Inter-linguistic Phonetic Composition (IPC): A Theoretical and Computational Approach to Enhance Second Language Pronunciation

Anonymous ACL submission

Abstract

Learners of a second language (L2) often unconsciously substitute unfamiliar L2 phonemes with similar phonemes from their native language (L1), even though native speakers of the L2 perceive these sounds as distinct and non-interchangeable. This phonemic substitution leads to deviations from the standard phonological patterns of the L2, creating challenges for learners in acquiring accurate English pronunciation. To address this, we propose **Inter-linguistic Phonetic Composition (IPC)**, a novel computational method designed to minimize incorrect phonological transfer by reconstructing L2 phonemes as composite sounds derived from multiple L1 phonemes. Tests with two automatic speech recognition models demonstrated that when L2 speakers produced IPC-generated composite sounds, the recognition rate of target L2 phonemes improved by 20% compared to when their pronunciation was influenced by original phonological transfer patterns. The improvement was observed within a relatively shorter time frame, demonstrating rapid acquisition of the composite sound.

1 Introduction

Second-language (L2) speakers often experience difficulty articulating certain phonemes absent in their native language (L1), unconsciously substituting phonologically distinct L1 phonemes for the target L2 phonemes (Kartushina and Frauenfelder, 2014). This issue is further complicated by the influence and limitation of the L1 writing system, through which learners unintentionally distort L2 phonemes to fit the phonetic framework of L1 (Kamiyama, 1994; Bassetti, 2017). These adaptations reduce pronunciation clarity and lower recognition rates by both native speakers and automatic speech recognition (ASR) models (Souza and Gottardi, 2022; Cumbal et al., 2024).

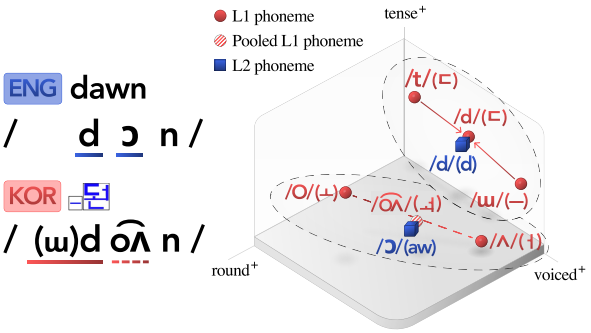


Figure 1: Illustration of inter-linguistic phonetic composition (IPC) in the phonological feature space. IPC approximates second-language (L2) phonemes absent from the native-language (L1) inventory by combining multiple L1 phonological feature vectors.

Previous attempts to address these challenges in L2 acquisition have involved time-intensive educational methods and achieved limited success. Such approaches include comprehensive phonology instruction (Kissling, 2013; Awadh et al., 2024), pronunciation training through audiovisual mimicry of L2 speech (Espinoza et al., 2021; Galimberti et al., 2023), and approaches that combine both strategies (González and Ferreiro, 2024).

To more effectively address this issue, we propose the **Inter-linguistic Phonetic Composition (IPC)**, a computational approach that synthesizes composite sounds from multiple L1 phonemes to closely approximate target L2 phonemes. In leveraging the speaker's existing phonological system, IPC significantly reduces the need for time-intensive instruction on entirely new phonemic representations.

2 Inter-linguistic Phonetic Composition

Across all languages, each phoneme can be represented as an N-dimensional binary vector within a phonological feature space, as shown in Equation 3. The value of N depends on the level of precision required for phoneme representation (Rubehn et al., 2024). In some cases, L2 phoneme vectors closely

$$\hat{\mathbf{v}}_{L2}^{(i)} = \begin{cases} \mathbf{v}_{L1}^{(i)} & \text{if } \mathbf{v}_{L2}^{(i)} = \mathbf{v}_{L1}^{(i)} \\ \mathbf{v}_{L1}^{(i_1)} \oplus \mathbf{v}_{L1}^{(i_2)} & \text{otherwise,} \quad \text{where } (i_1, i_2) = \arg\min_{i_1,i_2} \Big\{ \parallel \mathbf{v}_{L2}^{(i)} - \left(\mathbf{v}_{L1}^{(i_1)} \oplus \mathbf{v}_{L1}^{(i_2)}\right) \parallel_2 \\ & \qquad + \lambda \left(\parallel \mathbf{v}_{L2}^{(i)} - \mathbf{v}_{L1}^{(i_1)} \parallel_2 + \parallel \mathbf{v}_{L2}^{(i)} - \mathbf{v}_{L1}^{(i_2)} \parallel_2 \right) \Big\} \end{cases} \tag{1}$$

