# Pronunciation Deviation Analysis Through Voice Cloning and Acoustic Comparison

Andrew Valdivia[1], YueMing Zhang[1], Hailu Xu[1], Amir Ghasemkhani[1], Xin Qin[1]

California State University Long Beach, Long Beach, CA, USA
{andrew.valdivia01,simon.zhang01}@student.csulb.edu
{hailu.xu,amir.ghasemkhani,xin.qin}@csulb.edu

**Abstract.** This paper presents a novel approach for detecting mispronunciations by analyzing deviations between a user's original speech and their voice-cloned counterpart with corrected pronunciation. We hypothesize that regions with maximal acoustic deviation between the original and cloned utterances indicate potential mispronunciations. Our method leverages recent advances in voice cloning to generate a synthetic version of the user's voice with proper pronunciation, then performs frame-by-frame comparisons to identify problematic segments. Experimental results demonstrate the effectiveness of this approach in pinpointing specific pronunciation errors without requiring pre-defined phonetic rules or extensive training data for each target language.

**Keywords:** pronunciation analysis · voice cloning, · Tspeech processing · language learning · acoustic comparison

## 1 Introduction

Computer-assisted language learning programs have been extensively researched and widely adopted to overcome deficiencies inherent in traditional English language learning resources. These tools enable ESL (English as a Second Language) learners and tutors to systematically address nuanced linguistic challenges, particularly those related to phonological differences between English phonemes and those present in a learner's primary language. In this context, Computer-Assisted Pronunciation Training (CAPT) has emerged as a crucial resource, offering learners and instructors accessible methods to practice and refine English pronunciation [1].

For such systems to be effective, they must robustly detect subtle mispronunciations and provide immediate, actionable feedback to learners. However, existing CAPT systems often underperform due to their reliance on generalized pronunciation models, which fail to account for individual variability in a learner's idiolect and accent characteristics. Standard reference models typically lack personalization, limiting their sensitivity to nuanced pronunciation errors [2,3]. Moreover, CAPT systems tend to overlook phonological transfer effects from learners' first languages (L1), which vary widely and require targeted modeling approaches [4] .

To address these limitations, our approach leverages personalized, synthetically generated voices, finely tuned to replicate each learner's unique vocal traits under ideal pronunciation conditions. We employ advanced voice cloning technologies, explicitly utilizing the ElevenLabs platform, renowned for its sophisticated deep-learning algorithms and realistic synthetic speech generation. ElevenLabs' state-of-the-art neural models and extensive training datasets enable precise replication of individual vocal nuances, including subtle variations in intonation, rhythm, and articulation. This personalized synthetic reference serves as a tailored benchmark, substantially enhancing mispronunciation detection and feedback mechanisms' sensitivity, accuracy, and real-time responsiveness.

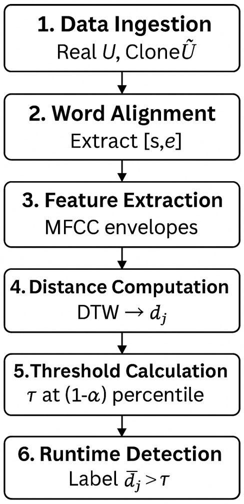

**Fig. 1.** Data-processing pipeline.

Recent research has shown that integrating speech synthesis into CAPT workflows enables large-scale generation of native-like references for training error detection models, overcoming the scarcity of annotated mispronounced data [4] . Furthermore, work by Das and Gutierrez-Osuna [5] demonstrated that combining text-to-speech (TTS) with speech reconstruction in a multi-task learning framework improves mispronunciation detection accuracy by modeling the mismatch between learner speech and native-like reconstruction. Nguyen et al. [6] extended this idea by synthesizing native-accented versions of non-native speech using knowledge distillation and TTS-based ground truth, allowing more effective accent and pronunciation correction.

