Deep Supervised Contrastive Learning of Pitch Contours for Robust Pitch Accent Classification in Seoul Korean

Hyunjung Joo[1,3], GyeongTaek Lee[2]*
[1]Department of Linguistics, Rutgers University
[2]Department of Smart Factory, Gachon University
[3]Hanyang Institute for Phonetics and Cognitive Sciences of Language (HIPCS)
hyunjung.joo@rutgers.edu, leegt@gachon.ac.kr

Abstract

The intonational structure of Seoul Korean has been defined with discrete tonal categories within the Autosegmental-Metrical model of intonational phonology. However, it is challenging to map continuous F_0 contours to these invariant categories due to variable F_0 realizations in real-world speech. Our paper proposes Dual-Glob, a deep supervised contrastive learning framework to robustly classify fine-grained pitch accent patterns in Seoul Korean. Unlike conventional local predictive models, our approach captures holistic F_0 contour shapes by enforcing structural consistency between clean and augmented views in a shared latent space. To this aim, we introduce the first large-scale benchmark dataset, consisting of manually annotated 10,093 Accentual Phrases in Seoul Korean. Experimental results show that our Dual-Glob significantly outperforms strong baseline models with state-of-the-art accuracy (77.75%) and F1-score (51.54%). Therefore, our work supports AM-based intonational phonology using data-driven methodology, showing that deep contrastive learning effectively captures holistic structural features of continuous F_0 contours.

1 Introduction

Seoul Korean is an edge-prominence language (Jun, 1998, 2005), where fundamental frequency F_0, the acoustic correlate of pitch, is used to organize prosodic structures to encode grammatical and pragmatic distinctions. Within the Autosegmental-Metrical (AM) theory of intonation (e.g., Beckman and Pierrehumbert, 1986; Ladd, 2008) and its Korean Tones and Break Indices (K-ToBI; Jun, 2000) transcription system, a continuous F_0 contour is modeled as a sequence of discrete tonal targets such as Lows (L) and Highs (H), which are interpolated with one another.

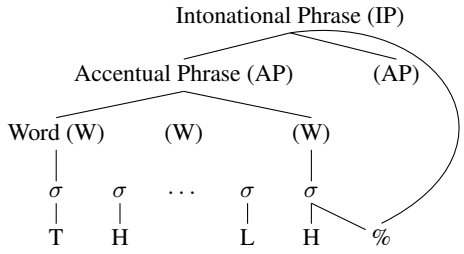


Figure 1: Intonational structure of Seoul Korean (Jun, 1998). The AP-initial tone (T) is realized as H for aspirated and tense consonants, otherwise L. The % symbol refers to a boundary tone (e.g., L% or H%) at the end of an IP.

The intonational structure of Seoul Korean is hierarchically organized, with an Accentual Phrase (AP) as the basic unit, one or more of which are grouped into an Intonational Phrase (IP). According to Jun (1998), APs with more than three syllables typically surface as LHLH or HHLH, depending on the phrase-initial segment: APs beginning with aspirated or tense consonants surface as HHLH, while others surface as LHLH (Figure 1). Shorter APs show fourteen possible tonal patterns: LH, HH, LL, HL, LLH, LHH, HLH, LHL, HHL, HLL, LHLL, HHLL, LHLH, and HHLH (See **Appendix A** for the schematic contours of these patterns.).

Despite this well-established theoretical characterization of intonation, there is a significant gap between phonological models (Jun, 1998) and real-world acoustic data. Crucially, it is highly complicated to map continuous F_0 contours onto discrete and invariant tonal categories due to the inherent variability of F_0 coming from gender differences, speech styles, and phonetic contexts (e.g., Cole and Shattuck-Hufnagel, 2016). While intonational research has been conducted based on expert annotations using the ToBI system (Beckman and Hirschberg, 1994), such perception-based transcriptions may fall into subjective bias and are difficult to deal with large-scale data for computational mod-

* Corresponding author.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 39602–39621

eling.

To fill this gap, we propose a **Dual-Glob**, a deep supervised contrastive learning approach to characterize continuous F_0 variations into fine-grained and invariant pitch accent categories in Seoul Korean. Unlike traditional supervised models that are prone to overfitting and sensitive to measurement noise, our contrastive framework enables the model to learn robust and discriminative representations of raw F_0 contours. Therefore, we maximized the similarity between F_0 contour shapes within the same pitch accent category while contrasting them across different pitch accent categories in the latent space.

This study supports the discrete tonal categories defined within AM theory with the representional power of data-driven deep learning approaches. By using large-scale acoustic data, we show that our model effectively captures fine-grained pitch accent categories in Seoul Korean. Our proposed framework offers a more robust and scalable approach to classifying the pitch accent patterns than conventional models.

The contributions of our work are summarized as follows:

- **Large-Scale Prosodic Benchmark Dataset:** We construct the first large-scale benchmark dataset for classifying pitch accent categories in Seoul Korean, paving the way for future intonational research.
- **Connecting AM Theory with Deep Learning Framework:** To our knowledge, we propose the first deep learning model that account for the AM's discrete tonal representation with holistic F_0 contours. This approach enriches AM-based intonational phonology with continuous and data-driven representation learning.
- **State-of-the-Art Performance:** Experimental results show that our **Dual-Glob** method significantly outperforms other competitive baseline models with state-of-the-art accuracy in characterizing continous F_0 contours.

2 Related Work

Intonational Structure of Seoul Korean The intonational structure of Seoul Korean has been mainly formalized within the AM framework (e.g., Beckman and Pierrehumbert, 1986; Ladd, 2008), with an AP as the basic unit of tonal organization. (Jun, 1998, Jun, 2005). (Jun, 2000) identified up to fourteen AP tonal patterns (e.g., HHLH, LH), conditioned by the laryngeal features of the phrase-initial segment and syllable count. Studies have empirically shown that the tonal patterns of APs in Seoul Korean exhibit strong edge-prominent characteristics (e.g., Hatcher et al., 2024; Kim, 2008). Importantly, Kim (2008) found that in both read and radio speech, most APs begin with a rising tone and end with a high tone (LH...LH) to signal prosodic edges for word segmentation. Despite this basic pattern, other tonal patterns were distributed variably across registers: radio speech showed a more varied distribution than read speech.