$$\hat{\mathbf{c}}_{L2}^{(j)} = \begin{cases} \mathbf{c}_{L1}^{(j)} & \text{if } \mathbf{c}_{L2}^{(j)} = \mathbf{c}_{L1}^{(j)} \quad \text{and} \quad \mathbf{c}_{L1}^{(j)} \in \mathcal{A}(\mathbf{c}_{L1}) \\ \mathbf{c}_{L1}^{(j)} \oplus V_{neighbor} & \text{if } \mathbf{c}_{L2}^{(j)} = \mathbf{c}_{L1}^{(j)} \oplus V_{neighbor} \quad \text{and} \quad \mathbf{c}_{L1}^{(j)} \oplus V_{neighbor} \in \mathcal{A}(\mathbf{c}_{L1}) \\ \mathbf{c}_{L2}^{(j)} & \text{otherwise} \end{cases} \tag{2}$$

Figure 2: Inter-linguistic phonetic composition algorithm (IPC) for representing second-language (L2) vowels and consonants using first-language (L1) phonemes. If an L2 phoneme shares identical phonological features with an L1 phoneme, we use the L1 phoneme directly. If not, we either represent the L2 phoneme with a combination of L1 phonemes or adopt it as is.

align with their nearest L1 counterparts, enabling direct substitution while maintaining a high recognition rate. IPC converts a single L2 phoneme vector into a composite of multiple L1 phoneme vectors in a way that preserves the essential acoustic features of the L2 phonemes, as shown in figure 2. In this figure, $\mathbf{v}_L^{(i)}$ and $\mathbf{c}_L^{(m)}$ represent the binary vectors for i-th vowel and m-th consonant in a specific word from language L, respectively.

$$\text{Features} = \begin{bmatrix} \text{+syllabic} \\ \text{+high} \\ \text{-back} \\ \text{-low} \\ \text{-anterior} \\ \text{-coronal} \\ \text{-round} \\ \vdots \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ \vdots \end{bmatrix} \tag{3}$$

2.1 Vowels

The IPC method draws inspiration from the phonological phenomenon of *Monophthongization*, where two adjacent vowels, or a vowel and a semivowel, merge to form a single monophthong. In some cases, only the primary vowel remains, resulting in a long vowel (Moosmüller, 1997; Forghema et al., 2022). In other cases, the blending of both vowels creates a new vowel—a process observed in languages like Korean, Mongolian, Dutch, and Arabic dialects. The nature of the new vowel that is created depends on the height of the primary vowel and the backness of the secondary vowels (Ahn and Iverson, 2005; Janhunen, 2012; Philippa et al., 2017). If one phoneme is rounded, the new vowel will also be rounded.

To reconstruct an L2 vowel that is absent from the L1 phoneme inventory, we first select a pair of phonemes—denoted as $\mathbf{v}_{L1}^{(i_1)}$ and $\mathbf{v}_{L1}^{(i_2)}$ in Equation 1—whose combined phonetic features most closely approximate the target L2 vowel. The selection process is governed by the method of Lagrange multipliers, an optimization technique that utilizes the multiplier λ as shown in Equation 1 to ensure that selected phonemes are minimally distant from the L2 vowel in the phonological feature space.

$$(\mathbf{v}_L^{(i_1)} \oplus \mathbf{v}_L^{(i_2)})_k = \begin{cases} (\mathbf{v}_L^{(i_1)})_k & \text{if } k \in \{\text{front}, \text{back}\} \\ (\mathbf{v}_L^{(i_2)})_k & \text{if } k \in \{\text{high}, \text{low}, \text{tense}\} \\ 1 & \text{if } k \in \{\text{round}\} \text{ and } (\mathbf{v}_L^{(i_1)})_k + (\mathbf{v}_L^{(i_2)})_k \geq 1 \\ (\mathbf{v}_L^{(i_2)})_k & \text{otherwise} \end{cases} \tag{4}$$

Phonetic features in consideration when selecting L1 phoneme pairs to approximate an L2 vowel are derived from the backness of the front vowel ($\mathbf{v}_{L1}^{(i_1)}$) and the height of the back vowel ($\mathbf{v}_{L1}^{(i_2)}$). If either vowel is rounded, this feature is retained in the composite vowel. Since all the other features remain constant across vowels, derivation from any vowel will yield the same result for these shared features. The operator $\oplus$ combines the feature vectors of the resulting vowel and is defined in Equation 4, where k indicates each phonological feature. In this way, IPC harnesses natural phonological processes to recreate L2 vowels in a form that L1 speakers can easily produce.

2.2 Consonants

Consonants often undergo phonetic shifts due to adjacent vowels (Mitterer et al., 2018). When a single L1 consonant cannot accurately represent an L2

consonant, we leverage this influence of neighboring vowels to derive an allophone vector that more closely approximates the intended L2. Figure 2 illustrates this consonant selection process, where $\mathcal{A}(\mathbf{c}_{L1})$ represents the L1 consonant inventory and $\mathbf{c}_{L1}^{(j)} \oplus V_{neighbor}$ denotes the derived allophone that, through the addition of neighboring vowel features, aligns more closely with the L2 consonant.

