement, and its underlying diagnostics have to remain reliable in realistic conditions. We therefore recommend three complementary validation protocols. First, **shadowing efficacy** should be tested via randomized crossover studies comparing TTS-guided (Chen et al., 2024) and unguided rehearsal, measuring whether exemplar-based practice yields measurable gains on the five taxonomy dimensions. Second, **diagnostic precision** should be evaluated by comparing GOP (Witt and Young, 2000) and clone-and-compare methods against expert annotations, since the choice of diagnostic directly affects the feedback a learner receives. Third, **robustness** should be assessed by measuring metric drift under environmental noise at multiple SNR levels, because real rehearsal settings rarely match studio-quality recording conditions.

7 Open Problems and Research Opportunities

The preceding sections reveal that while individual components of presentation coaching (pronunciation diagnostics, prosody modeling, exemplar generation) have matured considerably, integrating them into comprehensive, deployable systems remains an open challenge. Rather than restating these gaps, we identify five concrete research directions with specific actionable proposals.

Closing the lexical stress and discourse prosody gap. Lexical stress remains the most under-addressed taxonomy dimension (Table 1), yet mis-stressed words are a leading cause of reduced L2 comprehensibility (Munro and Derwing, 1995). A concrete next step is to extend self-supervised mispronunciation detectors (Xu et al., 2021; Gong et al., 2022) with syllable-level prominence classifiers trained on forced-aligned lexical stress annotations. For discourse prosody, current systems assess intonation at the utterance level, but presentations require section-level prosodic planning (e.g., pitch resets at topic boundaries, rising contours for rhetorical questions) (Hirschberg, 2004; Xu, 2005). We propose training prosodic planning models that condition on slide structure and discourse markers, enabling feedback such as “lower your pitch at the start of a new section.”

Real-time, personalized coaching. Only two of 15 surveyed systems provide real-time feedback (Table 1), yet immediate feedback is pedagogically more effective than post-hoc reports (Golonka et al., 2014). Sub-second diagnostics during live rehearsal require neural audio codecs (Défossez et al., 2022) and efficient alignment algorithms (e.g., sub-sequence DTW or streaming CTC). Beyond latency, systems should learn user-specific pacing bands and prosodic targets from minimal enrollment data rather than enforcing fixed norms (Trofimovich and Baker, 2006). A concrete proposal is to develop few-shot personalization modules that adapt pacing and prosody targets from a 2–3 minute calibration recording, enabling coaching that improves clarity without erasing the speaker’s natural style.

Evaluation infrastructure and benchmark creation. The field lacks standardized corpora with slide-aligned scripts, emphasis annotations, and diverse L1 coverage (Section 6). We recommend a community effort to create *presentation mini-sets*:

20–30 speakers from 5+ L1 backgrounds, each delivering a 5–10 minute technical talk with time-stamped slide boundaries, phone-level transcriptions, and expert prosody ratings. Such a benchmark would enable head-to-head comparison of coaching systems on all five taxonomy dimensions and support shared evaluation campaigns analogous to SUPERB (Yang et al., 2021) for speech processing. Equally important is validating that proposed metrics predict actual learner improvement, not just reference similarity; longitudinal studies measuring pre/post coaching gains on each taxonomy dimension remain rare and are critically needed.