Our experiments leverage the L2-ARCTIC corpus [7] a comprehensive dataset designed explicitly for voice conversion, accent modification, and mispronunciation detection research. The corpus includes 26,867 utterances from 24 non-native English speakers representing diverse linguistic backgrounds, such as Arabic, Chinese, Hindi, Korean, Spanish, and Vietnamese. Each speaker provided approximately one hour of phonetically balanced read speech, amounting to over 27 hours of recorded data. Critically, the dataset offers detailed annotations, comprising over 238,000 word segments and approximately 852,000 phone segments, with explicit identification of more than 14,000 phone substitutions, 3,400 deletions, and 1,000 additions. These meticulous annotations facilitate the robust development and evaluation of innovative pronunciation assessment tools.

Integrating personalized voice cloning with a comprehensive, rigorously annotated corpus like L2-ARCTIC enables our method to pinpoint problematic phonemes and subtle pronunciation errors reliably. Consequently, our approach offers targeted, precise feedback and fosters more effective pronunciation training interventions. This personalized feedback strategy also aligns with emerging trends in CAPT personalization, where learner-specific acoustic profiles, L1 influences, and synthetic benchmarking are used to enhance training efficacy [8], [9].

## 2 Related Work

Recent advances in voice cloning and neural TTS have reshaped pronunciation training. Modern systems leverage synthetic speech for high-variability phonetic training (Korzekwa et al., 2022), accent conversion (Zhou et al., 2022), and personalized feedback generation. Notably, Korzekwa et al. (2022) demonstrated that synthetic mispronunciations created through phoneme manipulation and voice cloning significantly improve error detection accuracy in CAPT systems, addressing data scarcity in non-native speech corpora. Their work aligns with the broader trend of using generative models to bypass dependency on annotated L2 data – a key enabler for our approach.

Traditional phoneme-level feedback methods rely on acoustic models like GOP scoring or end-to-end neural classifiers (Richter et al., 2023). While effective, these techniques require extensive training data and struggle with interpretability, often necessitating post-hoc transformations to map phonetic errors to orthographic representations. Recent innovations in voice identity preservation through neural voice conversion (Khaustova et al., 2023) suggest new possibilities for creating personalized pronunciation references, though prior work has not exploited acoustic deviation analysis between native and learner-specific voice clones.

Standard CAPT systems face three key limitations our method circumvents: (1) reliance on a large number of samples for model training, (2) rule-based error detection requiring phonetic expertise, and (3) generic feedback lacking voice personalization. By combining voice cloning with frame-level acoustic analysis, our approach eliminates the need for predefined phonetic rules while maintaining native pronunciation benchmarks in the learner's own vocal characteristics.

## 3 Proposed Method

Our mispronunciation detector runs in two phases—calibration and runtime—over six steps:

### 3.1 Data Preparation

Load real utterances $U = \{u_i\}$ with TextGrids and their TTS clones $\hat{U} = \{\hat{u}_i\}$.

### 3.2 Word Alignment

For each $(u_i, \hat{u}_i)$, extract word spans

$$\left\{ \left(w_j,\, [s_j^{\text{real}}, e_j^{\text{real}}],\, [s_j^{\text{clone}}, e_j^{\text{clone}}]\right) \right\}. \tag{1}$$

### 3.3 Feature Extraction

Compute 13-dimensional MFCC envelopes $E_j^{\text{real}}$, $E_j^{\text{clone}}$ over each span.

### 3.4 Distance Computation

Use DTW to get

$$d_j^{\text{DTW}} = \min_{\pi} \sum_t \left\| e_{j,t}^{\text{real}} - e_{j,\pi(t)}^{\text{clone}} \right\|, \quad \bar{d}_j = \frac{d_j^{\text{DTW}}}{\max_t \left| u_i(t) \right|}. \tag{2}$$

$$d_j^{\text{DTW}} = \min_{\pi} \sum_t \left\| e_{j,t}^{\text{real}} - e_{j,\pi(t)}^{\text{clone}} \right\|, \quad \bar{d}_j = \frac{d_j^{\text{DTW}}}{\max_t \left| u_i(t) \right|}. \tag{3}$$