While vowels exist on a continuum, consonant features such as place of articulation, manner, and voicing are more discrete (Anderson, 2018). Consequently, even slight differences in features can have a significant impact on consonant recognition. To properly recreate L2 consonants, we insert neutral neighboring vowels at the beginning, middle, or end of a syllable formed from L1 phonemes. This process allows for the creation of new phonological relationships and formation of allophones from L1 sounds that can meet the nuanced features of the target L2 consonant. A consonant that forms a new syllabic relationship solely with the added vowel is classified as a singleton consonant, while a consonant that relates to both an existing vowel and an added vowel—or to two newly added vowels—is classified as a geminate consonant.

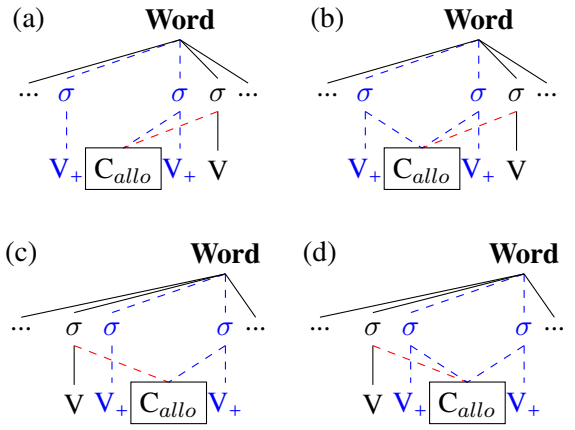


Figure 3: The tree diagram (Hammond, 1995) illustrates the introduction of two surrounding vowels V_+ that creates a new allophone C_{allo} for the existing consonant (C: consonant, V: vowel). (a) a new singleton consonant in a CV structure. (b) a new geminate consonant in CV. (c) a new singleton in VC. (d) a new geminate in VC. Other cases are explained in Appendix A.

Because singleton and geminate consonants have different values, they can function as distinct allophones in various phonological environments. Figure 3 illustrates a situation where two new neighboring vowels V_+ are added around a consonant to create an intervocalic allophone C_{allo}. Blue dashed lines indicate new syllabic relationships that can be created, while red dashed lines show disruptions of existing ones (σ: represents syllables). Geminate consonants relate to two vowels, so they are connected by two lines.

To minimize negative impact on phoneme pronunciation, the newly added vowels must be as weak and neutral as possible. To ensure this, the most unmarked vowel vector is used to weaken the pronunciation (Crosswhite, 2004). When multiple vowel candidates $\mathcal{V}$ are available, preference is given to the vowel with the most zero-valued phonological features (unspecified or neutral features), as formalized in Equation 5. $V_{neighbor} = \{\mathbf{v}_{pre}, \mathbf{v}_{suc}\}$ can contain $\mathbf{v}_{suc}$ (a preceding V_+ vector) or $\mathbf{v}_{pre}$ or (a succeeding V_+ vector). If only one vowel is added, the other vector is treated as a zero vector, exerting no influence on pronunciation.

$$V_{neighbor} = \underset{\{\mathbf{v}_{pre},\mathbf{v}_{suc}\}\in\mathcal{V}}{\arg\min} \left(\| \mathbf{v}_{pre} \|_1 + \| \mathbf{v}_{suc} \|_1 \right) \tag{5}$$

3 Experiments

The Korean and English languages exhibit significant differences in both phoneme inventories and writing systems. We conducted a statistical analysis to evaluate how closely the IPC-generated composite sound, created from Korean phonemes, can approximate an English phoneme.

3.1 Experimental Setup

We selected 20 English words containing phonemes absent from the Korean phonetic inventory. Based on the critical period hypothesis (Abello-Contesse, 2009), we recruited 20 participants who had never lived in an English-dominant environment before the age of 13, ensuring that English is their L2. Each participant was presented with three types of visual cues, which they read aloud three times in sequence: (1) the English word alone, (2) the English word with its Koreanized transcription, and (3) the English word with an IPC-based Korean grapheme. We assessed the phonemes in each case by estimating both confidence scores and accuracy using two phoneme recognition models: `Wav2Vec2Phoneme` (Baevski et al., 2020) and fine-tuned `XLSR-53` (Xu et al., 2022). The confidence scores were calculated by applying softmax functions to the logit values of the top-5 highest-scoring phonemes during phoneme decoding, and accuracy was determined as the ratio of instances where the top-scoring phoneme matched the ground truth