LLM-augmented coaching and multimodal integration. The success of LLM-based coaching (Aiba et al., 2024; Chen et al., 2025) raises broader questions about how large language models can be integrated across all five taxonomy dimensions. Concrete opportunities include: (a) generating contextualized drill prompts (“Try saying *algorithm* with stress on the first syllable”) that target specific taxonomy dimensions identified by diagnostics; (b) evaluating content faithfulness through semantic similarity rather than word-level WER, enabling tolerance for paraphrasing while still catching substantive omissions; and (c) adapting coaching tone to individual learner profiles and proficiency levels. Recent work shows that speech LLMs can outperform prior baselines for L2 oral proficiency assessment (Ma et al., 2025), suggesting a path toward end-to-end systems that jointly assess pronunciation, prosody, and fluency. However, grounding LLM outputs in acoustically verified evidence, rather than generating plausible-sounding but inaccurate feedback, remains a key challenge. Future systems should also incorporate visual cues (slides, gestures) via vision-language models (Radford et al., 2021; Girdhar et al., 2023) to enable truly multimodal coaching. Extending coaching beyond English to other academic presentation languages (e.g., Mandarin-accented English, multilingual code-switching) requires corpora, phoneme sets, and G2P systems that do not yet exist at scale (Comneau et al., 2021).

Bridging the research and industry gap. Commercial L2 pronunciation and presentation coaching applications have grown rapidly. Platforms such as ELSA Speak and Speechling target segmental pronunciation, while Yoodli and Orai focus on pacing and filler-word reduction for general

public speaking. However, a significant gap exists between research prototypes and deployed systems. Commercial platforms typically address one or two taxonomy dimensions and rarely publish their technical approaches, making systematic comparison with research systems difficult. Furthermore, industry systems face practical constraints, including device heterogeneity, diverse noise environments, and the need for sub-second feedback on mobile hardware, that are rarely addressed in academic evaluations. We recommend that: (a) future surveys explicitly compare commercial and research systems where possible; (b) researchers engage with deployed platforms to identify real-world failure modes that controlled experiments may miss; and (c) standardized evaluation protocols be developed that span both research and commercial systems. The fragmentation between research (which advances methods) and industry (which addresses deployment constraints) could be reduced through shared benchmarks and open evaluation challenges.

8 Conclusion

We have surveyed automated presentation coaching systems, organizing them through a five-dimensional taxonomy (pronunciation, lexical stress, prosody, pacing, and content faithfulness) and mapping 15 representative systems onto this framework. Our comparison (Table 1) reveals that lexical stress and content faithfulness are dramatically under-addressed, that real-time feedback remains rare, and that no existing system integrates all five dimensions. We reviewed the core technical methods, including TTS-based exemplar generation and diagnostic approaches such as GOP/CTC, clone-and-compare, and prosody/pacing metrics, and showed how they relate to the taxonomy. Open challenges include the scarcity of annotated presentation corpora, accent-fair evaluation across diverse L1 backgrounds, the latency constraints of real-time coaching, and the gap between research prototypes and industry deployment. We hope this survey and its associated taxonomy serve as a useful reference for researchers developing the next generation of integrated, evidence-based presentation-coaching systems.

Limitations

This survey has several limitations that should be acknowledged. First, while we aim to cover representative work at the intersection of CAPT, TTS,

and L2 speaking systems, our coverage is not exhaustive; for space reasons we emphasize studies that directly inform pronunciation, prosody, pacing, and comprehensibility for slide-based L2 English presentations. Second, while we synthesize methods from TTS and CAPT domains, we focus primarily on English presentation training; the applicability to other languages and presentation styles (e.g., storytelling, persuasive speaking) requires further investigation. Third, our taxonomy and metrics are derived from existing literature and may not capture all dimensions relevant to real-world coaching scenarios. Fourth, the datasets we curate and recommend are limited in scale; large-scale presentation corpora with diverse accents and domains would strengthen evaluation. Fifth, we discuss voice cloning for personalized references but do not provide empirical comparisons of clone-and-compare vs. expert references across different speaker populations. Finally, while we address ethics and fairness, the practical implementation of accent-fair thresholds and privacy-preserving systems requires more extensive validation in deployed settings.

Ethics Statement

As automated coaching systems become more widely deployed, ethical concerns around fairness, privacy, and data governance are core design requirements. Many surveyed systems collect sensitive voice data and generate personalized feedback, making the following principles essential for trustworthy deployment.

