Unsupervised Discovery of Non-Categorical L2 Error Patterns Using Wav2Vec2.0 Code Vectors

Eunsoo Hong* Sunhee Kim† and Minwha Chung‡
* Seoul National University, Seoul, Korea
E-mail: stxlla13@snu.ac.kr Tel/Fax: +82-2-8806162
† Seoul National University, Seoul, Korea
E-mail: sunhkim@snu.ac.kr Tel/Fax: +82-2-8807693
‡ Seoul National University, Seoul, Korea
E-mail: mchung@snu.ac.kr Tel/Fax: +82-2-8809195

***Abstract*—L2 pronunciation is shaped by the interaction of two sound systems, which makes their identity more complex than a single phoneme category. The non-categorical nature demands assessment at a level finer than phonemes. As the granular requirement is highly labor-intensive, unsupervised methods emerged. Nevertheless, they either reverted to categorical diagnosis or used the supervised and phoneme-prescribed feature phonetic posterior-gram (PPG). Alternatively, this study adopts the unprescribed and unsupervised feature, the Wav2Vec2.0 code vector, to locate sub-phonemic variations. We first verify the features' L2 discernability by comparing their frequency across single-speaker data of L1 (CMU ARCTIC) and L2 (L2 ARCTIC). Clustering is performed on frequency vectors to test their separability on account of nativeness. Subsequently, sub-segmental patterns are analyzed among segmentally identical error samples in L2 Korean English NIA 037 data. After cataloging segmental errors detected by the model finetuned with L1 TIMIT, their corresponding code vector sequences are extracted by referencing the forced alignment result. We then derived dominant patterns of the sequences and compared them against L1 reference materials likewise constructed from TIMIT. Phoneme-code vector co-occurrence probability and code vector clustering were each used to check their attributes and uniqueness. The result confirmed the discernability, followed by linguistically interpretable common traits across patterns. (1) They formed a gradient error continuum along the changed articulatory value, reflecting the non-categorical nuanced understanding. (2) This trait is highlighted by intermediary typology, which assumed opposite values in two codebooks and was also rare in L1 for being L2 specific. Lastly, (3) distribution skewed towards the most approximate sound in the learner's L1, from which the patterns' complexity stems.**

I. Introduction

Articulation in L2 speech involves the mutual participation of the learner's native (L1) and target (L2) sound systems. Their complex interplay often spans the boundaries of two or more canonical phoneme categories. Cantonese English speakers, for instance, may utter a variation of [n] resembling both [l] and [n], mirroring their recent sound merger at the syllable initial position [1][2][3]. This between-categorical characteristic exists in a continuum, demanding gradient evaluation. Nevertheless, identifying subtleties of acoustic mismatch is an extremely laborious task. Unsupervised mispronunciation detection literature emerged under this setting, using the framework of acoustic pattern discovery[2][3][4][5][6][7][8][9][10]. If pattern discovery aims to extract recurring signals to substitute manual labeling [11], L2 errors could be set as one such target as they are recursive under systematic interaction. Nevertheless, the initial research focus reverted to categorical scope, as the obtained nuanced granular details were reduced to phonemic decoding [7][8] and binary decision-making [9][10]. The following [2][3][4][5][6] went beyond to describe the non-categorical attributes of error using the common analysis feature, phonetic posterior-gram (PPG). These posterior vectors were generated from MFCC-phoneme-label-trained neural networks, which contradicts two aspects. First, it is prescriptive of predefined phoneme thresholds, making the discovered pattern still tied to categorical circumscription. Requiring labels to map probability further notes its reliance on supervised learning. PPG is susceptible to the quality of labeled data used to train the instrumental model [2], which partially runs against the goal of sparing labor costs.