While AM theory has been a mainstream theoretical framework in intonational phonology, configurational approaches (e.g., Bolinger, 1951; Hart et al., 2003; Xu, 2005) view intonation as a holistic F_0 contour shape rather than a sequence of discrete tonal targets. Interestingly, recent studies have emphasized the importance of incorporating continuous F_0 information (Barnes et al., 2012, 2021; Joo and D'Imperio, 2025), suggesting that only considering discrete tonal targets within AM theory may be insufficient to fully capture the intonational patterns.

Intonational Modelling Inspired by configurational approaches, Levow (2005) used uniform representations of pitch, duration, and intensity within a support vector machine to classify Mandarin tones and English pitch accents, while also considering the preceding and following shape contexts. However, these methods only focused on discrete features, which are not enough to capture fine-grained dynamic patterns of F_0.

Recent work has modeled intonation as a continuous F_0 contour. The dynamical systems approach models English pitch accents as nonlinear trajectories toward phonological targets (Iskarous, 2017; Iskarous et al., 2023; Iskarous and Cole, 2026). However, it uses predefined parameters (e.g., F_0 peak and velocity) in a differential equation, rather than learning representations directly from raw F_0 time series.

In contrast, recent deep learning approaches have shown that directly modeling raw F_0 of Mandarin tones can outperform conventional feature-based methods (Chen et al., 2022).

Deep Learning Approaches in Korean Prosody. Recent deep learning approaches use F_0 contours for tasks like speech emotion recognition using

dual recurrent encoder (Yoon et al., 2018), or dialect identification via Bidirectional long short-term memory (BiLSTM) (Lee et al., 2021). However, these works largely focus on boundary detection or broad regional categories. To our knowledge, we are the first to apply deep contrastive learning to the fine-grained classification of tonal patterns in Seoul Korean (e.g., LHLH vs. HH), explicitly modeling holistic F_0 contour shapes.

Contrastive Learning for Robust Representations. While models like InceptionTime (Ismail Fawaz et al., 2020) handle time-series, their reliance on a large amount of labeled data limits scalability. Self-supervised predictive coding (Oord et al., 2018) offers an alternative but rests on Markovian assumptions not suitable for holistic tonal representations (e.g., HHLH). In contrast, supervised contrastive learning (SupCon) (Khosla et al., 2020) can optimize for the consistency of global F_0 contours. By clustering the same class samples in the latent space, SupCon learns invariant tonal representations robust to measurement noise and speaker variability.

3 Proposed Model

3.1 Motivation: Holistic Tonal representations vs. Temporal Prediction

Standard self-supervised time-series methods often rely on predictive coding, predicting future segments from past contexts (e.g., $x_{past} \rightarrow x_{future}$). While this works well for stochastic processes, we argue that it is suboptimal for modeling Seoul Korean pitch accent patterns.

In AM theory, APs are not Markovian sequences but **discrete tonal representations** where the entire contour determines linguistic meaning. Thus, treating intonation as a local predictive task risks fragmenting these global units. To address this, we propose a **holistic representation learning** framework by capturing the global shape of F_0 contours, at the same time, making distinct representations contrastive despite local temporal variations.

3.2 Dual-View Supervised Contrastive Learning

To robustly learn tonal representations, we employ a dual-branch architecture (Figure 2) that processes both original and augmented views via shared encoders. This design enforces *consistency regularization*, ensuring the model recognizes underlying tonal categories even under input perturbations.

3.2.1 Data Augmentation and Encoding

Given a clean input x_c, we generate an augmented view x_a via stochastic perturbations to simulate natural variability. Both x_c and x_a are processed by a shared encoder $E(\cdot)$ and projection head $P(\cdot)$ to yield:

$$z_c = P(E(x_c)), \quad z_a = P(E(x_a)) \tag{1}$$

These projections map F_0 contours into a normalized latent space for contrastive learning.

3.2.2 Objective Function

We employ the SupCon loss to structure the latent space. Specifically, we formulate a joint objective with two terms to balance intra-class compactness on clean data and robust representation learning against perturbations.

1. Clean-view SupCon ($\mathcal{L}_{Clean}$): This term ensures that the representation of the original F_0 contour correctly aligns with other instances of the same tonal category. Formally, the loss is defined as:

$$\mathcal{L}_{Clean} = -\sum_{i \in \mathcal{B}} \frac{1}{|P(i)|} \sum_{j \in P(i)} \log \frac{\exp(z_{i,c} \cdot z_{j,c}/\tau)}{\sum_{k \in \mathcal{B} \setminus \{i\}} \exp(z_{i,c} \cdot z_{k,c}/\tau)} \tag{2}$$

where $z_{i,c}$ is the clean projection of the i-th sample, $P(i)$ is the set of indices of positive samples in batch $\mathcal{B}$, and τ is a temperature parameter.