Accent fairness. Calibrate thresholds per L1 cohort to account for systematic phonetic differences (Flege, 1995; Best and Tyler, 2007); report subgroup metrics (e.g., precision/recall by L1) to identify potential biases; and avoid penalizing identity-marking phonetic features (e.g., rhoticity patterns, vowel quality (Wells, 1982; Ladefoged and Johnson, 2015)) when intelligibility is unaffected (Munro and Derwing, 1995; Derwing and Munro, 2005). Systems should help learners improve clarity without erasing accent identity (Isaacs and Trofimovich, 2018).

Voice cloning and data governance. Obtain explicit consent before creating voice clones, clearly explaining how embeddings will be used and stored; encrypt voice embeddings both in transit and at rest; support deletion/expiry mechanisms;

and allow non-clone baselines (e.g., generic TTS or expert references) for users who opt out. For data governance, disclose TTS licenses, avoid retaining raw audio longer than needed (consider on-device processing), and clearly communicate data collection practices.

Prosodic perception. Listener-impression studies (Shoda et al., 2023) highlight that minor pitch/timing adjustments can change perceived competence and speaker traits. Ethical coaching systems should avoid reinforcing biased mappings between prosody, accent, and personas, presenting feedback as optional stylistic guidance rather than prescriptive judgments.

Acknowledgments

We thank Ziwei Gong for reviewing the paper and providing valuable feedback.

References

- Mayuko Aiba, Daisuke Saito, and Nobuaki Minematsu. 2024. A chatgpt-based oral q&a practice system for first-time student participants in international conferences. In *Proceedings of the 25th Annual Conference of the International Speech Communication Association (Interspeech 2024)*, Kos, Greece. ISCA.
- Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proc. LREC*, pages 4218–4222.
- Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems 33*.
- Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2019. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443.
- Catherine T. Best and Michael D. Tyler. 2007. Nonnative and second-language speech perception: Commonalities and complementarities. In *Language Experience in Second Language Speech Learning: In Honor of James Emil Flege*, pages 13–34. John Benjamins.
- Xinwei Cao, Zijian Fan, Torbjørn Svendsen, and Giampiero Salvi. 2024. A framework for phoneme-level pronunciation assessment using etc. In *Proc. Interspeech*.

