Compositional Phoneme Approximation for L1-Grounded L2 Pronunciation Training

Jisang Park[1*] **Minu Kim**[2*] **DaYoung Hong**[3] **Jongha Lee**[3†]
[1]Stanford University [2]KAIST [3]Independent Researchers
jisangp@stanford.edu, minus@kaist.ac.kr, {dayoung.hong, jongha.lee}@posthangul.com

Abstract

Learners of a second language (L2) often map non-native phonemes to similar native-language (L1) phonemes, making conventional L2-focused training slow and effortful. To address this, we propose an L1-grounded pronunciation training method based on compositional phoneme approximation (CPA), a feature-based representation technique that approximates L2 sounds with sequences of L1 phonemes. Evaluations with 20 Korean non-native English speakers show that CPA-based training achieves a 76% in-box formant rate in acoustic analysis, 17.6% relative improvement in phoneme recognition accuracy, and over 80% of speech being rated as more native-like, with minimal training. Project page: https://gsanpark.github.io/CPA-Pronunciation.

1 Introduction

This paper explores how leveraging a learner's native-language (L1) phonological system can guide the acquisition of non-native L2 phonemes. Learners often substitute such L2 phonemes with the closest yet non-interchangeable L1 phonemes, resulting in pronunciation errors (Kartushina and Frauenfelder, 2014; Shi et al., 2019; Wayland, 2021). This phenomenon undermines the effectiveness of conventional pronunciation training methods (Grimaldi et al., 2014), such as audiovisual mimicry of L2 speech (Espinoza et al., 2021; Galimberti et al., 2023; González and Ferreiro, 2024) and explicit phonological instruction (Karhila et al., 2019; Awadh et al., 2024). These approaches focus solely on the L2 target and overlook the learner's L1 background, often resulting in time-intensive training requirements.

Foundational theories have shown that L1 background strongly shapes how learners perceive L2 phonemes (Best et al., 1994; Flege, 1995; Flege et al., 2021). Empirical studies further demonstrate that when an L2 sound is perceptually assimilated to an existing L1 category, this perceptual confusion is mirrored in production, resulting in systematic substitutions and a noticeable foreign accent (Flege, 1993; Baker and Trofimovich, 2005). Consequently, many pronunciation training approaches have emphasized making L1–L2 phonetic contrasts more salient in order to counteract these overlaps.

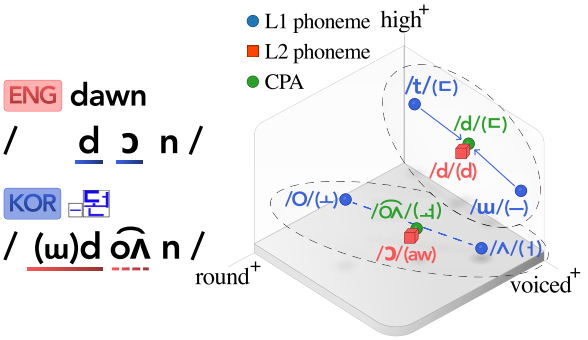

Figure 1: Compositional phoneme approximation represents L2 phonemes absent in the learner's L1 as composite sounds derived from multiple L1 phonemes.

Common approaches include contrasting L1 and L2 sound pairs to highlight phonological distinctions (Carey et al., 2015; Leppik et al., 2022), integrating signal processing techniques such as foreign accent conversion (Felps et al., 2009) and L1-adaptive automatic speech recognition (ASR) (Arora et al., 2018; Khaustova et al., 2023), and leveraging L1-specific error corpora (Husby et al., 2011) to provide personalized computer-assisted pronunciation training (CAPT). However, these approaches primarily focus on drawing attention to the differences between L1 and L2 sounds, rather than leveraging the learner's existing L1 articulatory knowledge as a resource to support the acquisition of unfamiliar L2 phonemes.

To this end, we propose an L1-grounded pronunciation training method that leverages **compo-**

*These authors contributed equally.
†Corresponding author.

Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 434–443
December 20-24, 2025