2. Augmented-view SupCon ($\mathcal{L}_{Aug}$): This term addresses the inherent instability of pitch extraction in real-world environment. Raw F_0 contours are accompanied by discontinuations (e.g., breath, plosives) and pitch tracking errors. Models trained solely on raw F_0 contours risk overfitting to surface-level irregularities. To mitigate this, $\mathcal{L}_{Aug}$ enforces invariant representation learning, enabling the model to ignore noise and capture robust phonological representations. Formally, the loss minimizes the discrepancy between augmented views and clean positives:

$$\mathcal{L}_{Aug} = -\sum_{i \in \mathcal{B}} \frac{1}{|P(i)|} \sum_{j \in P(i)} \log \frac{\exp(z_{i,a} \cdot z_{j,c}/\tau)}{\sum_{k \in \mathcal{B} \setminus \{i\}} \exp(z_{i,a} \cdot z_{k,c}/\tau)} \tag{3}$$

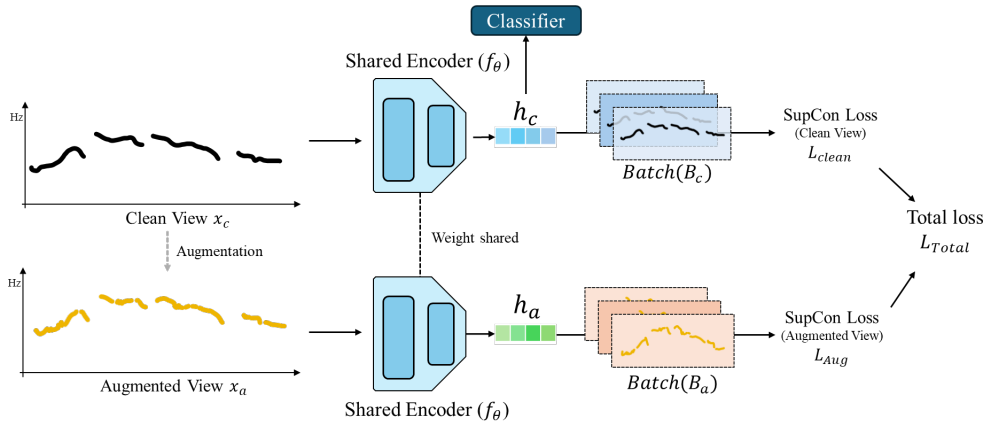


Figure 2: Overview of the proposed **Dual-Glob** framework. The model processes entire F_0 contours via parallel clean (x_c) and augmented (x_a) views using a shared encoder. A composite supervised contrastive objective ($\mathcal{L}_{Total}$) enforces structural consistency across both views to learn robust representations.

Here, $z_{i,a}$ denotes the embedding of the augmented anchor, while $P(i)$ is the set of indices for clean samples belonging to the same class as i. The terms $z_{j,c}$ and $z_{k,c}$ represent these clean positive prototypes and the entire pool of clean embeddings in the batch, respectively. This asymmetric formulation explicitly encourages the model to project distorted representations onto the stable manifold formed by the clean signals, effectively denoising the intonational features.

The final training objective integrates both clean-signal consistency and augmented-view robustness. We define the total loss as a weighted sum of the two components:

$$\mathcal{L}_{Total} = \lambda_1 \mathcal{L}_{Clean} + \lambda_2 \mathcal{L}_{Aug} \quad (4)$$

where λ is a balance parameter. By minimizing this total loss, the model learns to capture the global F_0 contour shape (e.g., HHLH), ignoring small local errors that often mislead predictive models.

3.2.3 Downstream Classification

After training, we freeze the encoder and discard projection heads to extract latent pitch features. These are fed into standard classifiers, ensuring that performance reflects robustness of representation rather than classifier complexity.

4 Experiments

4.1 Dataset and Preprocessing

We constructed a new dataset of APs in Seoul Korean[1], using the broadcasting conversational data, a large-scale corpus provided by AI Hub (National Information Society Agency, 2022). In order to make sure clear prosodic realization with precise articulation, we selected the recordings produced by 18 professional broadcasters (11 females, 7 males).

The raw audio was manually segmented into APs based on perceptual judgment and visual inspection of F_0 contours, since automated forced alignment often fails to detect prosodic boundaries and varying speech rates can result in different phrasing (Jun, 2003).

Each AP was then annotated by two trained K-ToBI (Jun, 2000) transcribers. A total of 10,093 APs were categorized into 16 distinct tonal patterns, including monosyllabic APs with either a L or an H tone. The distribution of these categories is provided in Table 1.

Feature Extraction We extracted F_0 contours using the pYIN algorithm from 22.05 kHz audio. Frame and hop lengths were set to 1024 and 256, with a range of 80–400 Hz. All sequences were fixed to 200 frames ($T = 200$) for consistency.

Normalization Since absolute pitch varies by speaker (e.g., gender), raw F_0 introduces bias. To mitigate this, we applied speaker-wise Min-Max normalization: $x' = \frac{x - \min_k}{\max_k - \min_k}$, scaling pitch values to the range $[0, 1]$.

4.2 Experimental Design

Baselines. We compared our framework against diverse competitive models. First, we employed standard sequence encoders: **1D-CNN** (Wang et al., 2017), **BiLSTM** (Schuster and Paliwal, 1997;

[1]Our dataset isavailable at our GitHub repository

Table 1: Distribution of the 16 tonal labels in the dataset.

Label	Count	Label	Count
H	200	L	15
HH	1,168	LH	1,318
HHL	77	LHH	1,084
HHLH	1,463	LHL	81
HHLL	784	LHLH	2,705
HL	57	LHLL	431
HLH	259	LL	8
HLL	26	LLH	417
Total	**10,093**		

Hochreiter and Schmidhuber, 1997), and **Transformer** (Vaswani et al., 2017). Second, we benchmarked against state-of-the-art time-series models: **InceptionTime** (Ismail Fawaz et al., 2020), **TimesNet** (Wu et al., 2022), and **DLinear** (Zeng et al., 2023), which specialize in trend decomposition. Finally, we included **MiniRocket** (Dempster et al., 2021) as an efficient non-deep learning baseline.

Evaluation Protocol. For the proposed model, we froze the pre-trained encoder and evaluated the representations using **LightGBM** (Ke et al., 2017), **logistic regression (LR)** (Cox, 1958), and **random forest (RF)** (Breiman, 2001). All experiments were conducted using 5-fold cross-validation. For the performance measurement, we employed both **accuracy (Acc)** and **macro-F1 score (F1)**. The detailed experimental design and the architectural specifications for our model and all comparative baselines are provided in **Appendix B**. In addition, we employed a stochastic data augmentation strategy for contrastive learning. Detailed information regarding this procedure is provided in **Appendix C**.