- Edresson Casanova, Julian Weber, Christopher D Shulby, Arnaldo Candido Junior, Eren Gölge, and Moacir A Ponti. 2022. Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone. In *Proc. ICML*, pages 2709–2720.
- Guoguo Chen, Shuzhou Chai, Guanbo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, Mingjie Jin, Sanjeev Khudanpur, Shinji Watanabe, Shuaijiang Zhao, Wei Zou, Xiangang Li, Xuchen Yao, Yongqing Wang, Yujun Wang, Zhao You, and Zhiyong Yan. 2021. GigaSpeech: An evolving, multi-domain ASR corpus with 10,000 hours of transcribed audio. In *Interspeech 2021*, pages 3670–3674.
- Lei Chen, Gary Feng, Chee Wee Leong, Blair Lehman, Michelle Martin, Hannah Graf, Aaron Clauset, and Su-Youn Yoon. 2014. Automated evaluation of presentation skills using speech recognition and audience feedback. In *Proc. Workshop on Intelligent Tutoring Systems for Ill-Defined Domains*, pages 1–10.
- Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. 2022. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.
- Sirui Chen, Jinsong Zhou, Xinli Xu, Xiaoyu Yang, Litao Guo, and Ying-Cong Chen. 2025. Presentcoach: Dual-agent presentation coaching through exemplars and interactive feedback. *arXiv preprint arXiv:2511.15253*.
- Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, Jian Zhao, Kai Yu, and Xie Chen. 2024. F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching. *arXiv preprint arXiv:2410.06885*.
- Cameron Churchwell, Max Morrison, and Bryan Pardo. 2024. High-fidelity neural phonetic posteriorgrams. *arXiv preprint arXiv:2402.17735*.
- Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2021. Un-supervised cross-lingual representation learning for speech recognition. In *Interspeech 2021*, pages 2426–2430.
- Catia Cucchiaroni, Ambra Neri, and Helmer Strik. 2009. Oral proficiency training in dutch L2: The contribution of ASR-based corrective feedback. *Speech Communication*, 51(10):853–863.
- Ionut Damian, Chiew Seng Sean Tan, Tobias Baur, Johannes Schöning, Kris Luyten, and Elisabeth André. 2015. Logue: A real-time feedback system for nonverbal presentation skills. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, pages 1835–1840.
- Steven Davis and Paul Mermelstein. 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4):357–366.
- Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. 2022. High fidelity neural audio compression. *arXiv preprint arXiv:2210.13438*.
- Tracey M. Derwing and Murray J. Munro. 2005. Second language accent and pronunciation teaching: A research-based approach. *TESOL Quarterly*, 39(3):379–397.
- Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, Fan Yu, Huadai Liu, Zhengyan Sheng, Yue Gu, Chong Deng, Wen Wang, Shiliang Zhang, Zhijie Yan, and Jingren Zhou. 2024. Cosyvoice 2: Scalable streaming speech synthesis with large language models. *arXiv preprint arXiv:2412.10117*.
- James Emil Flege. 1995. Second language speech learning: Theory, findings, and problems. In Winifred Strange, editor, *Speech Perception and Linguistic Experience: Issues in Cross-Language Research*, pages 233–277. York Press.
- Horacio Franco, Leonardo Neumeyer, Yoon Kim, and Orith Ronen. 1997. Automatic pronunciation scoring for language instruction. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 2, pages 1471–1474.
- Tania Amanda Nkoyo Frederick Eneye, Chukwuebuka Fortunate Ijezue, Ahmad Imam Amjad, Maaz Amjad, Sabur Butt, and Gerardo Castañeda-Garza. 2025. Advances in auto-grading with large language models: A cross-disciplinary survey. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 477–498. Association for Computational Linguistics.
- John S. Garofolo, Lori F. Lamel, William M. Fisher, Jonathan G. Fiscus, David S. Pallett, and Nancy L. Dahlgren. 1993. TIMIT acoustic-phonetic continuous speech corpus. LDC93S1.
- Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Manat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. Imagebind: One embedding space to bind them all. In *Proc. CVPR*, pages 15180–15190.
- Ewa M. Golonka, Anita R. Bowles, Victor M. Frank, Dorna L. Richardson, and Suzanne Freynik. 2014. Technologies for foreign language learning: A review of technology types and their effectiveness. *Computer Assisted Language Learning*, 27(1):70–105.