Concerning these drawbacks, this work adopts an alternative analysis feature, Wav2Vec2.0 code vectors. The choice is grounded on three factors. First, their generation process is fully unsupervised and unaffected by phonemic stipulation. Representation learning in self-supervised learning (SSL) does not accompany label training but rather self-discovers the operating units from the data's structural property. Second, this learning procedure conceptually resembles acoustic pattern discovery, which shares the goal of discovering lexically meaningful units to express speech content. As code vectors are already a rendition of discovered patterns, using them reinforces the research agenda. Third, linguistic probings in model documentation verify their phonetic relevancy. Ref. [12] proved that each feature specializes in representing different phonemes, while [13] showed that the usage pattern of multilingual pre-trained latent overlaps for close languages. The probing of [12] also implies that code vectors are more granular units than phonemes as the same phoneme is represented by multiple varieties.

With this understanding, we aim to survey the scope of variation capturable by code vectors to propose them as an

alternative means to discover non-categorical patterns. Accordingly, we first test the differentiability among L1 and L2 and subsequently use the feature to describe variations within each segmentally defined L2 error. This pattern discovery pipeline combines the aforementioned probing methods with frameworks of previous literature. To answer the question (1) Does the feature encode L2 acoustics differently from L1? codeword frequency probing in [13] is adopted. Featural usage of L1 and L2 speakers are compared through a frequency vector that maps occurrences of different code vector inventories. For question (2) How are non-categorical characteristics in L2 encoded in them? sequence analysis in [6] is applied to derive dominant patterns among segmentally identical error samples. The samples were pre-selected from the auxiliary segmental detection task. To interpret these patterns, two L1 reference materials are created. Co-occurrence-based phonetic probing in [12] is used to infer their phonetic attribute. KMeans clustering, a common similarity measure in [6][3][4][5], is performed on the union of available features to check relationships among discovered patterns.

II. Methods

A. Code Vector Inventory Probing

Code vectors' L2 discernability can be proved through the difference in the used inventory between L1 and L2 speech. Since product quantization in Wav2Vec2.0 concatenates two codebook entries, usage can be analyzed at the level of 1) individual entry and 2) entry pair. The former approach was taken in [13] to prove the feature's inter-language discernability. Among data of languages the model was pre-trained on, the degree of overlap in codeword distribution was in line with language similarity.

The same strategy is applied with the target of frequency calculation now set as individual L1 and L2 speakers. After segmenting speech by phoneme through forced alignment, the most representative codeword or codeword pair of the divided field was recorded as an index. At the entry-level illustrated in Fig. 1, codeword frequency vectors of size 1*320 are constructed per codebook whose values are normalized into probability. 320 is a size of codebook representation pre-defined by the model. These vectors are ultimately concatenated to a size of 1*640 for comparison. At the pair-level, the number of used pairs is not fixed, with a theoretical maximum reaching 102.4k i.e. 320*320 [12]. To plot the frequency in mutual space, we took the union of all available paired indices across the dataset and set them as counting bins to map occurrences. This amounted to 1*5712-sized vectors that were likewise normalized from the raw count.

KMeans clustering was eventually performed on each set of frequency vectors with the cluster number k set as 2. The expected result was to have L1 and L2 speaker data develop separate clusters, reflecting the usage pattern differences.

B. L2 Error Pattern Discovery

We next expand our survey to sub-segmental variations within L2. A list of error segments is first gathered through the

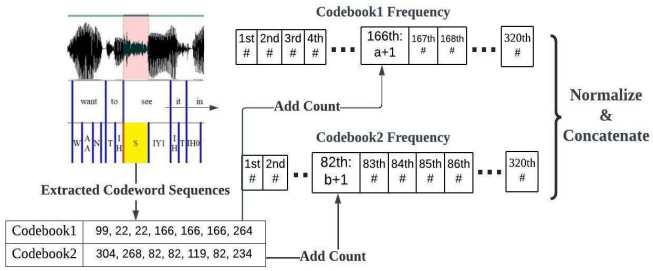


Fig. 1. Construction of Codeword Frequency Vectors. a and b are the hypothetical number of occurrences before adding count

standard practice of comparing recognized results with ground truth labels. The same pre-trained model used for code vector extraction was fine-tuned to create the recognition tool.