4.2.1 Ablation Study Design

To validate the synergy between augmentation and training objectives, we evaluated representation quality using classifiers trained on frozen features. We denoted predictive loss as $\mathcal{L}_{Pred}$, with superscripts $Clean$ and $Augment$ indicating clean and augmented views. The variants were categorized as follows:

- **Local Predictive Models:** To assess the impact of learning local temporal transitions, we evaluate two variants of predictive coding. **Pred-C** (based on Vanilla SimTS (Zheng et al., 2023)) employs a unidirectional contrastive loss ($\mathcal{L}_{Pred}^{Clean}$) that predicts the future half of a pitch sequence from its past, using only clean data. In contrast, **Pred-A** extends this approach by utilizing bidirectional prediction between clean and augmented views ($\mathcal{L}_{Pred}^{Augment}$). Specifically, it simultaneously optimizes the prediction of the clean future from an augmented past, and vice versa, to enforce stronger local consistency across stochastic perturbations.

- **Global Contrastive Models: Glob-Clean** and **Glob-Augment** apply contrastive objectives to clean ($\mathcal{L}_{Clean}$) and augmented ($\mathcal{L}_{Aug}$) sequences, respectively, to capture holistic structural features. Each objective follows the contrastive formulation defined in **Eq. (4)**, applied to its respective data view.

- **Hybrid:** This model integrates the global contrastive objective on augmented views with a cross-view predictive task. Specifically, it combines $\mathcal{L}_{Aug}$ with a loss that forecasts the future pitch features of an augmented view from the corresponding clean future segment ($\mathcal{L}_{Aug} + \mathcal{L}_{Pred}^{Clean_{fut} \to Aug_{fut}}$). This hybrid approach encourages the model to align global structural representations while maintaining local consistency between clean and perturbed pitch contours.

Loss Formulation Strategies. To justify our selection of independent loss terms for clean and augmented views ($\mathcal{L}_{Clean} + \mathcal{L}_{Aug}$), we evaluate two alternative contrastive formulations that incorporate cross-view interactions:

1. **Cross-Branch Class-Aware Loss (Cross-view SupCon):** This variant explicitly calculates the contrastive loss between clean view embeddings (z) and augmented view embeddings (z') to enforce direct cross-view alignment. The objective is defined as:

$$\mathcal{L}_{cross} = -\frac{1}{|I|}\sum_{i\in I}\frac{1}{|P(i)|}\sum_{p\in P(i)} \log\frac{\exp(\text{sim}(z_i, z'_p)/\tau)}{\sum_{a\in I}\exp(\text{sim}(z_i, z'_a)/\tau)} \tag{5}$$

   where I denotes the set of indices in a batch, and $P(i) = \{p \in I \mid y'_p = y_i\}$ is the set of indices of augmented samples sharing the same class label as clean sample i.

2. **Unified SupCon:** This variant treats clean and augmented views as a single integrated batch. We define a combined set

Table 2: Comparison of classification performance between the proposed Dual-Glob framework and baseline models. All results represent the average of 5-fold cross-validation (Mean ± SD).

Model	Acc	F1
Standard Deep Learning Baselines		
1D-CNN	0.7410 ± 0.0104	0.4930 ± 0.0134
BiLSTM	0.7568 ± 0.0156	0.4915 ± 0.0290
Transformer	0.7177 ± 0.0107	0.4680 ± 0.0248
State-of-the-Art Time-Series Models		
InceptionTime	0.7426 ± 0.0106	0.5043 ± 0.0147
TimesNet	0.6794 ± 0.0180	0.3759 ± 0.0191
MiniRocket	0.7303 ± 0.0152	0.4322 ± 0.0179
DLinear	0.6461 ± 0.0078	0.3892 ± 0.0242
Proposed (Dual-Glob)		
w/ LightGBM	0.7743 ± 0.0052	0.5086 ± 0.0064
w/ RF	0.7740 ± 0.0069	0.5051 ± 0.0061
w/ LR	**0.7775** ± 0.0064	**0.5154** ± 0.0151

$Z_{all} = \{z_1, \ldots, z_{|I|}, z'_1, \ldots, z'_{|I|}\}$ with indices J. The objective is to optimize the supervised contrastive loss over this unified batch:

$$\mathcal{L}_{unified} = -\frac{1}{|J|}\sum_{i\in J}\frac{1}{|P(i)|}\sum_{p\in P(i)} \log\frac{\exp(\text{sim}(r_i, r_p)/\tau)}{\sum_{a\in J\setminus\{i\}}\exp(\text{sim}(r_i, r_a)/\tau)} \quad (6)$$

where r denotes an embedding in Z_{all} and $P(i) = \{p \in J \setminus \{i\} \mid y_p = y_i\}$.

Gender-Specific Analysis. Previous studies (Henton, 1989; Pépiot, 2014) show that F_0 from female speakers ranged wider and is more variable than that from male speakers. We hypothesized these different F_0 realizations could hinder learning the shared tonal patterns. To investigate this, we designed two settings:

1. **Disaggregated Evaluation:** We evaluated the unified model on male and female subsets separately to detect potential bias.

2. **Gender-Specific Training:** We trained independent models on gender-split data to assess if domain splitting outperforms learning a shared invariant space.

4.3 Result Analysis

Comparative Analysis with Baselines. Table 2 shows the main result of the proposed model and baseline models. Our **Dual-Glob** (w/ LR) achieved state-of-the-art Acc (**0.7775**) and F1 (**0.5154**), outperforming the strongest baseline, **BiLSTM** (0.7568). While **BiLSTM** remained competitive, our method demonstrated superior stability with lower standard deviations, suggesting contrastive learning effectively stabilized performance. **InceptionTime** (0.7426) and **MiniRocket** (0.7303) showed strong results but fall short, while **TimesNet** (0.6794) failed to capture pitch variations.