- Yuan Gong, Ziyi Chen, Iek-Heng Chu, Peng Chang, and James Glass. 2022. Transformer-based multi-aspect multi-granularity non-native english speaker pronunciation assessment. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.
- Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *Proc. ICML*, pages 369–376.
- Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. Conformer: Convolution-augmented transformer for speech recognition. In *Interspeech 2020*, pages 5036–5040.
- Yo Hamada. 2018. Shadowing for pronunciation development: Haptic-shadowing and IPA-shadowing. *Journal of Asia TEFL*, 15(1):167–183.
- François Hernandez, Vincent Nguyen, Sahar Ghannay, Natalia Tomashenko, and Yannick Estève. 2018. Tedlium 3: Twice as much data and corpus repartition for experiments on speaker adaptation. In *Proc. SPECOM*, pages 198–208.
- Rebecca Hincks. 2005. *Computer Support for Learners of Spoken English*. Ph.D. thesis, KTH Royal Institute of Technology.
- Julia Hirschberg. 2004. Pragmatics and intonation. In Laurence R. Horn and Gregory Ward, editors, *The Handbook of Pragmatics*, pages 515–537. Blackwell.
- T. Hori. 2008. *Exploring Shadowing as a Method of English Pronunciation Training*. Ph.D. thesis, Kwansei Gakuin University, Nishinomiya, Japan.
- Chih-Chieh Hsieh, Damin Dong, and Hsien-Chin Wang. 2013. A preliminary study of applying shadowing technique to english intonation instruction. *Taiwan Journal of Linguistics*, 11(2):43–66.
- Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.
- Talia Isaacs and Pavel Trofimovich. 2018. Cognition, intelligibility, and the boundaries of phonology: A look at l2 speech. *Annual Review of Applied Linguistics*, 38:125–143.
- Ye Jia, Yu Zhang, Ron Weiss, Quan Wang, Jonathan Shen, Fei Ren, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, and Yonghui Wu. 2018. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. In *Proc. NeurIPS*, pages 4485–4495.
- Daniel Jurafsky and James H. Martin. 2026. *Speech and language processing*. Online manuscript, Jan. 6, 2026 release.
- Shuhei Kadota. 2019. *Shadowing as a Practice in Second Language Acquisition: Connecting Inputs and Outputs*. Routledge.
- Kate Knill, Diane Nicholls, Mark J.F. Gales, Mengjie Qian, and Pawel Strojnski. 2024. Speak & improve corpus 2025: An L2 english speech corpus for language assessment and feedback. *arXiv preprint arXiv:2412.11986*.
- Daniel Korzekwa, Roberto Barra-Chicote, Szymon Zaporowski, Grzegorz Beringer, Jaime Lorenzo-Trueba, Jasha Droppo, Thomas Drugman, and Bożena Kostek. 2021. Detection of lexical stress errors in non-native (L2) english with data augmentation and attention. *arXiv preprint arXiv:2012.14788*.
- Peter Ladefoged and Keith Johnson. 2015. *A Course in Phonetics*, 7th edition. Cengage Learning.
- Matthew Le, Apoorv Vyas, Bowen Shi, Brian Karrer, Leda Sari, Rashel Moritz, Mary Williamson, Vimal Manohar, Yossi Adi, Jay Mahadeokar, and Wei-Ning Hsu. 2023. Voicebox: Text-guided multilingual universal speech generation at scale. In *Advances in Neural Information Processing Systems* 36.
- Chen-Chou Lo, Szu-Wei Fu, Wen-Chin Huang, Xin Wang, Junichi Yamagishi, Yu Tsao, and Hsin-Min Wang. 2019. MOSNet: Deep learning-based objective assessment for voice conversion. In *Proc. Interspeech*, pages 1541–1545.
- Stephen E. Lucas. 2014. *The Art of Public Speaking*, 12th edition. McGraw-Hill.
- Rao Ma, Mengjie Qian, Siyuan Tang, Stefano Banno, Kate M. Knill, and Mark J.F. Gales. 2025. Assessment of L2 oral proficiency using speech large language models. In *Proc. Interspeech*.
- Sherwyn Morreale, Philip Backlund, and Leyla Sparks. 2014. Communication education and instructional communication: Genesis and evolution as fields of inquiry. *Communication Education*, 63(4):344–354.
- Murray J. Munro and Tracey M. Derwing. 1995. Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning*, 45(1):73–97.
- Ambra Neri, Catia Cucchiari, Helmer Strik, and Lou Boves. 2002. The pedagogy-technology interface in computer assisted pronunciation training. *Computer Assisted Language Learning*, 15(5):441–467.
- Kentaro Onda, Keisuke Imoto, Satoru Fukayama, Daisuke Saito, and Nobuaki Minematsu. 2025. Prosodically enhanced foreign accent simulation by discrete token-based resynthesis only with native speech corpora. In *Proc. Interspeech*.