	Wav2Vec2Phoneme									XLSR-53								
	ENG			ENG + KOR			**ENG + IPC (ours)**			ENG			ENG + KOR			**ENG + IPC (ours)**		
IPA	AVG	STD	ACC	AVG	STD	ACC	AVG	STD	ACC	AVG	STD	ACC	AVG	STD	ACC	AVG	STD	ACC
/ɔ/	**0.071**	0.099	**7%**	0.042	0.071	2%	0.066	0.087	**7%**	0.174	0.151	29%	0.131	0.139	15%	**0.179**	0.151	**32%**
/oʊ/	0.041	0.106	5%	0.008	0.025	1%	**0.077**	0.161	**10%**	0.033	0.045	3%	0.021	0.031	1%	**0.057**	0.064	**9%**
/æ/	0.066	0.180	7%	0.010	0.055	1%	**0.116**	0.247	**14%**	0.093	0.126	8%	0.029	0.052	1%	**0.169**	0.193	**25%**
/ə/	0.221	0.367	26%	0.028	0.139	3%	**0.236**	0.368	**29%**	**0.335**	0.388	**41%**	0.099	0.230	15%	0.295	0.368	37%
/b/*	0.519	0.398	54%	0.092	0.192	6%	**0.624**	0.395	**68%**	0.437	0.391	46%	0.074	0.198	8%	**0.582**	0.399	**63%**
/d/*	0.274	0.327	30%	0.107	0.223	11%	**0.362**	0.364	**40%**	0.260	0.348	29%	0.113	0.225	11%	**0.374**	0.403	**41%**
/dʒ/*	0.240	0.320	31%	0.048	0.135	5%	**0.466**	0.324	**60%**	0.323	0.340	43%	0.069	0.161	8%	**0.533**	0.313	**71%**
/l/*	**0.754**	0.336	83%	0.603	0.393	68%	0.734	0.319	**86%**	**0.826**	0.297	**89%**	0.676	0.417	73%	0.750	0.357	82%
/m/*	0.770	0.369	80%	0.474	0.427	51%	**0.813**	0.336	**84%**	0.781	0.368	82%	0.584	0.429	62%	**0.841**	0.302	**89%**
/n/*	0.690	0.378	73%	0.451	0.417	51%	**0.818**	0.291	**89%**	0.793	0.339	83%	0.518	0.421	59%	**0.895**	0.239	**93%**
/ʃ/	0.565	0.399	63%	0.312	0.346	39%	**0.681**	0.370	**75%**	0.470	0.370	58%	0.235	0.274	34%	**0.589**	0.338	**73%**
/tʃ/	**0.526**	0.282	72%	0.435	0.300	65%	0.497	0.270	**75%**	0.459	0.290	68%	0.318	0.280	55%	**0.466**	0.240	**77%**
/dʒ/	0.184	0.268	23%	0.134	0.249	19%	**0.186**	0.265	**24%**	**0.299**	0.311	**45%**	0.244	0.294	33%	0.295	0.289	38%
Total	0.276	0.375	31%	0.135	0.278	15%	**0.326**	0.392	**37%**	0.312	0.366	37%	0.171	0.290	20%	**0.360**	0.373	**45%**

Table 1: Quantitative evaluation of phoneme confidence scores and accuracy across distinct pronunciation conditions. A superscript * on IPA symbols denotes phonemes that occur at the initial position of a word.

phoneme. Additional methodological details are provided in Appendix A.

3.2 Result

Table 1 demonstrates that IPC-based Korean grapheme improves the pronunciations of the majority of IPC-applied phonemes by approximately 17% compared to the English word alone and by 181% compared to the Koreanized transcription, on average. Given the Koreanization principle, which transcribes a single foreign phoneme using a single symbol, the improvement over the Koreanized transcription can be interpreted as the reduction in explicit phonological interference, while the improvement over the English word alone suggests mitigation of implicit phonological interference among native Korean speakers. This mitigation ultimately results in a 20% and 135% enhancement of overall pronunciation accuracy, respectively.

The vowel systems of English and Korean are significantly different, so when only Korean graphemes were used, vowel recognition exhibited very low accuracy. However, with IPC, the accuracy considerably improved. It can be seen that IPC adjusts the distance within the oral space where the vowels are articulated, thereby alleviating the differences between the two languages. On the other hand, in the case of consonants, IPC leads to a considerably higher accuracy, as it maps the correct allophones suitable for each environment.

Table 2 further validates that the enhancement in phoneme recognition is attributable to the application of IPC, as demonstrated by the t-test results with highly significant t-statistics and extremely low p-values across all conditions. The application of IPC shows statistically significant improvements in phoneme confidence scores over both baseline conditions. The greatest gains are seen over the ENG + KOR baseline, demonstrating IPC's effectiveness in reducing phonological transfer between languages. Moreover, the fact that training to read IPC-based graphemes took only 10 minutes underscores IPC's practical advantages. These results indicate that IPC not only improves phoneme recognition but also offers a user-friendly approach to second language acquisition, facilitating better pronunciation with minimal training.