Then, code vector sequences of the error samples were retrieved as indices. The process first entails forced-aligning the L2 analysis data and spotting the time frames of falsely recognized phonemes. This was viable by employing the same phoneme set during forced alignment and grapheme-to-phoneme (G2P) conversion of fine-tuning labels. Subsequently, the sequence analysis method in [6] is applied to spot recurring patterns. The study [6] shared the goal of finding patterns within a canonically designated segment represented by summarizing indices. The analysis first involved filtering the sequence by removing minor presences and summing adjacent overlaps into one. Then, dominant patterns were selected based on frequency as their representation was once again refined by merging subsequences with their subsuming counterparts. In all analyzed cases, applying this procedure resulted in the 3 most dominant patterns, each represented by a single paired index.

To interpret these patterns, two reference materials were created from the same L1 data used for fine-tuning. We recorded 1) what phonemes each codeword (pair) statistically represents and 2) the raw numerical values of code vectors as opposed to indices. The former is a revisit to the phonetic probing in [12] that plotted the conditional probability of phoneme distribution into a graph. The same method of counting the feature's co-occurrence with human-annotated phoneme boundaries was executed, representing a cross-section of this graph along the vertical axis. Raw vectors were used to evaluate the distance among patterns. As our L1 data utilized 2202 pairs in total, the recorded 2202 concatenated vectors were clustered into 39 groups. The number 39 reflects the quantity of the mutual phoneme set as it serves as a minimum division criterion. Cluster IDs of the three dominant indices were compared to check if each discovered pattern formed a separate identity. If two or more indices belong to the same cluster, the patterns are merged into one. Fig. 2 illustrates the overall framework. While segmental detection is supervised, it is an optional step to limit the scope of sub-segmental analysis. In practice, the unsupervised discovery is applicable to any segment of interest.

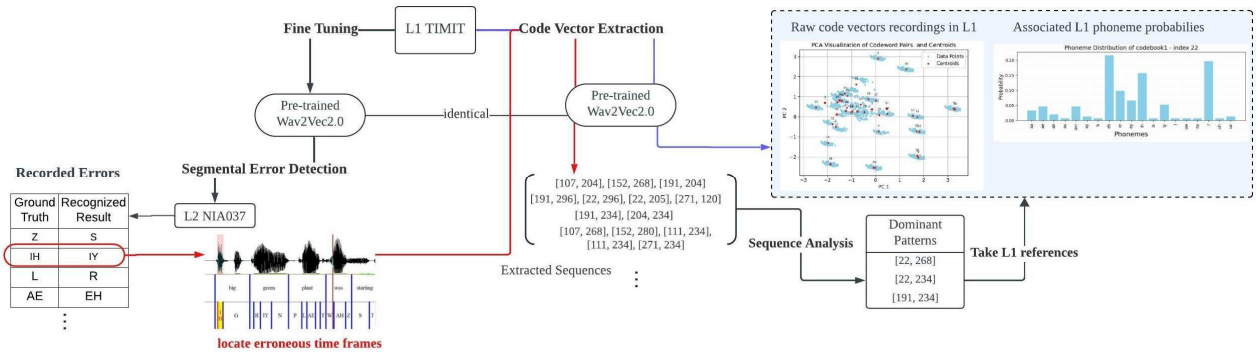


Fig. 2. Overview of L2 Error Pattern Discovery

III. Experiments

A. Data

Two different L1-L2 datasets were used for each task. For inventory comparison, an identical speech prompt across L1 and L2 was important to provide content-wise regulation. Speaker-level recordings were also needed to spot trends concerning a particular demographic. Accordingly, CMU ARCTIC (L1) and L2-ARCTIC (L2) v5.0 were chosen. For CMU ARTIC, only 5 speakers (bdl, rms, jmk, rms, clb) of North American accents were used as this research concerns pronunciations of US English. For pattern analysis, encompassing multiple speakers of diverse backgrounds was desirable to create robust speech recognition and references (L1), and make generalizable discoveries (L2). Manually confirmed phoneme duration was also preferred to extract code vector sequences accurately. Accordingly, TIMIT (L1) and NIA037, Korean Learners' L2 Speech corpus (English) developed by the National Information Society Agency (NIA), Korea, were used.