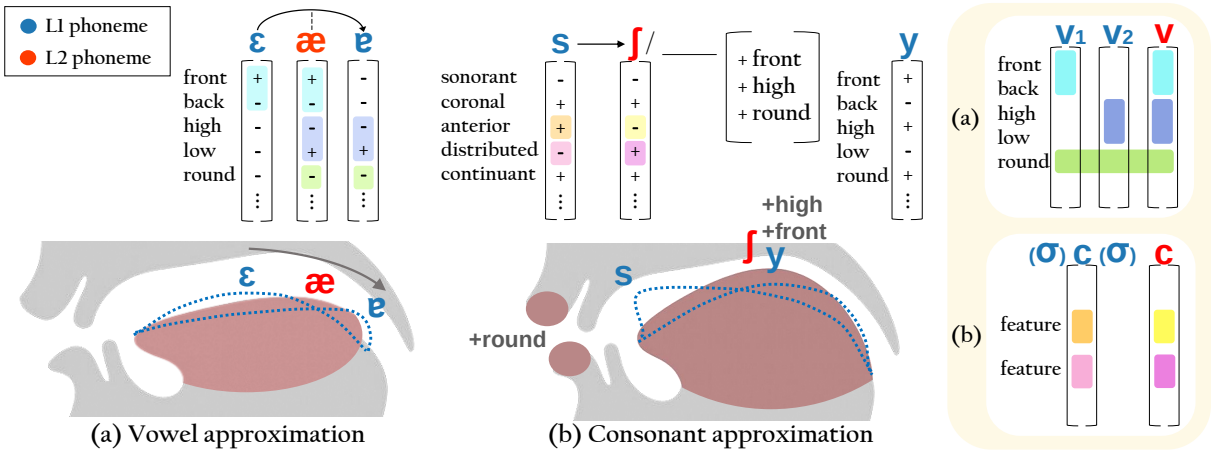

Figure 2: (a) An L2 vowel is approximated by combining two L1 vowels whose features jointly mirror the phonological identity of the target vowel. (b) An L2 consonant is approximated by inserting one or two L1 segments, forming allophones that more closely match the phonological features of the target consonant.

sitional phoneme approximation (CPA), a representation technique that approximates non-native L2 phonemes using sequences of L1 phonemes as in Figure 1. Building on articulatory proximity and theoretical linguistics, CPA is formulated over a phonological feature space. We apply CPA in a single 10-minute pronunciation training session with 20 Korean learners of English. Computational and quantitative evaluations across acoustic, phoneme, and word levels demonstrate measurable improvements with minimal instruction.

2 Method

To formulate CPA, we represent phonemes as 22-dimensional feature vectors (Mortensen et al., 2016)[1]. To approximate the feature vector of the target phoneme, we then define a rule for composing these vectors in the feature space. We also show how the combination is phonetically realized (Fig. 2). We apply CPA only where it adds value. It is skipped (i) when the target segment is already present in L1 with the same feature vector as no approximation is needed and (ii) when an identical match cannot be constructed in the feature space; a forced composite would add little guidance, and the large acoustic gap itself helps learners notice and acquire the new sound (Flege, 1995).

2.1 Vowel Approximation

The composition of vowels follows the principle of monophthongization, a phonological process that reduces two vowel sounds to a single vowel (Philippa et al., 2017; Elramli, 2018; Alahdal, 2019). Formally, it is defined as the following operation in the feature vector space: we take backness ([front], [back]) from the first vowel, height ([high], [low]) from the second, and assign rounding ([round]) if either source vowel is rounded, as illustrated in Table 1. This composition covers three of the four dimensions of vowel identity, excluding tenseness due to its lack of consistent articulatory grounding (Raphael and Bell-Berti, 1975). Among candidate pairs with an exact match to the L2 target, we select those whose individual vowels exhibit fewer unmatched features.

Feature	Target	V1	V2	CPA
French /y/ ~ Spanish /i/+/u/				
IPA	/y/	/i/	/u/	/i/+/u/
front	–	–	+	–
back	–	–	+	–
high	+	+	+	+
low	–	–	–	–
round	+	–	+	+
English /ɒ/ ~ Mongolian /ɔ/+/a/				
IPA	/ɒ/	/ɔ/	/a/	/ɔ/+/a/
front	–	–	+	–
back	+	+	+	+
high	–	–	–	–
low	+	–	+	+
round	+	+	–	+

Table 1: CPA-based vowel approximation. Each block shows how L2 vowels are approximated using a combination of two L1 vowels, with feature-wise comparisons.

2.2 Consonant Approximation

Consonant approximation in CPA draws on patterns of allophonic variation, in which consonant re-

[1]We add [front] as part of the backness dimension.