	Wav2Vec2Phoneme		XLSR-53	
Baseline	t_stat	p_val	t_stat	p_val
ENG	9.567	2.04E-21	8.909	8.20E-19
ENG + KOR	29.876	1.86E-174	29.563	3.05E-171

Table 2: T-test results on phoneme confidence scores from the baseline and IPC-applied pronunciations.

4 Conclusion

This study introduced **IPC**, a novel system that maps L2 phonemes onto combinations of L1 phonemes. Empirical evidence demonstrates that by leveraging existing phonological knowledge and minimizing the need to learn the phonological information of a new language, IPC significantly enhances phonemic perception and pronunciation accuracy.

5 Limitations

While the Interlinguistic Phonetic Composition (IPC) approach has shown promising results in facilitating phoneme recognition and pronunciation between Korean and English, its reliance on mapping L2 phonemes to intermediary phonetic points between two closely related L1 phonemes presents certain limitations. This can be problematic when no equivalent L1 phoneme exists, or when the target phoneme does not align well with a midpoint between two L1 phonemes, leading to less accurate pronunciation and reducing the overall effectiveness of the method.

Although this paper provides a detailed discussion on transferring L2 phoneme vectors to L1 phoneme vectors, it lacks an in-depth examination of the phoneme-to-grapheme mapping process. The primary goal of modifying phoneme vectors is to help L1 speakers pronounce L2 phonemes more accurately, making the transcription into the L1 grapheme system a crucial step. However, L1 graphemes may not always adequately represent all phoneme vectors. For instance, consonant sounds in L2 that have no counterparts in L1 are directly represented in the L2 phoneme vector. When mapping these phonemes to graphemes, L2 characters are used, which may cause articulation difficulties if the speaker is unfamiliar with the L2 writing system.

Furthermore, the effectiveness of IPC may vary significantly across different language pairs due to differences in phonological systems and writing scripts. For example, in languages with more ambiguous writing systems like Arabic, where vowels are often not explicitly represented (Abu-Rabia and Taha, 2013), phoneme mapping becomes more challenging. Similarly, in languages that share the same script, such as those using the Latin alphabet, orthographic similarities might not correspond to phonetic similarities, limiting IPC's effectiveness.

In addition to these challenges, languages are often categorized based on their rhythmic timing patterns, including syllable-timed, stress-timed, and mora-timed languages (Hoequist, 1983; Nespor et al., 2011). Syllable-timed languages, such as French, Spanish, and Italian, exhibit relatively even syllable durations. Stress-timed languages, such as English, German, and Russian, emphasize the rhythmic grouping of stressed syllables. Mora-timed languages (Kubozono, 2017), such as Japanese and Hawaiian, follow an even more intricate rhythm, based on the mora, a smaller sub-syllabic unit. These rhythmic categories have a profound impact on how phonemes are perceived, articulated, and mapped between languages. For effective IPC, it is essential to consider not only phonemic transfer but also rhythmic aspects, such as syllable length and stress patterns. Phoneme vectors need to account for how these timing structures influence speech perception and articulation. This becomes particularly challenging when mapping between languages with drastically different rhythmic systems. On the other hand, tone languages such as Mandarin Chinese or Vietnamese use pitch contours to distinguish words (Yip, 2002), which is not represented through the standard phoneme vector system used in IPC.

To address these limitations, future iterations of IPC could explore incorporating tonal or non-phonemic features into the model. This might include modeling pitch contours, clicks, or other non-phonemic features that are linguistically significant in these languages, thereby broadening the applicability of IPC.

These considerations highlight the need for further research to adapt and refine IPC for broader linguistic contexts. Future studies could explore modifications to accommodate languages with non-phonetic writing systems or significantly different phonemic inventories. Additionally, developing alternative strategies for phoneme mapping—especially when direct L1 equivalents are lacking—would increase IPC's versatility and enhance its applicability in second language acquisition.

References

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

Salim Abu-Rabia and Haitham Taha. 2013. Reading in arabic orthography: Characteristics, research findings, and assessment. In *Handbook of orthography and literacy*, pages 321–338. Routledge.

Sang-Cheol Ahn and Gregory K Iverson. 2005. Structured imbalances in the emergence of the korean vowel system. *Historical linguistics*.

Catherine Anderson. 2018. *Essentials of linguistics*. McMaster University.

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.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Bene Bassetti. 2017. Orthography affects second language speech: Double letters and geminate production in english. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 43(11):1835.