B. Implementation Detail

All SSL-related works and the model finetuning used fairseq **Framework**, a sequence modeling toolkit developed by the Facebook research group. The **Pretrained Model** used to extract code vectors was LARGE architecture trained on LibriVox (LV-60k) data. This is also the version phonetic probing in [12] was performed on. Using the TIMIT train split, **Finetuning** was conducted under the parameter of 40000 max update, 3e-4 learning rate, mask probability 0.65, mask channel probability 0.5. The final validation phoneme error rate was 1.63%. The mutual phoneme set used for finetuning label and forced alignment was the 39 ARPAbet symbols. For **G2P**, the set was derived by excluding the sentence stress numeric from the g2p tool kit [1]. For **Forced Alignment**, excluding the stress marker in English (US) ARPA dictionary v3.0.0 of Montreal Forced Aligner resulted in the same union of notation. The corresponding version of the acoustic model was also used.

[1]https://github.com/Kyubyong/g2pGitHub - Kyubyong/g2p: g2p: English Grapheme To Phoneme Conversion

Clustering of paired vectors and **Visualization** of individual codewords used FAISS library.

IV. Results

A. L1 to L2 Code Vector Usage Comparison

Fig. 3 displays the clustering results after performing the principal component analysis. Speakers of the frequency vectors are annotated next to the plotted location, and each element is color-coordinated according to its cluster membership. As can be noted from the color scheme, the cluster grouping coincides with the division between native and non-native speech. The concatenated format offers a clearer separation, allowing for a more nuanced examination of subtle pronunciation deviations. This motivates us to conduct detailed cross-speaker comparisons in a paired setting.

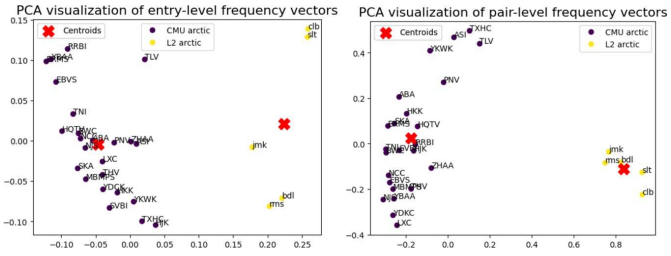
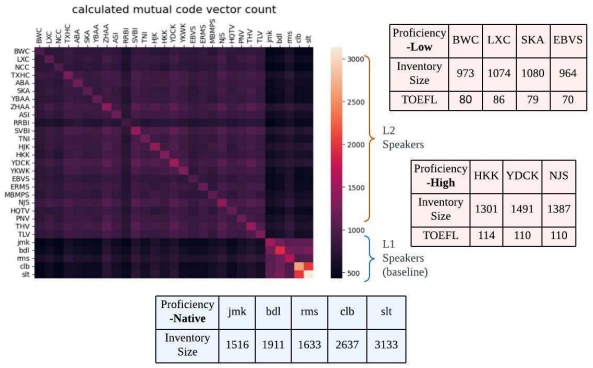


Fig. 3. Frequency Vector Clustering Results

The heatmap in Fig. 4 counts mutual codeword pair inventory between two speakers. Reflecting the former clustering result, the shared amount between native and non-native speakers is far below the rate within each speaker group. This sharing rate asymmetry is further marked by higher numbers of utilized pairs in L1. The left L2-ARCTIC grids have darker shades associated with fewer tokens compared to the rightmost five grids of CMU ARCTIC speakers. Interestingly, the inventory size increases as a function of the speaker's language proficiency. From the demographic information provided in [14], we have selected two speaker groups of opposing proficiency levels. The lower-level group had TOEFL iBT scores ranging under 90, whereas the higher-level group had scores over 110. The total usage-

-counting on the right shows that higher-level speakers overall use more paired tokens than lower-level counterparts. One possible explanation is that to fully utilize representations encoding English, one has to be phonetically aware of the sounds in the language. This awareness is marked by the utilization ratio of the code vector inventory acquired in the L1 standard. If the amount used by the native speakers is the full range of available acoustics, the less adept one is at articulating these sound units, the less amount of inventory in use there will be.