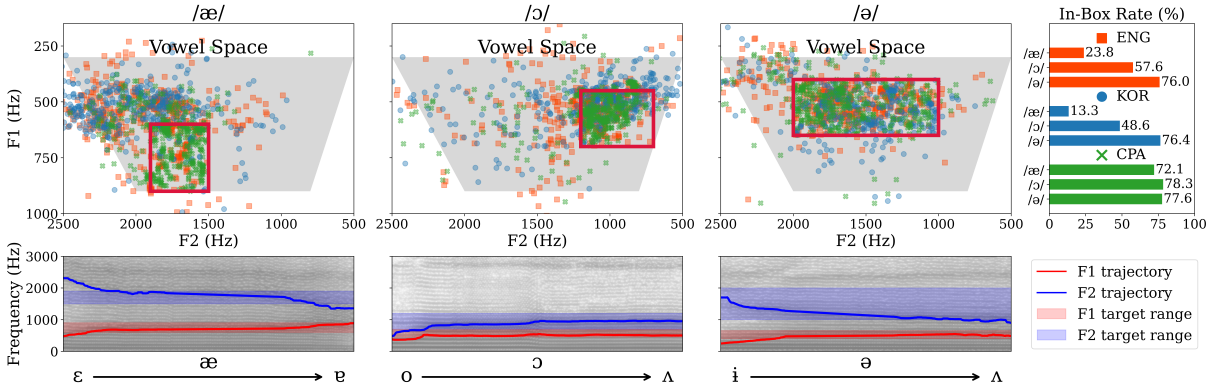

Figure 3: Vowel production and formant trajectories for /æ/, /ɔ/, and /ə/. Top: Distributions of speaker productions across conditions (ENG, KOR, CPA), with in-box rates (%). Red boxes show target F1–F2 regions; gray trapezoids indicate canonical vowel space. Bottom: CPA productions shown with spectrograms and smoothed F1 (red) and F2 (blue) trajectories. Shaded bands indicate target formant ranges; arrows show intended transitions.

alization systematically shifts depending on neighboring segments (Hayes, 2011). CPA selects a base L1 consonant and applies a feature modification conditioned by L1 phonological contexts, such as palatalization before front vocoids, labialization before rounded vowels, and spirantization under reduced closure. Table 2 summarizes the articulatory domains involved in these context-driven shifts. In feature space, these adjustments correspond to changes in place or manner features that move the base consonant toward the L2 target while remaining grounded in familiar articulatory gestures.

For example, in a language such as Japanese, which has no direct phonemic counterpart to the Korean sequence /tɕe/, the target can be approximated through coarticulation: the alveolar stop /t/ becomes palatalized before the high front vowel /i/, producing /te/ that can be realized as [tɕ(i)e] in this configuration, closely resembling the Korean sound. See Table 3 for the corresponding feature shift. Rather than introducing a new segment outright, CPA enables learners to approximate the L2 consonant using coarticulatory patterns already present in the L1.

3 Experiments

Through a 10-minute training session that targets Korean English-learners, we evaluate whether CPA-based pronunciation training leads to improvements within a short time frame. The Korean-English language pair was chosen due to their substantial phonological differences (Ha et al., 2009). For objective and comprehensive evaluation, we adopt computational methods to assess pronunciation across acoustic, phoneme, and word levels.

Category	Transformation	Core feature changes
Laryngeal	Voicing	[+voice]
	Fortition	[+constricted glottis]
	Aspiration	[+spread glottis]
Place	Velarization	[+back]
	Labialization	[+labial], [+round]
	Dentalization	[+distributed]
	Palatalization	[–anterior], [+distributed]
Manner	Nasalization	[+nasal]
	Lateralization	[+lateral]
	Spirantization	[+continuant], [+strident]

Table 2: Phonological transformations categorized by articulatory domain. Listed features indicate core changes required to license each transformation.

Transformation	Feature	t	→	tɕ	/ _i
Palatalization	anterior	+		–	
	distributed	–		+	

Table 3: Feature shift from /t/ to /tɕ/ in Japanese, triggered by a following /i/.

3.1 Experimental Setup

We selected 18 English words containing phonemes absent from the Korean phonemic inventory as shown in Table 5. We recruited 20 native Korean speakers and presented three types of visual cues: (1) the English word alone (ENG), (2) the English word and its Hangul transcription (KOR), and (3) the English word with a CPA-based Korean grapheme (CPA). Here, the KOR cue follows Korea's official Loanword Transcription Rules (Ministry of Culture, Sports and Tourism, 2017). In each condition, participants read each word aloud three times (nine total). Details on

Target	Approximation		Accuracy (%)		
ENG	KOR	CPA	KOR	ENG	CPA
/ɔ/	/o/	/o/ + /ʌ/	4.8	10.4	**10.9**
/æ/	/e/	/ɛ/ + /ɐ/	0.7	7.4	**14.5**
/ə/	/ʌ/	/i/ + /ʌ/	11.0	39.3	**46.0**
/b/*	/p/	/i/ + /p/	9.2	57.5	**73.3**
/d/*	/t/	/i/ + /t/	41.9	63.9	**78.1**
/g/*	/k/	/i/ + /k/	16.7	45.8	**72.5**
/dʒ/*	/tɕ/	/i/ + /tɕ/ + /y/	5.8	33.3	**64.2**
/l/*	/ɾ/	/il/ + /ɾ/	91.7	96.7	**99.2**
/m/*	/m^b/	/im/ + /m^b/	93.9	98.3	**98.3**
/n/*	/n^d/	/in/ + /n^d/	95.8	99.2	**100.0**
/ʃ/	/ɕ/	/s/ + /y/	60.0	77.0	**87.0**
/tʃ/	/$tɕ^h$/	/$tɕ^h$/ + /y/	71.7	73.3	**83.3**
/dʒ/	/dʑ/	/dʑ/ + /y/	42.5	25.0	**25.0**
Weighted Average			31.1	45.4	**53.4**

Table 4: ASR-based phoneme recognition accuracy for each target English phoneme absent from Korean. Asterisks (*) denote word-initial consonants.

the experimental setup and grapheme design are provided in Appendix A.