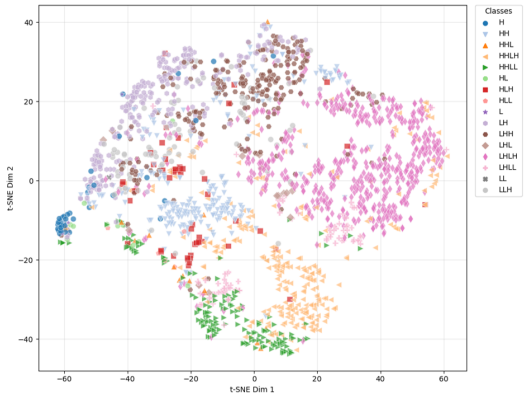


Figure 3: t-SNE visualization of the validation set. Distinct clusters form for most categories, while overlaps between similar classes (e.g., HHL vs. HHLL) reflect their F_0 contour shape resemblance.

Limitations of Attention and Decomposition. **Transformer** (0.7177) and **DLinear** (0.6461) performed poorly. DLinear's trend decomposition proves too rigid for complex intonation. Similarly, Transformer's global attention failed to capture fine-grained local transitions effectively compared to our approach. For detailed error analysis, refer to **Appendix D**.

Feature Visualization Figure 3 visualizes representations via t-SNE (Maaten and Hinton, 2008), showing our model maps distinct tonal patterns into clusters, capturing global pitch shapes. However, similar categories like HHL and HHLL overlap. Even though the final pitch patterns are different, the initial pitch patterns are the same so that those categories are placed closely in the latent space.

4.4 Analysis of Ablation Studies

Limitations of Predictive Contrastive Modeling. Table 3 highlights the limitations of the predictive contrastive paradigm (SimTS) for this task. Using **LightGBM**, both **Pred-C** (0.5521) and **Pred-A** (0.6901) showed significantly lower Acc compared to the global contrastive approaches. This suggests that the predictive goal itself is not suitable for this

task, regardless of how it predicts the future or past sequence. Since Seoul Korean tones are defined by the overall shape of the pitch, methods that rely on predicting only parts of the sequence fail to learn the global structure.

Efficacy of Global Constraints. All global contrastive models consistently exceed 0.76 Acc, outperforming predictive baselines by a wide margin. This confirms that applying **SupCon** on the full sequence effectively captures the holistic tonal patterns essential for classification.

Superiority of the Proposed Method. Proposed model (Dual-Glob) achieved the highest Acc (0.7775), showing that simply enforcing consistency between the global representations of different views is the most effective strategy. Interestingly, the **Hybrid** model (w/ LR) performed slightly worse (0.7712). This model combines the global contrastive objective with a predictive task that forecasts the augmented features from the clean ones. The results suggest that this explicit predictive constraint is unnecessary; it enforces the model to learn exact feature mappings between views, potentially distracting it from learning the invariant global patterns that are critical for classification.

Furthermore, our investigation into loss formulation strategies (Table 3) shows that the Dual-Glob approach yields favorable results compared to **Cross-View SupCon** (w/ LR) (0.7679) and **Unified SupCon** (w/ LR) (0.7732). While explicit cross-view alignment is common in self-supervised learning, these findings suggest that optimizing clean and augmented views through separate loss terms may provide a more balanced training signal for prosodic analysis. This independent supervision appears to encourage the model to capture the shared global patterns effectively while maintaining robustness against specific variations in perturbed pitch contours.

Effect of Gender Difference. Table 4 shows that in the unified model, female speakers achieved significantly higher Acc than male speakers (**0.8075** vs. **0.7130**). As shown in Figure 4, female speakers form a compact and high-performance cluster, whereas male speakers exhibit lower median Acc. Adopting a gender-specific approach further improved performance, increasing female speakers' Acc to **0.8120** and male speakers' Acc to **0.7288**. This difference is likely because female speakers use a wider pitch range, which makes their intonation patterns more distinct and easier for the model to recognize. In contrast, male speakers tend to use a narrower range, creating more similar and ambiguous patterns that are harder to tell apart. For a visual look at how these patterns are separated in the model, please see **Appendix E**.

Table 3: Performance comparison of contrastive models including loss formulation variants. Performance is measured via 5-fold cross-validation on frozen features.

Method	LightGBM		RF		LR	
	Acc	F1	Acc	F1	Acc	F1
Pred-C	0.5521	0.3231	0.5469	0.3193	0.5275	0.32970
Pred-A	0.6901	0.3722	0.5976	0.3549	0.5981	0.3064
Glob-Clean	0.7688	0.4892	0.7677	0.4822	0.7708	0.4931
Glob-Augment	0.7654	0.4838	0.7656	0.4844	0.7697	0.4918
Hybrid	0.7679	0.4956	0.7659	0.4868	0.7712	0.4947
Cross-View SupCon	0.7670	0.4877	0.7661	0.4835	0.7679	0.4878
Unified SupCon	0.7721	0.4970	0.7719	0.4950	0.7732	0.5051
Proposed (Dual-Glob)	**0.7743**	**0.5051**	**0.7740**	**0.5051**	**0.7775**	**0.5154**

Table 4: Performance comparison between unified and gender-specific models.

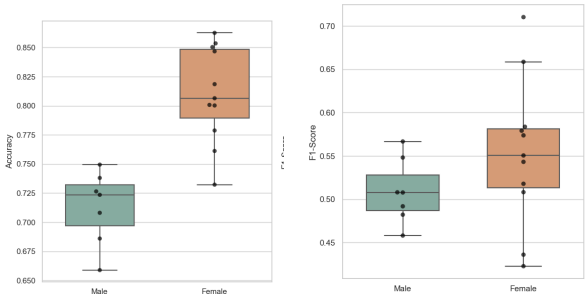
Gender	Unified Model		Gender-Specific Model	
	Acc	F1	Acc	F1
Male	0.7130 $\pm$0.04	0.4784 $\pm$ 0.04	0.7288 $\pm$ 0.02	0.5539 $\pm$ 0.02
Female	**0.8075** $\pm$ 0.01	**0.5286** $\pm$ 0.02	**0.8120** $\pm$ 0.02	**0.6103** $\pm$ 0.03

4.5 Ambiguity in Sustained Tones

A detailed error analysis reveals a limitation in modeling tonal patterns ending with sustained L tones. Figure 5 illustrates failure cases where the model incorrectly predicts HHLL for ground truth labels ending in single L tones.