Proficiency -Low	BWC	LXC	SKA	EBVS
Inventory Size	973	1074	1080	964
TOEFL	80	86	79	70

Proficiency -High	HKK	YDCK	NJS
Inventory Size	1301	1491	1387
TOEFL	114	110	110

Proficiency -Native	jmk	bdl	rms	clb	slt
Inventory Size	1516	1911	1633	2637	3133

Fig. 4. Code Vector Inventory Counting

B. L2 Error Pattern Discovery Result

Subject of Gradience Our analyzed segmental errors concerned 5 different substitution routes. Patterns of each trajectory could be scaled by the assumed degree of changed articulatory value, forming a gradient error continuum. *Substitution of voicing identity*: in Z to S, the second highest probability in native phoneme distribution moved from sh[-voicing] to z[+voicing]. *Substitution of manner of articulation*: in DH to D, V to B, and F to P, native phoneme distribution moved from plosives and silence[-continuity] to fricatives, vowels, and approximants[+continuity]. This was a recurring dynamic across all three examples. *Substitution of laterality*: in L to R and R to L, native phoneme distribution moved from having more lateral[+laterality] to rhotic[-laterality] association (or vice versa). *Substitution of vowel height*: in AH to AA, the association ratio moved from allotting higher probability to AA[-high] to AO[+high]. *Substitution of tenseness*: in IH to IY, IY to IH, AE to EH, and EH to AE, tenseness calculation reflected Euclidean distance among raw code vectors. Fig. 6 charts these gradient movements. The intermediary patterns were decomposed into two codewords for the reason that will be explained below.

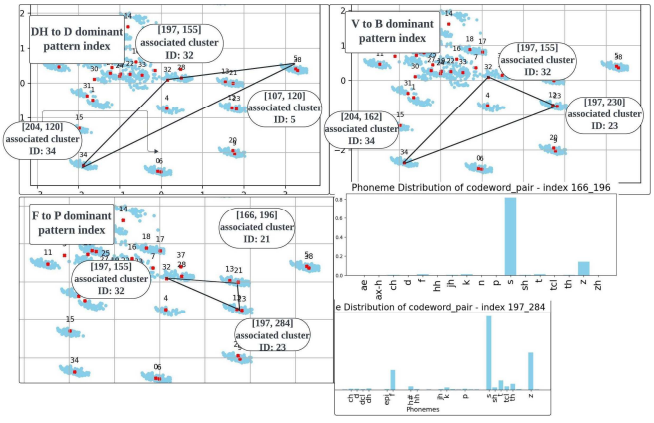


Fig. 5. Fricative to Plosive Dominant Pattern Dynamics

Intermediary Typology Patterns at the intermediate position were non-categorical by assuming opposite identities in two codebooks. If one displayed a positive value of the changed articulatory trait, the other displayed a negative value. Such contradictory pairing was rare in L1 data, attesting to the L2 particular nature of non-categoricity. In *Z to S*, intermediary [18, 51] had codebook1 bearing [-voicing], while codebook2 bearing [+voicing] identity. The paired index had no native speech presence. In *DH to D*, intermediary [204, 120] had codebook1 bearing [+continuity], while codebook2 bearing [-continuity] identity. It records a single co-occurrence with nasal [n], a sound that exemplifies duality with airflow obstructed in the oral cavity yet unhindered in the nasal cavity. In *R to L*, intermediary [191, 235] had codebook1 bearing [-laterality], while codebook2 bearing [+laterality] attribute. The pair had only three L1 occurrences. In *AH to AA*, codebook1 and 2 each assumed [-high] and [+high] identity. The middling pink figures in Fig. 6 showcase this ambivalence. Note that the ranking is comparative.