3.2 Acoustic-Level Evaluation

To analyze vowel acoustics across different reading conditions, we aligned recordings with IPA transcriptions using the Montreal Forced Aligner (MFA) (McAuliffe et al., 2017). For each vowel token, we extracted its spectral segment and estimated F1 and F2 using formant tracking (Markel and Gray, 2013). We tracked F1–F2 trajectories over time and checked whether they fell within the reference formant range.

Figure 3 shows the F1–F2 distributions under the three conditions, with red boxes indicating reference formant ranges (Yang, 2019). Trajectories passing through these boxes are more likely to be perceived as the target vowel. CPA consistently yields higher in-box rates across vowels, with an overall rate of 76.0%. Gains were especially notable for /æ/ and /ɔ/, whereas /ə/, which has a broader canonical range (Flemming, 2009), showed only moderate improvement. Furthermore, representative CPA spectrograms with overlaid F1 and F2 trajectories are also shown in the bottom row of Figure 3, illustrating why vowel sequences are perceived as realizations of the target phoneme.

3.3 Phoneme-Level Evaluation

We evaluate phoneme-level intelligibility using an automatic speech recognition (ASR) model. Specifically, we use Wav2Vec2Phoneme (Xu et al., 2022), a multilingual speech-to-IPA model, to decode each utterance into a phoneme sequence by selecting the most probable English phoneme at each timestep. We then compute accuracy as the proportion of cases where the target phoneme appeared in the correct position. Table 4 shows the phoneme recognition accuracy for each target segment under the three cue conditions. The CPA cue consistently leads to higher accuracy across individual segments. The overall average, computed across all target segments, is highest for CPA (53.4%), followed by ENG (45.4%) and KOR (31.1%).

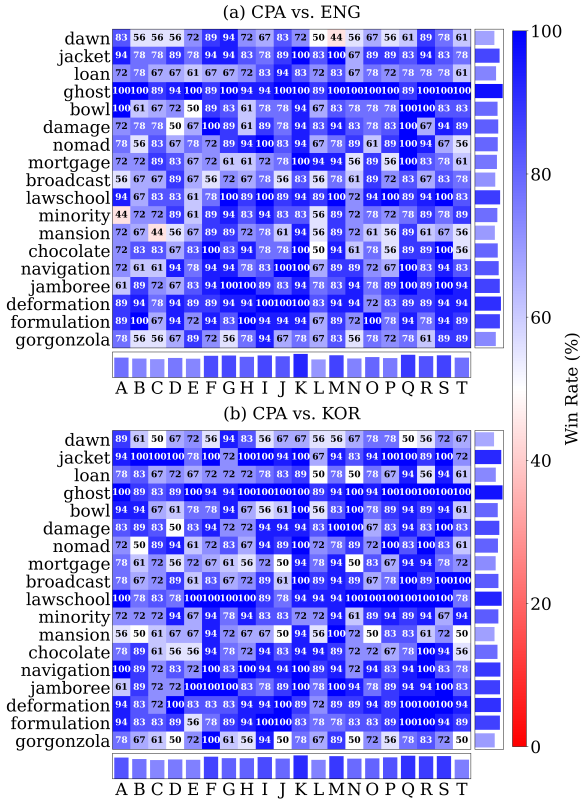

Figure 4: LLM-based word-level nativeness comparison: (a) CPA vs. ENG and (b) CPA vs. KOR. Each cell summarizes the CPA win rate (%) from 18 pairwise comparisons per word and participant. Bars show average win rates across words and participants.

3.4 Word-Level Evaluation

To assess word-level perceptual nativeness, we utilize the LLM-as-a-judge (Parikh et al., 2025). We perform pairwise comparisons between utterances using deterministic decoding settings and apply debiasing techniques to mitigate order effects (Zheng et al., 2023; Liusie et al., 2024). For each participant and word, GPT-4o (Hurst et al., 2024) conducted 18 pairwise comparisons between the two

methods with the prompt provided in Appendix A.4

As shown in Figure 4, CPA-based utterances were preferred over ENG in nearly all cases, with 357 out of 360 cells showing a win rate above 50%; only 3 cells (0.8%) fell below this threshold, with 44% win rates. Against KOR, CPA achieved over 50% win rates in all 360 cells. On average, CPA was preferred in 80.6% of comparisons against ENG and 81.9% against KOR, indicating a consistent perception of greater nativeness. These results suggest that accurately approximating L2 phonemes through CPA significantly improves word-level perceptual nativeness in L2 speech. Supplementary human evaluation results are provided in Appendix C.