### 3.5 Threshold Calculation

From all training distances split by true label, $\{\, \bar{d}_j : \text{correct}\}$ and $\{\, \bar{d}_j : \text{incorrect}\}$, compute class-specific thresholds:

$$\tau_C = \text{Percentile}\big(\{\bar{d}_j\}_{\text{correct}},\, 90\%\big), \quad \tau_I = \text{Percentile}\big(\{\bar{d}_j\}_{\text{incorrect}},\, 90\%\big). \tag{4}$$

### 3.6 Distribution Selection

For each new distance $\bar{d}_j$, compute its empirical CDF in each class:

$$p_C = F_{\text{correct}}(\bar{d}_j), \quad p_I = F_{\text{incorrect}}(\bar{d}_j). \tag{5}$$

Measure distance from the median (0.5):

$$\delta_C = \left| p_C - 0.5 \right|, \quad \delta_I = \left| p_I - 0.5 \right|. \tag{6}$$

Choose the distribution whose center it's closest to:

$$D^* = \begin{cases} \text{correct}, & \delta_C < \delta_I, \\ \text{incorrect}, & \text{otherwise.} \end{cases} \tag{7}$$

### 3.7 Runtime Detection

For a new utterance $u$: synthesize $\hat{u} = V(u)$, repeat steps 2–4 to get each $\bar{d}_j$, then:

- Select threshold $\tau^* = \begin{cases} \tau_C, & D^* = \text{correct}, \\ \tau_I, & D^* = \text{incorrect}, \end{cases}$
- Assign

$$\hat{y}_j = \begin{cases} \text{CORRECT}, & D^* = \text{correct} \ \wedge \ \bar{d}_j \le \tau_C, \\ \text{INCORRECT}, & D^* = \text{incorrect} \ \wedge \ \bar{d}_j \ge \tau_I, \\ \text{AMBIGUOUS}, & \text{otherwise.} \end{cases} \tag{8}$$

### 3.8 Algorithm

**Algorithm 1** Mispronunciation Detection via TTS-Clone Comparison

```
Require: Real utterances $U = \{u_i\}_{i=1}^{M}$ with TextGrids, base TTS model $V$, significance levels $\alpha_C, \alpha_I$
Ensure: Class-specific thresholds $\tau_C, \tau_I$ and mispronunciation labels for new utterances
Calibration phase:
for $i \leftarrow 1$ to $M$ do
    $\hat{u}_i \leftarrow V(u_i)$                                   ▷ Speaker-adapted TTS clone
    Extract word alignments $\{(w_j, [s_j^{\text{real}}, e_j^{\text{real}}], [s_j^{\text{clone}}, e_j^{\text{clone}}])\}_j$
    for each aligned token $w_j$ do
        Extract MFCC envelopes $E_j^{\text{real}}, E_j^{\text{clone}}$
        Compute DTW distance $d_j^{\text{DTW}} = \min_\pi \sum_t \|e_{j,t}^{\text{real}} - e_{j,\pi(t)}^{\text{clone}}\|$
        Normalize: $\bar{d}_j \leftarrow d_j^{\text{DTW}} \ / \ \max_t |u_i(t)|$
        Append $\bar{d}_j$ to $\mathcal{D}_{\text{correct}}$ or $\mathcal{D}_{\text{incorrect}}$ by true label
    end for
end for
Compute thresholds:
    $$\tau_C \leftarrow \text{Percentile}(\mathcal{D}_{\text{correct}},\ 1 - \alpha_C), \quad \tau_I \leftarrow \text{Percentile}(\mathcal{D}_{\text{incorrect}},\ 1 - \alpha_I)$$
Detection phase (new utterance $u$):
$\hat{u} \leftarrow V(u)$
Repeat alignment, feature extraction, distance computation to obtain $\{\bar{d}_j\}$
for each word $w_j$ with distance $\bar{d}_j$ do
    Compute empirical CDFs: $p_C \leftarrow F_{\text{correct}}(\bar{d}_j)$, $p_I \leftarrow F_{\text{incorrect}}(\bar{d}_j)$
    Compute median-distances: $\delta_C \leftarrow |p_C - 0.5|$, $\delta_I \leftarrow |p_I - 0.5|$
    Select distribution:
        $$D^* \leftarrow \begin{cases} \text{correct}, & \delta_C < \delta_I, \\ \text{incorrect}, & \text{otherwise.} \end{cases}$$
    Set threshold $\tau^* \leftarrow \begin{cases} \tau_C, & D^* = \text{correct}, \\ \tau_I, & D^* = \text{incorrect} \end{cases}$
    Assign label:
        $$\hat{y}_j \leftarrow \begin{cases} \text{CORRECT}, & D^* = \text{correct} \ \wedge \ \bar{d}_j \leq \tau_C, \\ \text{INCORRECT}, & D^* = \text{incorrect} \ \wedge \ \bar{d}_j \geq \tau_I, \\ \text{AMBIGUOUS}, & \text{otherwise.} \end{cases}$$
end for
```