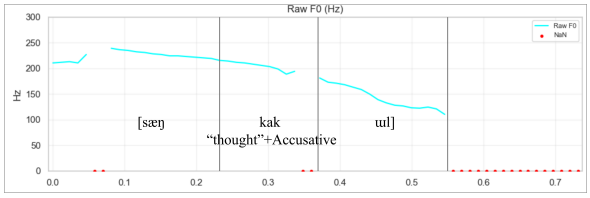
In Figure 5a, the AP, [sæŋ.kak + ɯl] ("thought"+Accusative), labeled as HHL, is misclassified as HHLL. Similarly, in Figure 5b, the AP [t^{h}oŋ.hæ] ("through"), labeled as HL, is also predicted as HHLL. These errors come from the acoustic ambiguity of the final L tone. When a speaker lengthens the final syllable (e.g., [hæ] in [t^{h}oŋhæ]), the resulting long and flat F_0 contour is geometrically indistinguishable from a sequence of multiple L tones to a model that relies solely on F_0 shapes.

This ambiguity is because the model only looks at the F_0 shape and does not know the length of each sound or where the syllables end. Without this information, the model often mistakes a long and flat low pitch for several different patterns. For more details on these errors, please see **Appendix D**.

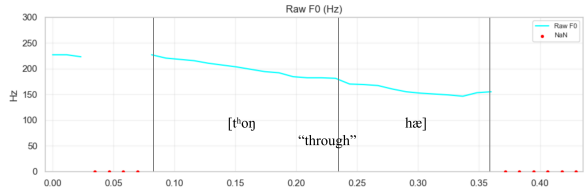


(a) Acc distribution. (b) F1 distribution.

Figure 4: Performance distribution of unified model across Acc and F1.



(a) Misclassification of [sæŋ.kak + ɯl] ("thought"+Accusative): Predicted HHLL, Ground Truth HHL.



(b) Misclassification of [tʰoŋ.hæ] ("through"): Predicted HHLL, Ground Truth HL.

Figure 5: Failure cases demonstrating the ambiguity in sustained tones. In both cases, the model misinterprets the lengthened final L tone as a sequence of multiple L tones (LL).

5 Discussion

To address the aforementioned limitations, we incorporated syllable count constraints into the classification framework. Specifically, we encoded the syllable count (N_{syl}) of each AP into a structural vector. We clip the maximum length at 4 (i.e., any $N_{syl} \geq 4$ is encoded as the maximum category).

This syllable information is encoded as a one-hot vector $v_{syl} \in \mathbb{R}^4$ and concatenated with the frozen latent representation z derived from the encoder. The final input to the classifier becomes $[z; v_{syl}]$. This fusion strategy allows the model to distinguish whether a sustained contour corresponds to a single lengthened syllable or multiple tonal targets, thereby resolving the ambiguity in patterns like HL versus HHLL.

The results in Table 5 show the performance of our syllable-aware model. By combining syllable counts with F_0, the model reached its highest Acc of 0.894 with LR. This demonstrates that adding simple timing information helps the model better understand complex intonation patterns.

We also analyzed how the model performs for different genders in Table 6. The syllable-aware model with LightGBM worked particularly well for female speakers, achieving 0.9154 Acc and 0.7725 F1. This confirms that even after adding syllable information, the clearer and wider pitch ranges of female voices continue to provide more distinct cues for the model compared to male voices.

Table 5: Classification performance comparison of the proposed syllable-aware model.

Ours	**Acc**	**F1**
w/ LightGBM	0.891 ± 0.014	0.694 ± 0.020
w/ RF	0.865 ± 0.015	0.622 ± 0.024
w/ LR	$\mathbf{0.894 \pm 0.021}$	$\mathbf{0.689 \pm 0.013}$

Table 6: Gender-specific performance analysis for the proposed syllable-aware model with LightGBM.

Gender	Ours	
	Acc	F1
Male	0.8548 ± 0.04	0.7153 ± 0.02
Female	0.9154 ± 0.02	0.7725 ± 0.03

6 Limitations

Although the proposed model effectively classifies pitch contours, several factors limit the extent to which tonal categories can be fully captured. First, while we rely on F_0, intonation is closely related to other prosodic cues, such as segmental duration and intensity that also mark prominence.

In addition, F_0 measurement itself remains imperfect: (1) F_0 track is often lost due to vowel devoicing or at the phrase-final position, (2) pitch halving errors (sudden drop of F_0) may lead to tonal misinterpretation, (3) glottalization in phrase-final L tones lead to pitch tracking loss or errors, and (4) F_0 rise due to local pitch perturbations following the release of the obstruents. (For a detailed discussion on these measurement difficulties, see **Appendix F**.)

Furthermore, as illustrated in Table 1, the distribution of pitch labels exhibits a significant class imbalance. In particular, categories with fewer than 100 samples show relatively low precision and re-

call. Consequently, overall F1 across all models, including the baselines, remain constrained. To address this, it is essential to secure a more substantial dataset, for example, obtaining several hundred samples per class, and to employ specialized loss functions or training strategies designed for class-balancing. For a more detailed error analysis for each class, please refer to Appendix D.(b).

Despite limitations, the model performs strongly using F_0 alone, suggesting pitch contours encode substantial structural information. Note that our analysis highlights inherent gender differences: female speakers typically exhibit wider ranges and more dynamic excursions in F_0 than male speakers.

However, it should be noted our dataset was constructed based on broadcast speech, which is optimized for clear information delivery, but this may lead to a restricted set of pitch patterns, potentially constraining the observed variability of pitch patterns. Given that Kim (2008) found that pitch accent distributions differ by register, future work should therefore examine a broader range of speech styles and corpora to validate the generalizability of these findings.