Katherine Crosswhite. 2004. Vowel reduction. *Phonetically based phonology*, pages 191–231.

Ronald Cumbal, Birger Moell, José Lopes, and Olof Engwall. 2024. You don't understand me!: Comparing asr results for l1 and l2 speakers of swedish. *arXiv preprint arXiv:2405.13379*.

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.

Njuasi Ivo Forghema et al. 2022. The monophthongization of/ai/and/oi/in kom: An autosegmental perspective. *International Journal of Linguistics, Literature and Translation*, 5(9):01–13.

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.

Michael Hammond. 1995. Metrical phonology. *Annual review of anthropology*, 24(1):313–342.

Charles Hoequist, Jr. 1983. Syllable duration in stress-, syllable-and mora-timed languages. *Phonetica*, 40(3):203–237.

Juha Janhunen. 2012. *Mongolian*. John Benjamins.

T Kamiyama. 1994. Tuttle new dictionary of loanwords in japanese. *Rutland, Vermont: Charles E. Tuttle*.

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.

Elizabeth M Kissling. 2013. Teaching pronunciation: Is explicit phonetics instruction beneficial for fl learners? *The modern language journal*, 97(3):720–744.

Haruo Kubozono. 2017. Mora and syllable. *The handbook of Japanese linguistics*, pages 31–61.

Holger Mitterer, Eva Reinisch, and James M McQueen. 2018. Allophones, not phonemes in spoken-word recognition. *Journal of Memory and Language*, 98:77–92.

Sylvia Moosmüller. 1997. Diphthongs and the process of monophthongization in austrian german: a first approach. In *Proceedings of Eurospeech*, volume 97, pages 22–25.

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

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.

Arne Rubehn, Jessica Nieder, Robert Forkel, and Johann-Mattis List. 2024. Generating feature vectors from phonetic transcriptions in cross-linguistic data formats. *arXiv preprint arXiv:2405.04271*.

Hanna Kivistö de Souza and William Gottardi. 2022. How well can asr technology understand foreign-accented speech? *Trabalhos Em Linguística Aplicada*, 61(3):764–781.

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

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

A Appendix

A.1 Phonological feature vector table

Table 4 describes the phonological features of the selected phonemes [1]. The number of features may vary depending on the level of detail in characterizing the phonemes. Each phoneme has a phonological feature vector based on these features. Typically, when a feature is used for phonemic distinction, it is marked as + if present, - if absent, and 0 if the feature is not utilized for distinction. In this context, to represent binary features, + is denoted as 1, while both - and 0 are combined and represented as 0.

A.2 Tree diagrams for more cases

The main text only presents the diagram for cases where two new vowels are added to generate an allophonic consonant, but when only one vowel is added, the structure may differ slightly. Therefore, a tree to describe this can be constructed as follows. When a single vowel is added, it can be placed before or after the existing syllabic relationship as in Figure 4, or it can be inserted in the middle as in Figure 5.

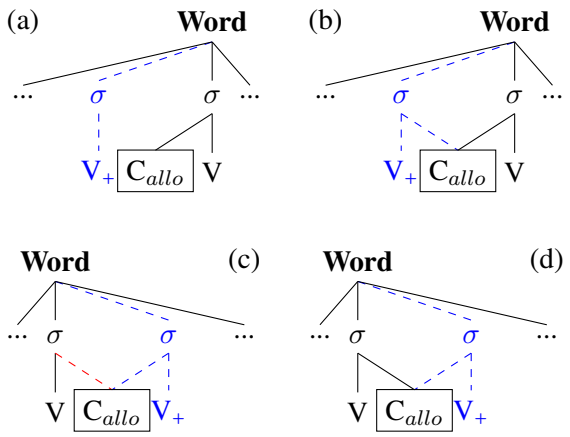


Figure 4: The tree diagram adds a new vowel V_+ before or after a specific consonant that changes to a new allophonic form C_{allo}. (a) a new singleton consonant in a CV structure. (b) a new geminate consonant in CV. (c) a new singleton in VC. (d) a new geminate in VC.