Distributional asymmetry Patterns were skewed toward the most approximate sound available in the learner's L1. This finding could be viewed at two levels: 1) dispersion rate across different substitution types and 2) within-substitution pattern distribution. The case of fricative to plosive substitution demonstrates the first view. Fig. 5 plots the distance among subsegmental patterns through their associated cluster IDs. Accordingly, *F to P* involves smaller between-pattern distances than the other two. The convergence gears towards the [+continuity] end as [197, 155] that once formed the [+continuity] end of the spectrum in *DH to D* in Fig. 6 is now the pattern bearing the least characteristic of continuity. The other two indices, [166, 196] and [197, 284], show an even higher association with the ground truth canonical value. The asymmetry is related to Korean fricative inventory consisting of voiceless but not voice fricatives, leading to fewer difficulties articulating a familiar phonation type. Namely, the more difficult target DH and V incurs greater dispersion. Meanwhile, the skewed distribution within substitution can be found in the case of liquid. Fig. 7 likewise visualizes the pattern dynamics with cluster IDs. The laterality spectrum is additionally superimposed by referencing Fig. 6. In *R to L*, the intermediary [191, 235] is closer to the [+laterality] pattern [204, 162] than [22, 101]. In *L to R*, two dominant patterns [191, 212] and [191, 162] merge into a single cluster that gears towards the same direction. The concentration on the

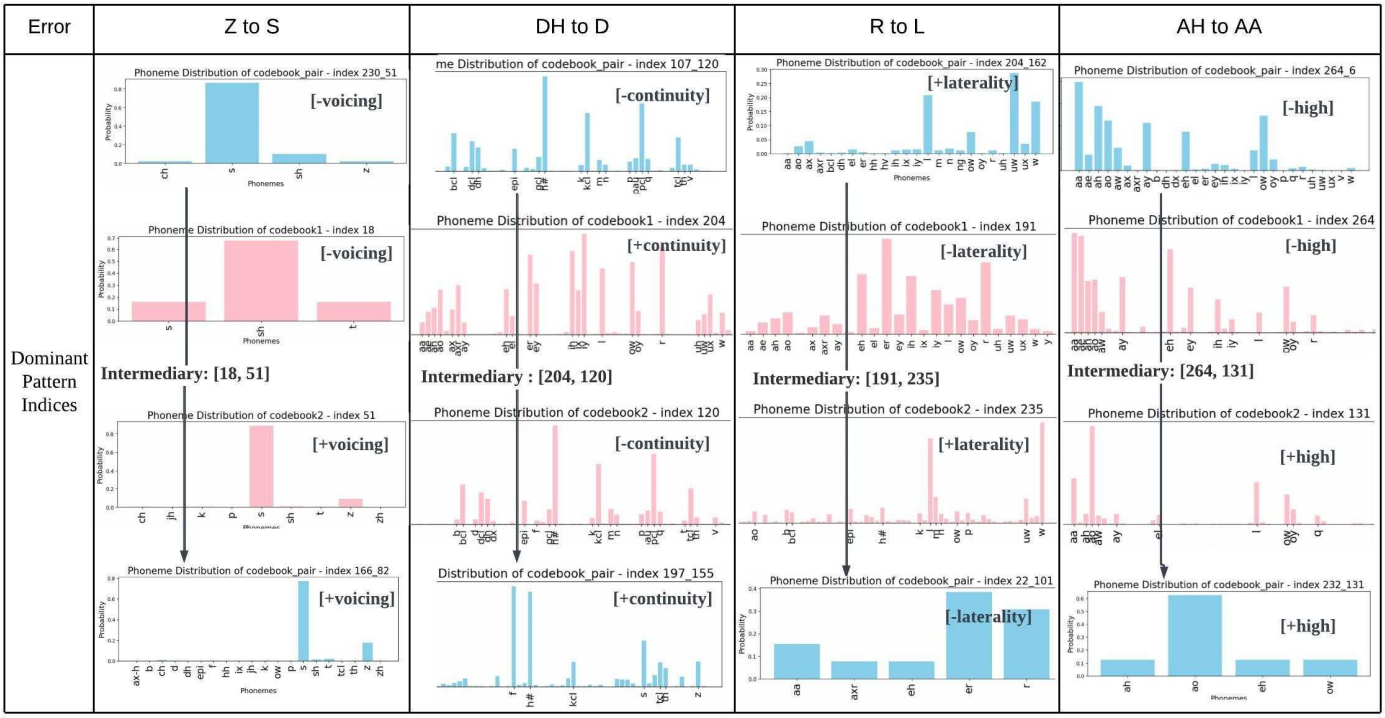


Fig. 6. Gradience of Error Continuum. Indices with numerous phoneme associations have been cropped for illustration purposes