4 Conclusion

This study introduces **compositional phoneme approximation (CPA)**, an approach that approximates L2 phonemes using compositional sequences of L1 phonemes. CPA operates over phonological feature representations grounded in phonetic articulatory knowledge, providing a principled framework for cross-linguistic phoneme mapping that enables efficient L2 production training.

5 Limitations

While CPA successfully models L2 phonemes through feature-based combinations of L1 segments, it currently focuses on phonemic-level approximation without incorporating suprasegmental elements such as stress, accent, or tone. These features often influence naturalistic pronunciation and perception, especially in tonal or rhythmically distinct languages (Yip, 2002; Nespor et al., 2011). As such, extending CPA to account for higher-level phonological features remains an open direction for broadening its applicability.

CPA itself is orthography-independent, operating only on phonological features. For classroom use, though, its composite sounds must still be written, and scripts differ in how transparently they encode pronunciation. Hangul's near one-to-one sound mapping simplifies the display, whereas scripts with less tight sound–symbol correspondence may call for different visual conventions. Adapting the cue to other writing systems (e.g., IPA symbols, romanization, or native characters) and testing how each variant supports learning is a sensible next step.

In implementing CPA-based instruction, a minor but practical consideration is to ensure that any epenthetic elements introduced in the compositional cue remain brief and soft (as in Appendix A.2). Without this care, added segments may become perceptually salient and distract from the intended phoneme target. As part of effective instructional design, this is a pedagogical detail worth attending to during training.

Acknowledgments

This work was conducted independently of the authors' past or present institutional affiliations and without external funding. We thank Professor Jieun Song of Korea Advanced Institute of Science and Technology (KAIST) and Professor Ho Young Lee of Seoul National University for invaluable consultation and guidance in linguistics.

References

Christian Abello-Contesse. 2009. Age and the critical period hypothesis. *ELT journal*, 63(2):170–172.

Ameen Alahdal. 2019. Vowel deletion in raimi and najdi arabic. *Utopía y praxis latinoamericana: revista internacional de filosofía iberoamericana y teoría social*, (6):243–253.

Vipul Arora, Aditi Lahiri, and Henning Reetz. 2018. Phonological feature-based speech recognition system for pronunciation training in non-native language learning. *The Journal of the Acoustical Society of America*, 143(1):98–108.

Fadhl Mohammed Awadh Mohammed Awadh, Nur Hidayanto Pancoro Setyo Putro, Albert Efendi Pohan, and Yasir Ayed Alsamiri. 2024. Improving english pronunciation through phonetics instruction in yemeni efl classrooms. *Journal of Languages and Language Teaching*, 12(2):930–940.

Wendy Baker and Pavel Trofimovich. 2005. Interaction of native-and second-language vowel system (s) in early and late bilinguals. *Language and speech*, 48(1):1–27.

Catherine T Best et al. 1994. The emergence of native-language phonological influences in infants: A perceptual assimilation model. *The development of speech perception: The transition from speech sounds to spoken words*, 167(224):233–277.

Paul Boersma. 2011. Praat: doing phonetics by computer [computer program]. *http://www. praat. org/*.

Michael David Carey, Arizio Sweeting, and Robert Mannell. 2015. An l1 point of reference approach to pronunciation modification: Learner-centred alternatives to 'listen and repeat'. *Journal of Academic Language and Learning*, 9(1):A18–A30.

Yousef Mokhtar Elramli. 2018. An optimality theoretic analysis of monophtongization in libyan arabic. *Kashmir Journal of Language Research*, 21(1):133–143.

Maria Gabriela Tobar Espinoza, Yola Indaura Chica Cárdenas, Claudia Valerie Piedra Martinez, and Francisco Israel Brito Saavedra. 2021. The use of audiovisual materials to teach pronunciation in the esl/efl classroom: El uso de materiales audiovisuales para enseñar la pronunciación en el aula de esl/efl. *South Florida Journal of Development*, 2(5):7345–7358.

Daniel Felps, Heather Bortfeld, and Ricardo Gutierrez-Osuna. 2009. Foreign accent conversion in computer assisted pronunciation training. *Speech communication*, 51(10):920–932.

James E Flege. 1995. Second language speech learning: Theory, findings, and problems. *Speech perception and linguistic experience: Issues in cross-language research*, 92(1):233–277.

James Emil Flege. 1993. Production and perception of a novel, second-language phonetic contrast. *The Journal of the Acoustical Society of America*, 93(3):1589–1608.