## 4 Experiments

From the L2-Arctic dataset, our model was applied to four individuals: EBVS, ERMS, MBMPS, and NJS. Classification performance was measured using precision, recall, F1-score, and accuracy.

| Dataset | Class | Precision | Recall | F1-score | Support | Accuracy |
|---|---|---|---|---|---|---|
| EBVS | 0 | 0.615 | 0.582 | 0.598 | 110 | 0.646 |
| | 1 | 0.669 | 0.699 | 0.684 | 133 | |
| ERMS | 0 | 0.581 | 0.541 | 0.560 | 133 | 0.537 |
| | 1 | 0.492 | 0.532 | 0.511 | 111 | |
| MBMPS | 0 | 0.604 | 0.416 | 0.492 | 154 | 0.461 |
| | 1 | 0.353 | 0.538 | 0.426 | 91 | |
| NJS | 0 | 0.634 | 0.479 | 0.545 | 163 | 0.486 |
| | 1 | 0.346 | 0.500 | 0.409 | 90 | |

**Table 1.** Classification performance of the proposed model on each dataset.

### 4.1 Data Distribution Summary

Table 4.1 summarizes per-individual distance distribution (mean, median, standard deviation, 25th and 75th percentiles) for correctly and incorrectly classified examples.

| Dataset | Outcome | Mean | Median | Std | 25% | 75% |
|---|---|---|---|---|---|---|
| EBVS | Correct | 0.3467 | 0.3343 | 0.0981 | 0.2694 | 0.4114 |
| | Incorrect | 0.3785 | 0.3795 | 0.0928 | 0.3130 | 0.4378 |
| ERMS | Correct | 0.3649 | 0.3575 | 0.1030 | 0.2903 | 0.4289 |
| | Incorrect | 0.3952 | 0.3852 | 0.0971 | 0.3261 | 0.4601 |
| MBMPS | Correct | 0.3252 | 0.3152 | 0.0936 | 0.2547 | 0.3809 |
| | Incorrect | 0.3379 | 0.3301 | 0.0907 | 0.2751 | 0.3957 |
| NJS | Correct | 0.3693 | 0.3551 | 0.1053 | 0.2936 | 0.4349 |
| | Incorrect | 0.3971 | 0.3913 | 0.0960 | 0.3270 | 0.4608 |

**Table 2.** Summary of distances across datasets for correct vs. incorrect examples.

This set of results demonstrates that our model performs best on EBVS in terms of both accuracy and confidence separation, with varying calibration across the other datasets.

## 5 Results and Discussion

To investigate explainability in our model, we employ the NJS speech sample `arctic_a0028.mp2` to illustrate the system's capacity for resolving an issue that is similar to the caught–cot vowel merger. This phenomenon—stemming from first-language phonological interference—results in non-native speakers conflating two distinct English low-back vowels. English, by contrast, maintains approximately twenty contrastive vowel phonemes, including the low back unrounded (/A/) and low back rounded (/O/) categories.