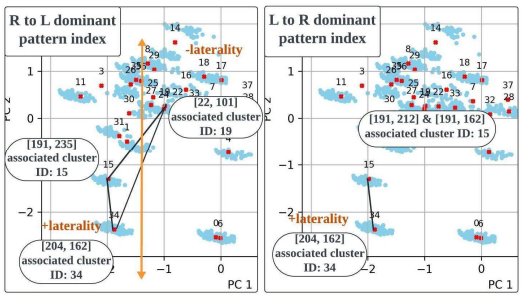


Fig. 7. Pattern Distribution in Liquid Substitution

[+laterality] end reflects greater difficulties in producing the rhotic variety, bunched r, unobserved in Korean compared to the existing lateral inventory.

V. Discussion

The three overarching findings are linguistically interpretable. First, it is expected for the variation spectrum to develop along the articulatory trait responsible for changes.

Second, since the non-categorical nature is L2-specific, the conflicting pairings were associated with non-mutual inventory between L1 and L2 speakers. As frequency probing confirmed the used inventory difference and its expanding range alongside articulation adeptness, one can expect that the more pronunciation deviation there is, the higher the chance for the speech to utilize these L1-unobserved non-mutuals. Accordingly, English front vowels reported to be particularly difficult for Korean

TABLE I
Dominant Index Pairs of Front Vowel Substitutions

IH to IY	[22,268] (1)	[191, 234]	[22, 268]
IY to IH	[191, 234] (4)	[22, 234]	[42, 234] (1)
AE to EH	[22, 268]	[42, 234]	[191, 234]
EY to EH	[42, 234]	[22, 234] (unobserved)	[22, 319]

learners in [15] commonly used code vector varieties rare in native speech. Their total count in TIMIT is written in parenthesis in Table I. Ref. [15] also notes how the difficulty is manifested as reduced frontal vowel space, when the articulatory confusion transfers to indistinguishability. In line with this understanding, patterns in Table. I experience overlap.

Third, relationships among sub-segmental patterns reflect the acoustic distance expected from L2 variations. The skewed distribution first implies that non-categorical traits are ultimately attributed to the learners' L1 influence. This between-categorical position of phonetic understanding further concurs with the numerical calculation between vectors. As our detected tenseness substitutions solely concerned front vowel pairs, they lacked enough representative samples for pair-level analysis. Hence, we have resorted to examining attributes of individual codewords, which revealed that the rarity stems from conflicting tenseness identity. That is, codebook1 displayed the identity of laxness while codebook2 displayed tenseness. This identity could be instantiated by calculating the tense-to-lax ratio of associated phoneme probabilities. Patterns can then be ranked on a tenseness spectrum with values multiplied from two codebooks. Inferred pattern-wise distance, hereafter, aligned with Euclidean distance. We have demonstrated this process with

a case of IH to IY substitution in Fig. 8.

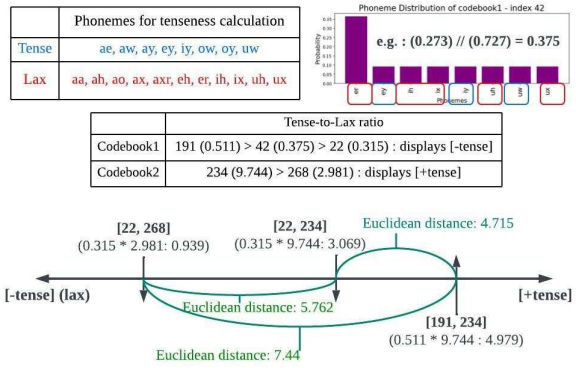


Fig. 8. Pattern Positioning in IH to IY. Although [22, 234] is unobserved, the distance between [22, 234] and [22, 268] is calculable through the second codeword vectors.