James Emil Flege, Katsura Aoyama, and Ocke-Schwen Bohn. 2021. The revised speech learning model (slm-r) applied. *Second language speech learning: Theoretical and empirical progress*, pages 84–118.

Edward Flemming. 2009. The phonetics of schwa vowels. *Phonological weakness in English*, 493:78–95.

Valeria Galimberti, Joan C Mora, and Roger Gilabert. 2023. Teaching efl pronunciation with audio-synchronised textual enhancement and audiovisual activities: Examining questionnaire data. In *Proceedings of the 7th International Conference on English Pronunciation: Issues and Practices (EPIP7)*, pages 70–82.

María de los Ángeles Gómez González and Alfonso Lago Ferreiro. 2024. Web-assisted instruction for teaching and learning efl phonetics to spanish learners: Effectiveness, perceptions and challenges. *Computers and Education Open*, 7:100214.

Mirko Grimaldi, Bianca Sisinni, Barbara Gili Fivela, Sara Invitto, Donatella Resta, Paavo Alku, and Elvira Brattico. 2014. Assimilation of l2 vowels to l1 phonemes governs l2 learning in adulthood: a behavioral and erp study. *Frontiers in human neuroscience*, 8:279.

Seunghee Ha, Cynthia J Johnson, and David P Kuehn. 2009. Characteristics of korean phonology: Review, tutorial, and case studies of korean children speaking english. *Journal of communication disorders*, 42(3):163–179.

Bruce P Hayes. 2011. *Introductory phonology*. John Wiley & Sons.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Olaf Husby, Åsta Øvregaard, Preben Wik, Øyvind Bech, Egil Albertsen, Sissel Nefzaoui, Eli Skarpnes, and Jacques C Koreman. 2011. Dealing with l1 background and l2 dialects in norwegian capt. In *SLaTE*, pages 133–136.

Yannick Jadoul, Bill Thompson, and Bart De Boer. 2018. Introducing parselmouth: A python interface to praat. *Journal of Phonetics*, 71:1–15.

Reima Karhila, Anna Riikka Smolander, Sari Ylinen, and Mikko Kurimo. 2019. Transparent pronunciation scoring using articulatorily weighted phoneme edit distance. In *Interspeech*, pages 1866–1870. International Speech Communication Association (ISCA).

Natalia Kartushina and Ulrich H Frauenfelder. 2014. On the effects of l2 perception and of individual differences in l1 production on l2 pronunciation. *Frontiers in psychology*, 5:1246.

Veronica Khaustova, Evgeny Pyshkin, Victor Khaustov, John Blake, and Natalia Bogach. 2023. Capturing accents: An approach to personalize pronunciation training for learners with different l1 backgrounds. In *International Conference on Speech and Computer*, pages 59–70. Springer.

Katrin Leppik, Cristian Tejedor-García, Eva Liina Asu, and Pärtel Lippus. 2022. Improving spanish l1 learners' perception and production of estonian vowels. In *Proc. ISAPh 2022*, pages 34–39.

Adian Liusie, Potsawee Manakul, and Mark Gales. 2024. LLM comparative assessment: Zero-shot NLG evaluation through pairwise comparisons using large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 139–151, St. Julian's, Malta. Association for Computational Linguistics.

John D Markel and AH Jr Gray. 2013. *Linear prediction of speech*, volume 12. Springer Science & Business Media.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Interspeech*, volume 2017, pages 498–502.

Ministry of Culture, Sports and Tourism. 2017. Loanword transcription rules [외래어 표기법]. Notification No. 2017-14.

David R Mortensen, Patrick Littell, Akash Bharadwaj, Kartik Goyal, Chris Dyer, and Lori Levin. 2016. Panphon: A resource for mapping ipa segments to articulatory feature vectors. In *Proceedings of COLING*

2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 3475–3484.

Marina Nespor, Mohinish Shukla, and Jacques Mehler. 2011. 49 stress-timed vs. syllable-timed languages.

Aditya Kamlesh Parikh, Cristian Tejedor-Garcia, Catia Cucchiarini, and Helmer Strik. 2025. Zero-shot speech llms for multi-aspect evaluation of l2 speech: Challenges and opportunities. *Proc. SLaTE 2025*, pages 11–15.

Kariem Philippa, Marlies Philippa, and Annelies Roeleveld. 2017. Monophthongization of ay/ai and aw/au: A comparison between arabic and germanic dialects. *Amsterdamer Beiträge zur älteren Germanistik*, 77(3-4):616–636.

Lawrence J Raphael and Fredericka Bell-Berti. 1975. Tongue musculature and the feature of tension in english vowels. *Phonetica*, 32(1):61–73.

Shuju Shi, Chilin Shih, and Jinsong Zhang. 2019. Capturing l1 influence on l2 pronunciation by simulating perceptual space using acoustic features. In *INTERSPEECH*, pages 2648–2652.