A.3 IPC-applied English Phonemes

Table 3 compares English phonemes with their IPC-applied counterparts, showing how IPC combines Korean phonemes to approximate English sounds. For vowels, the IPC approach combines similar

[1] https://github.com/anna-hope/phonemes

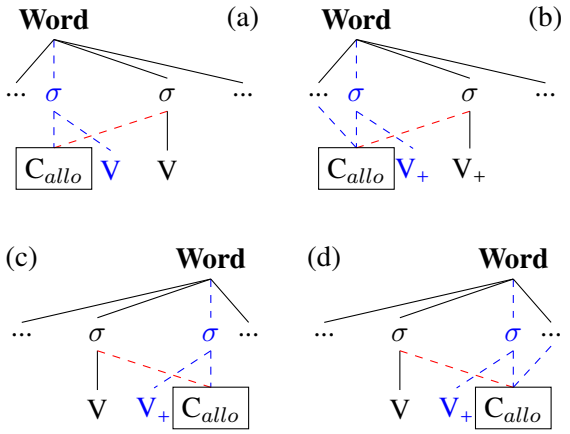


Figure 5: This inserts a new vowel V_+ in the middle of the existing syllabic relationship σ and generates a new allophone C_{allo}. (a) a new singleton consonant in a CV structure. (b) a new geminate consonant in CV. (c) a new singleton in VC. (d) a new geminate in VC.

Korean vowel sounds to approximate the target English phoneme. For example, the English vowel /æ/ is approximated using a combination of the Korean vowels for /ɛ/ and /ɐ/. Consonants are also adjusted through a combination of voicing, manner, and place of articulation. Voiced consonants, which become devoiced at word-initial in Korean, are represented by a combination with a preceding vowel to change the sound properly. Additionally, some consonants are adjusted for gemination to match the pronunciation more closely. In cases where the articulation position changes due to the addition of front vowels, this adjustment is also utilized (e.g., /ʃ/, /tʃ/, /dʒ/).

	English Phoneme	IPC-applied Phoneme
Vowel	/ɔ/	/o/(ㅗ) + /ʌ/(ㅓ)
	/oʊ/	/o/(ㅗ) + /ɯ(ㅡ)
	/æ/	/ɛ/(ㅐ) + /ɐ/(ㅏ)
	/ə/	/ɯ/(ㅡ) + /ʌ/(ㅓ)
Consonant (voicing)	/b/*	/ɯ/(ㅡ) + /p/(ㅂ)
	/d/*	/ɯ/(ㅡ) + /t/(ㄷ)
	/g/*	/ɯ/(ㅡ) + /k/(ㄱ)
	/dʒ/*	/ɯ/(ㅡ) + /tɕ/(ㅈ) + /y/(ㅟ)
Consonant (manner)	/l/*	/ɯ/(ㅡ) + geminate + /ɾ/(ㄹ)
	/m/*	/ɯ/(ㅡ) + geminate + /m^b/(ㅁ)
	/n/*	/ɯ/(ㅡ) + geminate + /n^d/(ㄴ)
Consonant (place)	/ʃ/	/s/(ㅅ) + /y/(ㅟ)
	/tʃ/	/tɕ^h/(ㅊ) + /y/(ㅟ)
	/dʒ/	/dʑ/(ㅈ) + /y/(ㅟ)

Table 3: English phonemes and corresponding ipc-applied phonemes. A superscript * on IPA symbols indicates phonemes that occur only in the initial position.

Feature	**t**	**d**	**s**	**z**	**ʃ**	**ʒ**	**p**	**b**	**f**	**v**	**k**	**g**	**h**	**tʃ**	**dʒ**	**m**	**n**	**l**	**ɹ**	**ɾ**	**i**	**ɪ**	**u**	**ʊ**	**e**	**ɛ**	**o**	**ɔ**	**a**	**æ**	**y**	**ʌ**	**ə**	**ɯ**
cons	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
son	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
syl	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
labial	0	0	0	0	0	0	1	1	1	1	0	0	0	0	0	1	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
round	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	1	0	0	1	0	0	0	0
coronal	1	1	1	1	1	1	0	0	0	0	0	0	0	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ant	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
dist	0	0	0	0	1	1	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
dorsal	0	0	0	0	0	0	0	0	0	0	1	1	0	1	1	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
high	0	0	0	0	0	0	0	0	0	0	1	1	0	1	1	0	0	0	0	0	1	1	1	1	0	0	0	0	0	0	1	0	0	1	1
low	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0
front	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	1	0	0	0	1	1	0	0	0	0
back	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	1	0	0	0	1	0	1	1
tense	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	1	0	1	0	0	1	1	0	0	1	1
pharyngeal	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
ATR	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	1	0	1	0	0	0	1	0	0	0	0
voice	0	1	0	1	0	1	0	1	0	1	0	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
SG	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CG	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
cont	0	0	1	1	1	1	0	0	1	1	0	0	1	1	1	0	0	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
strident	0	0	1	1	0	0	0	0	1	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
lateral	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
del_rel	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
nasal	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 4: Phonological features for selected phonemes.

A.4 Word Selection for the Experiment

We selected 20 English words that contain at least one IPC-applied phoneme outlined in Appendix A.3. Consonants were selected as allophones that vary in voicing, manner of articulation, and place of articulation depending on the neighboring vowel added. Table 5 presents the three types of visual cues provided to the participants.