- Kentaro Onda, Joonyong Park, Nobuaki Minematsu, and Daisuke Saito. 2024. [A pilot study of GSLM-based simulation of foreign accentuation only using native speech corpora](#). In *Interspeech 2024*, pages 3600–3604.
- Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An ASR corpus based on public domain audio books. In *Proc. ICASSP*, pages 5206–5210.
- Kaizhi Qian, Yang Zhang, Shiyu Chang, Xuesong Yang, and Mark Hasegawa-Johnson. 2019. Autovc: Zero-shot voice style transfer with only autoencoder loss. In *Proc. ICML*, pages 5210–5219.
- Lawrence R Rabiner. 1989. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proc. ICML*, pages 28492–28518.
- Vikram Ramanarayanan, Lei Chen, Chee Wee Leong, Gary Feng, and David Suendermann-Oeft. 2015. An analysis of time-aggregated and time-series features for scoring different aspects of multimodal presentation data. In *Proc. Interspeech*.
- Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2021. Fastspeech 2: Fast and high-quality end-to-end text to speech. In *Proc. ICLR*.
- Antony W Rix, John G Beerends, Michael P Hollier, and Andries P Hekstra. 2001. Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In *Proc. ICASSP*, volume 2, pages 749–752.
- Andrew Rosenberg. 2010. Autobi: A tool for automatic tobi annotation. In *Proc. Interspeech*, pages 146–149.
- Kazuya Saito, Konstantinos Macmillan, Magdalena Kachlicka, Takuya Kunihara, and Nobuaki Minematsu. 2023. Automated assessment of second language comprehensibility: Review, training, validation, and generalization studies. *Studies in Second Language Acquisition*, 45(1):234–263.
- Hiroaki Sakoe and Seibi Chiba. 1978. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1):43–49.
- Jan Schneider, Dirk Börner, Peter van Rosmalen, and Marcus Specht. 2015. Presentation trainer, your public speaking multimodal coach. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 539–546.
- Yang Shen, Ayano Yasukagawa, Daisuke Saito, Nobuaki Minematsu, and Kazuya Saito. 2021. [Optimized prediction of fluency of L2 english based on interpretable network using quantity of phonation and quality of pronunciation](#). In *2021 IEEE Spoken Language Technology Workshop (SLT)*.
- Chihiro Shoda, Yingxiang Gao, Yurun He, Nobuaki Minematsu, Noriko Nakanishi, and Daisuke Saito. 2023. [Learners’ prosodic control in the task of expressive storytelling and predicted native listeners’ impressions of the learners’ speech](#). In *9th Workshop on Speech and Language Technology in Education (SLaTE 2023)*, pages 46–50.
- Helmer Strik, Khiet Truong, Febe de Wet, and Catia Cucchiari. 2009. Comparing different approaches for automatic pronunciation error detection. *Speech Communication*, 51(10):845–852.
- M. Iftekhar Tanveer, Emy Lin, and Mohammed Ehsan Hoque. 2015. Rhema: A real-time in-situ intelligent interface to help people with public speaking. In *Proceedings of the 20th International Conference on Intelligent User Interfaces*, pages 286–295.
- Pavel Trofimovich and Wendy Baker. 2006. Learning second language suprasegmentals: Effect of 12 experience on prosody and fluency characteristics of 12 speech. *Studies in Second Language Acquisition*, 28(1):1–30.
- Jazmín Vidal, Luciana Ferrer, and Leonardo Brambilla. 2019. EpaDB: A database for development of pronunciation assessment systems. In *Proc. Interspeech*, pages 589–593.
- Steven H. Weinberger. 2015. Speech accent archive. *George Mason University*. <http://accent.gmu.edu>.
- John C. Wells. 1982. *Accents of English*. Cambridge University Press.
- Silke M. Witt. 2012. Automatic error detection in pronunciation training: Where we are and where we need to go. *Proc. International Symposium on Automatic Detection of Errors in Pronunciation Training (ISADEPT)*, pages 1–8.
- Silke M. Witt and Steve J. Young. 2000. Phonetic segment evaluation for automatic assessment of pronunciation quality. In *Proc. ICSLP*.
- Xiaoshuo Xu, Yueteng Kang, Songjun Cao, Binghuai Lin, and Long Ma. 2021. Explore wav2vec 2.0 for mispronunciation detection. In *Interspeech 2021*, pages 4428–4432.