Ratree Wayland. 2021. *Second language speech learning*. Cambridge University Press.

Qiantong Xu, Alexei Baevski, and Michael Auli. 2022. Simple and effective zero-shot cross-lingual phoneme recognition. In *Proc. Interspeech 2022*, pages 2113–2117.

Byunggon Yang. 2019. A comparison of normalized formant trajectories of english vowels produced by american men and women. *Phonetics and Speech Sciences*, 11(1):1–8.

Moira Yip. 2002. *Tone*. Cambridge University Press.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

A Experiment Details

A.1 Subject Recruitment and Payment

We recruited a total of 20 native Korean speakers, aged 20–70, who had not lived in an English-dominant environment before the age of 13, in line with the critical period hypothesis (Abello-Contesse, 2009). Among them, 12 were residing in Korea and 8 in Massachusetts, U.S., recruited through an anonymous platform. Participation was voluntary, with informed consent and the option to withdraw at any time. Each participant received upfront compensation equivalent to twice the minimum hourly wage for the one-hour session.

A.2 Data Collection

The study consisted of three recording sessions in a controlled setting. Before recording CPA-based pronunciations, participants underwent a 10-minute training using a single slide in Fig. 5. It introduces the CPA-based Korean grapheme system and instructs participants to (1) pronounce the symbols inside each box quickly, (2) articulate the contents of each box as a unit, and (3) pronounce smaller boxes including epenthetic vowels or transitional elements softly and briefly. To familiarize participants with the system, the slide featured five example words, each practiced once before recording.

Figure 5: An instructional slide for reading CPA-based Korean graphemes used in a 10-minute training session.

Each of the 18 target words was pronounced three times given each visual cue in Tab. 5, resulting in a total of 3,240 audio clips. Recordings were made using the `python-sounddevice`[2] module at a fixed sampling rate of 16,000 Hz to match model input requirements. No personal or identifying information was collected and all recordings were anonymized to protect participant privacy.

[2] https://github.com/spatialaudio/python-sounddevice

A.3 CPA-based Korean Grapheme Design

The grapheme representation of CPA-based pronunciation follows native principles of the Hangul, which makes it easily interpretable by native Korean readers without additional explanation. Mirroring Hangul's way of forming diphthongs by combining two monophthongs, we visualize approximated vowels in CPA by composing the two component vowels used in the approximation. For consonants, conditioning phonemes for allophonic variations are displayed in adjacent blocks at half size, indicating a quick, weak articulation rather than an independent sound.

A.4 LLM Prompt for Word-Level Evaluation

> Please act as an impartial judge and evaluate which of two pronunciations sounds closer to a native speaker's pronunciation.
> You will hear two audio samples, A and B, in that order.
> Both are recordings of the same speaker saying the word "[word]."
> Which pronunciation sounds more native-like?
> You must choose only one: A or B.
> Do not provide any explanation—just respond with the letter.

B Evaluation Models

B.1 Acoustic-level evaluation

We used the MFA (MIT License) with the pretrained `english_mfa` acoustic and pronunciation models to automatically generate phoneme-level alignments (McAuliffe et al., 2017). Then, we estimated F1 and F2 over time using `parselmouth` (Jadoul et al., 2018), a Python interface to the Praat software (Boersma, 2011) (GNU General Public License v3.0).

B.2 Phoneme-level evaluation

We adopted `Wav2Vec2Phoneme` (Xu et al., 2022) to transcribe the recorded speech of participants into phonemes. The model is open-source with an Apache-2.0 license at Huggingface[3]. The model is specified in a configuration file and was executed on an M1 MacBook Air with 16GB RAM.

B.3 Word-level evaluation

We utilized `gpt-4o-audio-preview`[4] (Hurst et al., 2024) via the OpenAI API to perform zero-shot pairwise comparisons, with `temperature=0` and `seed=0` for deterministic inference.