A.5 Experiment participation

We recruited a total of **20** native Korean speakers, aged between 20 and 70, based on the criterion that they had not resided in an English-dominant environment before the age of 13, in accordance with the Critical Period Hypothesis. Of these participants, 12 were residing in Korea, and 8 were residing in Massachusetts, U.S, recruited through an anonymous platform. The study was embedded within educational and game activities, targeting individuals who sought to improve their English pronunciation. Participation was voluntary, with the option to withdraw at any time without penalty, and consent for the use of the collected data in research was obtained. Each participant received upfront compensation equal to twice the minimum hourly wage for the 1-hour session, and an additional reward was given to the highest improvement scorer in the game-based task to encourage motivation. The study was conducted in three sessions, focusing on recording the pronunciation of target phonemes in a controlled setting. Before recording the IPC-applied pronunciations, we gave 10-minute training session teaching how to read out the IPC-based Korean graphemes, with a single slide as provided in Figure 6. All the participants reported high satisfaction, noting that the gamified structure made the experience both engaging and enjoyable as well as their pronunciation improvement within 1 hour session.

The resulting dataset contains a total of 3,600 audio clips from 20 participants, each pronouncing 20 words. Under three different conditions, participants pronounced each word three times within a three-second window. We used the `python-sounddevice` [2] module to record the pronunciations, with a fixed sampling rate of 16,000 Hz to meet the input requirements of the evaluation models. To protect participants' privacy, no additional personal information was collected. The data were recorded and stored without names, addresses, or other identifiable details, ensuring that individual participants could not be traced or identified.

Figure 6: A single slide instructing how to read Korean graphemes based on inter-linguistic phonetic composition.

A.6 Evaluation Models

We adopted two ASR models, `Wav2Vec2Phoneme` and fine-tuned `XLSR-53` to transcribe the recorded speech of participants into English phonemes and acquire confidence scores during the decoding process. Both models are open-source with an apache-2.0 license, and listed as `facebook/wav2vec2-lv-60-espeak-cv-ft` [3] and `facebook/wav2vec2-xlsr-53-espeak-cv-ft` [4] at an online platform named huggingface. These models were fine-tuned on multi-lingual datasets to recognize phonemes in multiple languages. Both models do not have /g/ in their vocabulary, so we had to omit the confidence score and accuracy of /g/ from the statistical evaluation. Model configurations are specified in the configuration file, and every transcription process was conducted in a single M1 Macbook Air with 16GB Memory.

A.7 Potential Risks

The Inter-linguistic Phonetic Composition (IPC) approach poses several challenges in second-language (L2) acquisition. One concern is pronunciation distortion, as the synthesized sounds may not fully match the target L2 phonemes, leading to inaccurate or distorted articulation. Additionally, the method risks an oversimplification of phonological complexity by relying on native-language (L1) phonemes, which may obscure the subtle distinctions of L2 sounds, potentially limiting learners' ability to grasp the intricacies of the L2 phonological system. Finally, learner over-reliance on the

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

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

[4] https://huggingface.co/facebook/wav2vec2-xlsr-53-espeak-cv-ft

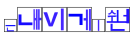
Word	IPA	Koreanized Transcription	IPC-based Korean Grapheme
dawn	/ dɔn /	돈	
jacket	/ dʒækɪt /	재킷	
loan	/ loʊn /	론	
rope	/ ɹoʊp /	로프	
ghost	/ goʊst /	고스트	
bowl	/ boʊl /	볼	
damage	/ dæmɪdʒ /	대미지	
nomad	/ noʊmæd /	노마드	
mortgage	/ mɔrgɪdʒ /	모기지	
broadcast	/ brɔdkæst /	브로드캐스트	
lawschool	/ lɔskul /	로스쿨	
minority	/ maɪnɔrɪti /	마이너리티	
mansion	/ mænʃən /	맨션	
chocolate	/ tʃɔkəlɪt /	초콜렛	
navigation	/ nævɪgɪʃən /	네비게이션	
revision	/ ɹəvɪʒən /	리비젼	
jamboree	/ dʒæmbəri /	잼버리	
deformation	/ dɪfɔrmeɪʃən /	디포메이션	
formulation	/ fɔrmjʊleɪʃən /	포뮬레이션	
gorgonzola	/ gɔrgənzɔlə /	고르곤졸라	

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

IPC framework could inhibit the development of authentic and natural L2 pronunciation skills over time.

A.8 Disclosure of AI Writing Assistance

We acknowledge the usage of the AI model GPT4o serviced by OpenAI [5] in writing codes within our research. However, the core algorithms of our proposed method and its evaluation were independently developed without AI assistance. Furthermore, we did not utilize AI for usage cases that should be disclosed: low-novelty text, new ideas, and new ideas with new text.

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