Yi Xu. 2005. Speech melody as articulatorily implemented communicative functions. *Speech Communication*, 46(3-4):220–251.

Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhota, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Tzu-Hsien Huang, Wei-Cheng Tseng, Kottik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung-yi Lee. 2021. SUPERB: Speech processing universal performance benchmark. In *Interspeech 2021*, pages 1194–1198.

Stephen A. Zahorian and Hongbing Hu. 2008. A spectral/temporal method for robust fundamental frequency tracking. *The Journal of the Acoustical Society of America*, 123(6):4559–4571.

Klaus Zechner, Derrick Higgins, Xiaoming Xi, and David M. Williamson. 2009. Automatic scoring of non-native spontaneous speech in tests of spoken english. *Speech Communication*, 51(10):883–895.

Junbo Zhang, Zhiwen Zhang, Yongqing Wang, Zhiyong Yan, Qiong Song, Yukai Huang, Ke Li, Daniel Povey, and Yujun Wang. 2021. speechocean762: An open-source non-native english speech corpus for pronunciation assessment. In *Interspeech 2021*, pages 3710–3714.

Guanlong Zhao, Sinem Sonsaat, Alif Silpachai, Ivana Lucic, Evgeny Chukharev-Hudilainen, John Levis, and Ricardo Gutierrez-Osuna. 2018. L2-ARCTIC: A non-native english speech corpus. In *Interspeech 2018*, pages 2783–2787.

A System Architecture and UI Patterns

Translating the methods and metrics discussed above into a usable coaching system requires careful attention to architecture, latency, and user interface design (Hincks, 2005; Chen et al., 2014). This section outlines practical technical considerations drawing on principles from public speaking pedagogy (Lucas, 2014; Morreale et al., 2014) and computer-assisted language learning (Golonka et al., 2014).

A typical coaching pipeline consists of seven stages: (1) script ingest & glossary extraction; (2) exemplar synthesis (anchor/target bands, emphasis) using modern TTS systems (Chen et al., 2024); (3) user recording in short chunks; (4) alignment (CTC timing (Graves et al., 2006); DTW (Sakoe and Chiba, 1978) fallback); (5) diagnostics (segmental GOP (Witt and Young, 2000)/DTW + prosody/pacing); (6) drill generation; and (7) progress tracking with per-slide thresholds. This architecture generalizes patterns found in prior L2 speaking and presentation systems (Aiba et al., 2024; Shen et al.,

2021; Saito et al., 2023), clarifying the interaction between TTS, diagnostics, and UI components.

To support interactive rehearsal, systems must meet strict latency constraints, targeting sub-second feedback per chunk. This includes TTS rendering in under 200ms for 5–8s text (achievable with non-autoregressive models like F5-TTS (Chen et al., 2024)), alignment and metric computation within 300ms, and UI updates in under 200ms. Privacy-sensitive applications should prioritize on-device inference and defer computationally intensive voice cloning to an offline enrollment step with explicit consent.

Finally, effective coaching interfaces should visualize these metrics through multiple synchronized views: (1) **waveform overlay** with log- F_0 contours (extracted using robust pitch trackers (Zahorian and Hu, 2008)) for comparing learner and reference prosody; (2) **word-level heatmap** color-coded by pronunciation scores (GOP (Witt and Young, 2000) or DTW (Sakoe and Chiba, 1978) distance) to identify problematic segments; (3) **twin playheads** for scrubbing corresponding moments in user and reference audio; (4) **per-slide checklist** highlighting top issues (e.g., “stress on *algorithm*”); (5) **pacing gauge** showing band compliance; and (6) **one-click drills** that auto-loop on error spans for focused practice.