[3] https://huggingface.co/facebook/wav2vec2-lv-60-espeak-cv-ft

[4] https://platform.openai.com/docs/models/gpt-4o-audio-preview

English Word (ENG)	IPA	Target Phoneme	Hangul Transcription (KOR)	CPA-based Korean Grapheme (CPA)
dawn	/ dɔn /	/d/*, /ɔ/	돈	
jacket	/ dʒækɪt /	/dʒ/*, /æ/	재킷	
loan	/ loʊn /	/l/*	론	
ghost	/ goʊst /	/g/*	고스트	
bowl	/ boʊl /	/b/*	볼	
damage	/ dæmɪdʒ /	/d/*, /æ/, /dʒ/	대미지	
nomad	/ noʊmæd /	/n/*, /æ/	노마드	
mortgage	/ mɔrgɪdʒ /	/m/*, /ɔ/, /dʒ/	모기지	
broadcast	/ brɔdkæst /	/b/*, /ɔ/, /æ/	브로드캐스트	
lawschool	/ lɔskul /	/l/*, /ɔ/	로스쿨	
minority	/ maɪnɔrɪti /	/m/*, /ɔ/	마이너리티	
mansion	/ mænʃən /	/m/*, /æ/, /ʃ/, /ə/	맨션	
chocolate	/ tʃɔkəlɪt /	/tʃ/, /ɔ/, /ə/	초콜릿	
navigation	/ nævɪgeɪʃən /	/n/*, /æ/, /ʃ/, /ə/	네비게이션	
jamboree	/ dʒæmbəri /	/dʒ/*, /æ/, /ə/	잼버리	
deformation	/ dɪfɔrmeɪʃən /	/d/*, /ɔ/, /ʃ/, /ə/	디포메이션	
formulation	/ fɔrmjʊleɪʃən /	/ɔ/, /ʃ/, /ə/	포뮬레이션	
gorgonzola	/ gɔrgənzoʊlə /	/g/*, /ɔ/, /ə/	고르곤졸라	

Table 5: Selected words for evaluation, target phonemes, and visual cues provided to participants.

C Supplementary Human Evaluation

C.1 Evaluation Setup

To complement the LLM-based perceptual evaluation in the main analysis, we conducted a small-scale human listener study. Ten native American English tutors were recruited from a commercial online English tutoring platform.[5] All participants provided informed consent, and responses were anonymized. Each rater evaluated two comparison types: (a) CPA vs. ENG and (b) CPA vs. KOR. For each type, six (speaker × word) combinations per rater were sampled such that speakers and words were evenly distributed across raters. Within each combination, raters provided judgments for the nine possible pairwise comparisons (3×3), with A/B order randomized to mitigate ordering effects. Every combination was evaluated by two independent raters to assess inter-rater consistency, yielding 270 unique utterance pairs per comparison type. Each pair was independently rated twice.

C.2 Results

CPA outperformed KOR, achieving an average win rate of 78.5% with strong agreement between LLM and human judgments at 76.5%. In contrast, CPA was preferred over ENG in only 45.9% of the cases, and the alignment between LLM and human judgments was notably lower at 46.4%. According to rater feedback, both CPA and ENG utterances were generally intelligible. However, CPA tended to produce more accurate phoneme realizations, while also sounding less native-like in some cases due to

[5] https://www.ringleplus.com/

Comparison	Successful	Unsuccessful
CPA vs. ENG	60.9%	38.2%
CPA vs. KOR	83.3%	76.1%

Table 6: CPA win rates against ENG and KOR baselines, conditioned on whether CPA productions met the prosodic training criterion.

elongated articulatory patterns.

Acoustic analyses further supported the prosodic basis of this discrepancy. A lower CPA-to-ENG word-duration ratio, indicating faster CPA delivery, was associated with higher CPA preference (Spearman's $\rho = -0.66$, $p = 4.77 \times 10^{-6}$). Additionally, speakers with CPA win rates above 50% exhibited significantly smaller CPA-to-ENG duration ratios compared to those below 50% (Mann–Whitney U, $p = 0.0002$).

To analyze the effect of training success on perceptual outcomes, we defined *successful CPA training* as productions whose CPA-to-baseline word-duration ratio was ≤ 1.0, consistent with realizing approximated phonemes with minimal epenthesis. Under this criterion, the aggregate CPA win rates were as shown in Table 6.

These findings suggest that CPA's segmental benefits offer perceptual advantages when accompanied by appropriate prosodic control. Incorporating explicit training on prosodic features such as stress and rhythm may further enhance nativeness beyond CPA's current segmental focus.

D Potential Risks

While CPA offers a structured approach to early-stage L2 pronunciation, it also poses potential pedagogical risks. One concern is pronunciation distortion, as the synthesized approximations may not fully capture the acoustic properties of target L2 phonemes, potentially leading to inaccurate articulation. Additionally, relying exclusively on L1 phonemes risks oversimplifying the phonological complexity of the L2, which may obscure subtle contrasts and hinder learners' ability to internalize the nuances of the L2 sound system. Finally, prolonged or uncritical use of CPA may lead to over-reliance, limiting the development of authentic and native-like pronunciation skills over time. Therefore, CPA should be viewed as a supportive tool rather than a standalone solution in pronunciation pedagogy. Future research should explore how to mitigate these risks while leveraging the pedagogical benefits of CPA.

E Disclosure of AI Writing Assistance

We acknowledge the use of `ChatGPT`[6], a chat-based AI assistant developed by OpenAI, for code-related assistance during our research. However, the core algorithms of our proposed method and its evaluation were independently developed without AI assistance. We also did not employ AI for use cases that require disclosure, such as generating low-novelty text, proposing new ideas, or creating new content based on original ideas.

[6] https://chat.openai.com/