The utterance under examination, *"Robbery, Bribery, Fraud,"* is presented in Figure 2 with its corresponding phonetic transcription. We extract the relevant

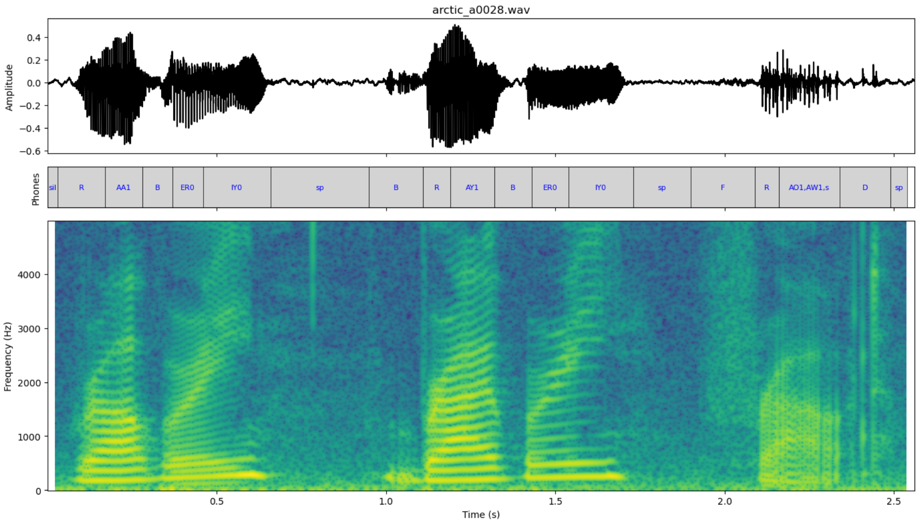

**Fig. 2.** Phonetic transcription of "Robbery, Bribery, Fraud."

temporal segments from both the original recording and the synthesized audio; these segments are depicted in Figure 4. Subsequent alignment is performed via Dynamic Time Warping (DTW), as demonstrated in Figure 3.

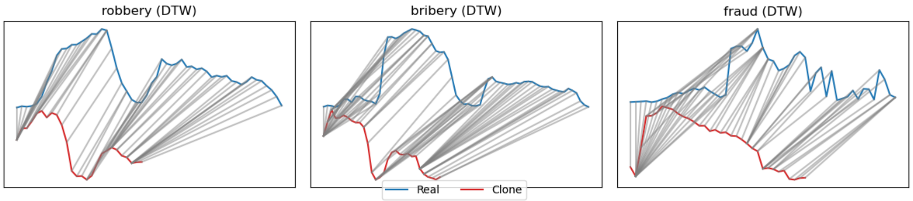

**Fig. 3.** DTW alignment between original and synthesized samples.

Focusing on the final word *"Fraud,"* the phonetic sequence can be analyzed as follows:

- /f/ (voiceless labiodental fricative)
- /r/ (voiced postalveolar approximant)
- **Vowel:** original sample (incorrectly realized as /AW1/), whereas the synthesized output correctly produces /AO1/
- /d/ (voiced alveolar plosive)

Notably, the original speaker's production employs /AW1/, indicative of the merger, while our model distinguishes the difference with /AO1/. This contrast