VI. Conclusions

[1] This work recreated the unsupervised L2 error pattern discovery experiments using an SSL representation, the Wav2Vec2.0 code vector. Leveraging its unprescribed yet phonetically relevant status, we have identified the range of variations within a single segmental error. While the inventory probing confirmed that encoding units of L2 differed from L1, the pattern discovery revealed that this difference results from an unlikely codebook combination in L2. Namely, the two codeword vectors assumed conflicting identities that formed opposite ends of the error pattern spectrum. As this conflict instantiates L2-specific non-categorical traits, increased usage of relevant index pairs coincided with a higher degree of phonetic divergence. Moreover, sub-segmental pattern distribution reflected the acoustic proximity of corresponding L1 phonemes to two segments participating in substitutions. Thus, the way Wav2Vec2.0 code vectors encode L2 variation is phonetically relevant, making it a valid tool to uncover sub-segmental gradience with the ability to quantify its details. Ultimately, the uncovered details can be employed for finer judgment and feedback that reflects the true nature of mispronunciations.

References

[1] C. L. C. Ng, "Merger of the syllable-initial [n-] and [l-] in hong kong cantonese," 2017.

[2] X. Li, X. Wu, X. Liu, and H. Meng, "Deep segmental phonetic posterior-grams based discovery of non-categories in l2 english speech," *arXiv preprint arXiv:2002.00205*, 2020.

[3] S. Mao, X. Li, K. Li, Z. Wu, X. Liu, and H. Meng, "Unsupervised discovery of an extended phoneme set in l2 english speech for mispronunciation detection and diagnosis," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2018, pp. 6244–6248.

[4] Y.-B. Wang and L.-S. Lee, "Toward unsupervised discovery of pronunciation error patterns using universal phoneme posteriorgram for computer-assisted language learning," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, 2013, pp. 8232–8236.

[5] Y.-B. Wang and L.-s. Lee, "Supervised detection and unsupervised discovery of pronunciation error patterns for computer-assisted language learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 3, pp. 564–579, 2015.

[6] X. Li, S. Mao, X. Wu, K. Li, X. Liu, and H. Meng, "Unsupervised discovery of non-native phonetic patterns in l2 english speech for mispronunciation detection and diagnosis.," in *INTERSPEECH*, 2018, pp. 2554–2558.

[7] A. Lee and J. Glass, "Mispronunciation detection without nonnative training data," in *Proc. Interspeech 2015*, 2015, pp. 643–647.

[8] A. Lee, N. F. Chen, and J. Glass, "Personalized mispronunciation detection and diagnosis based on unsupervised error pattern discovery," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 6145–6149.

[9] A. Lee and J. Glass, "A comparison-based approach to mispronunciation detection," in *2012 IEEE Spoken Language Technology Workshop (SLT)*, 2012, pp. 382–387.

[10] A. Lee, Y. Zhang, and J. Glass, "Mispronunciation detection via dynamic time warping on deep belief network-based posteriorgrams," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 8227–8231.

[11] A. S. Park and J. R. Glass, "Unsupervised pattern discovery in speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 186–197, 2007.

[12] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[13] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised cross-lingual representation learning for speech recognition," *arXiv preprint arXiv:2006.13979*, 2020.

[14] G. Zhao, E. Chukharev-Hudilainen, S. Sonsaat, *et al.*, "L2-arctic: A non-native english speech corpus," 2018.

[15] Y.-J. O. Hee-San Koo, "An Analysis of English Vowels of Korean Learners of English and English Native Speakers," *Korean Education Inquiry*, vol. 16, pp. 1–12, 2001.

[16] E. Hong, "Unsupervised Discovery of Non-Categorical L2 Error Patterns Using Wav2Vec2.0 Code Vector Features," M.A. thesis, College of Hum., Seoul National Univ., Seoul, Aug. 2024.

[1] This paper is based on a part of the first author's 2024 master's thesis [16]