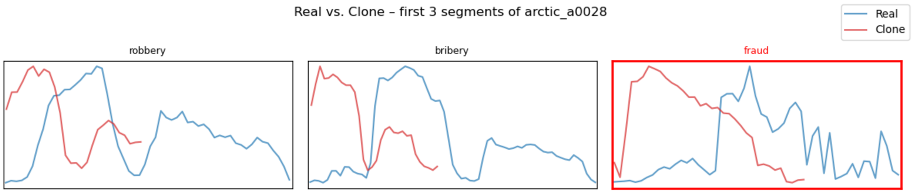

**Fig. 4.** Temporal segments from the original and synthesized signals. Mispronounced word "fraud" detected.

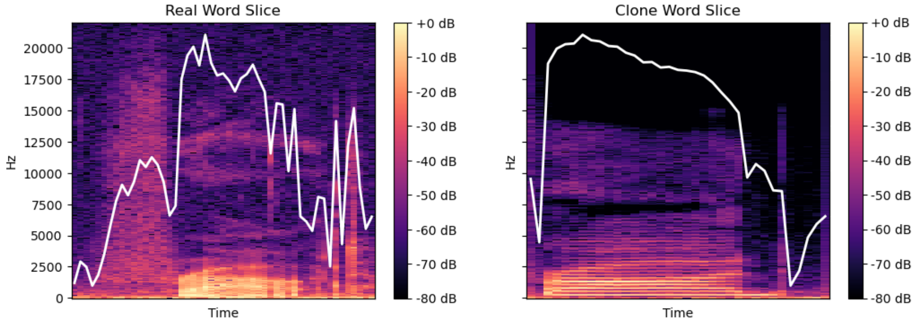

**Fig. 5.** STFT spectrograms of a real versus cloned word slice (color = –80 dB to 0 dB), with white curves showing frame-wise mean magnitude.

substantiates the model's ability to mitigate L1 interference effects by reinstating the phonetic distinction lost in non-native renditions.

## 6 Conclusion and Future Work

This paper introduces a novel paradigm for pronunciation error detection through voice cloning and differential acoustic analysis, demonstrating robust performance in identifying mispronunciations across diverse linguistic backgrounds. Our experiments validate the core hypothesis that deviations between original and pronunciation-corrected cloned speech correlate strongly with human-perceived errors. While the current framework operates effectively without language-specific rules, future work will integrate linguistic knowledge to enhance error type classification (e.g., distinguishing phonemic substitutions from prosodic errors) and pedagogical feedback. Further developments will focus on real-time implementation for interactive language learning applications and expansion to under-resourced languages, addressing broader cross-lingual pronunciation challenges. By bridging voice cloning technology with language pedagogy, this approach establishes a scalable foundation for adaptive pronunciation training systems that preserve speaker identity while targeting acoustic-phonetic improvements.

## References

1. M. Amrate and P.-H. Tsai, “Computer-assisted pronunciation training: A systematic review,” *ReCALL*, 2024, published online Sep. 2024.
2. D. Korzekwa, J. L. Trueba, T. Drugman, and B. Kostek, “Computer-assisted pronunciation training—speech synthesis is almost all you need,” *Speech Communication*, vol. 142, pp. 22–33, 2022.
3. C. Richter, R. Pálsson, L. O’Brien, K. Friðriksdóttir, B. Bédi, E. H. Magnúsdóttir, and J. Guðnason, “Orthography-based pronunciation scoring for better capt feedback,” in *Proc. Interspeech 2023*, 2023, pp. 1004–1008.
4. D. Korzekwa, J. Lorenzo-Trueba, T. Drugman, and B. Kostek, “Computer-assisted pronunciation training—speech synthesis is almost all you need,” *Speech Communication*, vol. 142, p. 22–33, Jul. 2022. [Online]. Available: http://dx.doi.org/10.1016/j.specom.2022.06.003
5. A. Das and R. Gutierrez-Osuna, “Improving mispronunciation detection using speech reconstruction,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 4420–4433, 2024.
6. 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 *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.
7. G. Zhao, S. Sonsaat, A. Silpachai, I. Lucić, E. Chukharev-Hudilainen, J. Levis, and R. Gutierrez-Osuna, “L2-arctic: A non-native english speech corpus,” in *Proc. Interspeech 2018*, 2018, pp. 2783–2787.
8. V. Khaustova, E. Pyshkin, V. Khaustov, J. Blake, and N. Bogach, “Capturing accents: An approach to personalize pronunciation training for learners with different l1 backgrounds,” in *Speech and Computer (SPECOM 2023)*, ser. Lecture Notes in Computer Science, vol. 14339, 2023, pp. 59–70.
9. Y. Zhou, Z. Wu, M. Zhang, X. Tian, and H. Li, “Accent conversion without parallel data using pretrained tts models,” *arXiv preprint arXiv:2212.10204*, 2022.