7 Conclusion

In this work, we make two main contributions to the study of Seoul Korean intonation. First, we constructed a **new benchmark dataset** by manually labeling AP in Seoul Korean. This dataset provides high-quality tonal labels, making it a rich resource for various fields such as speech synthesis and automated linguistic analysis.

Second, we presented **Dual-Glob**, a framework designed to capture the holistic shape of F_0 contours. Unlike previous methods that focus on local transitions, our approach learns the entire structure of the pitch pattern at once. Our results showed that this global perspective is much more effective, achieving significantly higher Acc than existing models.

Our analysis also revealed that while the narrower pitch ranges of male speakers can make classification more difficult, our model remains robust by capturing the global structure of the sounds. These findings confirm that focusing on the F_0 shapes as a whole is the most reliable way to analyze intonational patterns.

References

Jonathan Barnes, Alejna Brugos, Nanette Veilleux, and Stefanie Shattuck-Hufnagel. 2021. On (and off) ramps in intonational phonology: Rises, falls, and the tonal center of gravity. *Journal of Phonetics*, 85:101020.

Jonathan Barnes, Nanette Veilleux, Alejna Brugos, and Stefanie Shattuck-Hufnagel. 2012. Tonal center of gravity: A global approach to tonal implementation in a level-based intonational phonology. *Laboratory Phonology*, 3(2):337–383.

Mary E Beckman and Julia Hirschberg. 1994. The tobi annotation conventions. *Ohio State University*.

Mary E Beckman and Janet B Pierrehumbert. 1986. Intonational structure in japanese and english. *Phonology*, 3:255–309.

Dwight L Bolinger. 1951. Intonation: levels versus configurations. *Word*, 7(3):199–210.

Leo Breiman. 2001. Random forests. *Machine learning*, 45(1):5–32.

Yue Chen, Yingming Gao, and Yi Xu. 2022. Computational modelling of tone perception based on direct processing of f 0 contours. *Brain Sciences*, 12(3):337.

Jennifer Cole and Stefanie Shattuck-Hufnagel. 2016. New methods for prosodic transcription: Capturing variability as a source of information. *Laboratory Phonology*, 7(1).

David R Cox. 1958. The regression analysis of binary sequences. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 20(2):215–232.

Angus Dempster, Daniel F Schmidt, and Geoffrey I Webb. 2021. Minirocket: A very fast (almost) deterministic transform for time series classification. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pages 248–257.

Johan't Hart, JT Hart, R Collier, A Cohen, Rene Collier, and Antonie Cohen. 2003. *Perceptual Study of Intonation*. Cambridge.

Richard Hatcher, Hyunjung Joo, Sahyang Kim, and Taehong Cho. 2024. Focus-induced tonal distribution in seoul korean as an edge-prominence language. *Journal of Phonetics*, 107:101353.

Caroline G Henton. 1989. Fact and fiction in the description of female and male pitch. *Language and communication*, 9(4):299–311.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Khalil Iskarous. 2017. The relation between the continuous and the discrete: A note on the first principles of speech dynamics. *Journal of Phonetics*, 64:8–20.

Khalil Iskarous, J Cole, and J Steffman. 2023. American english pitch accent dynamics: A minimal dynamical model. In *Proceedings of the International Congress of Phonetic Sciences*. Guarant International.

Khalil Iskarous and Jennifer Cole. 2026. A quantal dynamical theory of f0 contours: Bridging the phonetics and phonology of intonation in: Developments in the modeling of speech prosody.

Hassan Ismail Fawaz, Benjamin Lucas, Germain Forestier, Charlotte Pelletier, Daniel F Schmidt, Jonathan Weber, Geoffrey I Webb, Lhassane Idoumghar, Pierre-Alain Muller, and François Petitjean. 2020. Inceptiontime: Finding alexnet for time series classification. *Data mining and knowledge discovery*, 34(6):1936–1962.

Hyunjung Joo and Mariapaola D'Imperio. 2025. The perception of lexical pitch accent in south kyungsang korean: The relevance of accent shape. *Language and Speech*, page 00238309251368294.

Sun-Ah Jun. 1998. The accentual phrase in the korean prosodic hierarchy. *Phonology*, 15(2):189–226.

Sun-Ah Jun. 2000. K-tobi (korean tobi) labelling conventions. *UCLA working papers in phonetics*, 99:149–173.

Sun-Ah Jun. 2003. The effect of phrase length and speech rate on prosodic phrasing. In *proceedings of the XVth international congress of phonetic sciences*, pages 483–486.

Sun-Ah Jun. 2005. Korean intonational phonology and prosodic transcription. *Prosodic typology: The phonology of intonation and phrasing*, 1:201.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673.

Sahyang Kim. 2008. Intonational pattern frequency of seoul korean and its implication to word segmentation. *Speech Sciences*, 15(2):21–32.

D Robert Ladd. 2008. *Intonational phonology*. Cambridge University Press.

Jooyoung Lee, Kyungwha Kim, and Minhwa Chung. 2021. Korean dialect identification based on intonation modeling. In *2021 24th Conference of the Oriental COCOSDA International Committee for the Coordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, pages 168–173. IEEE.

Gina-Anne Levow. 2005. Context in multi-lingual tone and pitch accent recognition. In *Interspeech*, pages 1809–1812. Lisbon.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605.

National Information Society Agency. 2022. Broadcasting content conversational data. https://aihub.or.kr/aihubdata/data/view.do?dataSetSn=71557.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Erwan Pépiot. 2014. Male and female speech: a study of mean f0, f0 range, phonation type and speech rate in parisian french and american english speakers. In *Speech prosody 7*, pages 305–309.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Zhiguang Wang, Weizhong Yan, and Tim Oates. 2017. Time series classification from scratch with deep neural networks: A strong baseline. In *2017 International joint conference on neural networks (IJCNN)*, pages 1578–1585. IEEE.

Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. 2022. Timesnet: Temporal 2d-variation modeling for general time series analysis. *arXiv preprint arXiv:2210.02186*.

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

Seunghyun Yoon, Seokhyun Byun, and Kyomin Jung. 2018. Multimodal speech emotion recognition using audio and text. In *2018 IEEE spoken language technology workshop (SLT)*, pages 112–118. IEEE.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. 2023. Are transformers effective for time series forecasting? In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 11121–11128.

Xiaochen Zheng, Xingyu Chen, Manuel Schürch, Amina Mollaysa, Ahmed Allam, and Michael Krauthammer. 2023. Simts: Rethinking contrastive representation learning for time series forecasting. *arXiv preprint arXiv:2303.18205*.

A Appendix A: Schematic F_0 contours of an AP

The schematic F_0 contours of an AP in Seoul Korean are provided in Figure 6 (Redrawn from Jun (2000)). The APs in Seoul Korean are typically realized with the following sixteen pitch accent categories.

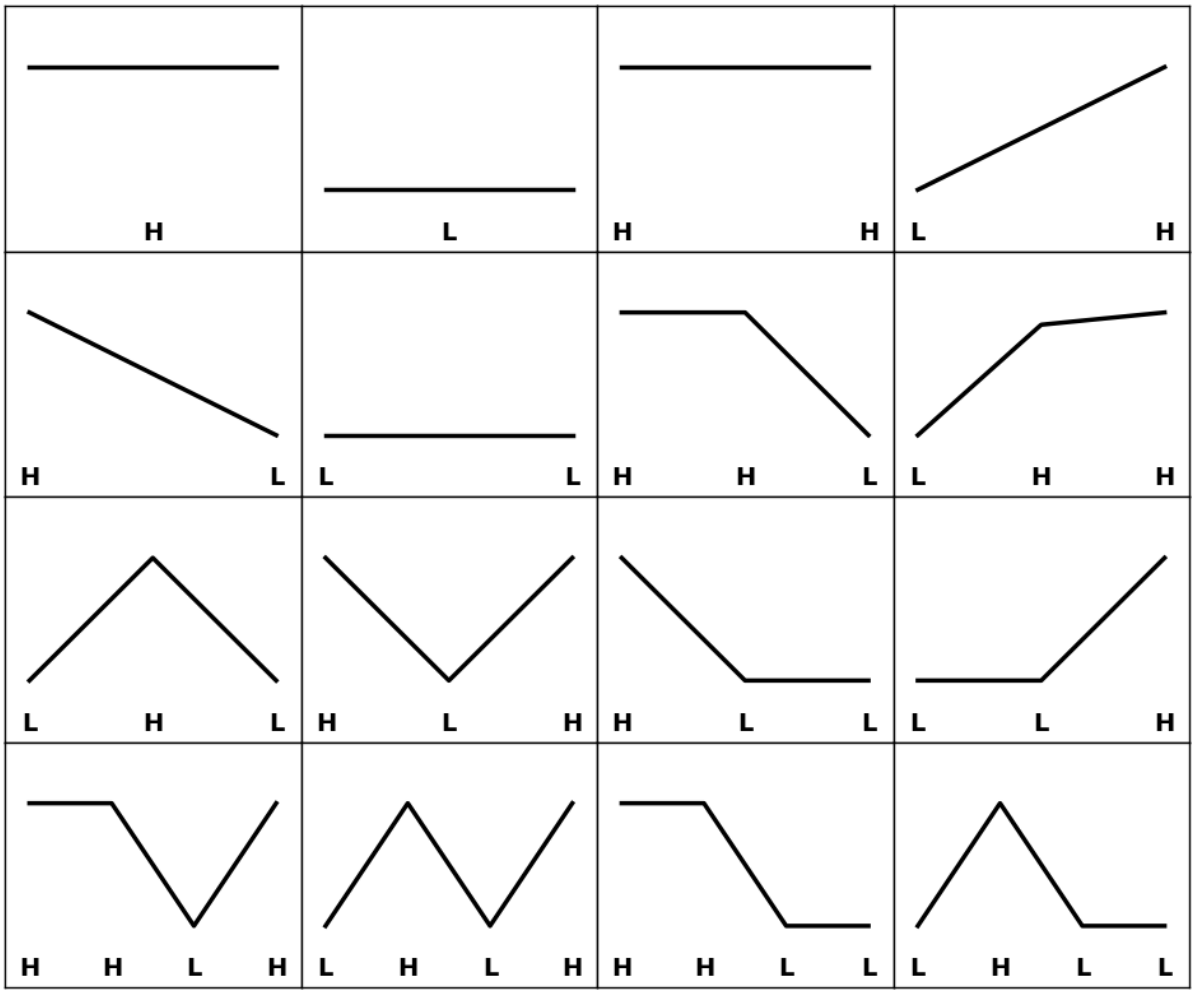


Figure 6: Schematic F_0 contours of sixteen pitch accent patterns for an AP in Seoul Korean (Jun, 2000)

Appendix B: Implementation Details

Environment and Data Split. All models were implemented using **PyTorch** and trained on an NVIDIA GPU RTX 2070. To ensure a robust evaluation, we employed **5-fold stratified cross-validation** with a fixed random seed (42). For each fold, the dataset was split into training and validation sets, and the final performance is reported as the mean and standard deviation across the five folds.

Model Architecture. For the proposed **Dual-Glob** framework, we utilized a **6-layer 1D CNN encoder** with kernel sizes of $[16, 12, 9, 6, 6, 6]$, strides of $[1, 2, 2, 1, 1, 1]$, and increasing channel sizes of $[16, 32, 64, 128, 256, D_{emb}]$. After the convolutional layers, a masked global average pooling (GAP) was applied to obtain the latent representation. Both the **projection and prediction heads** were configured as 2-layer MLPs with a hidden and output dimension of 64 and ReLU activation. To evaluate the robustness of the learned embeddings, we systematically varied the **embedding size** (D_{emb}) from **64 to 1024** during our experiments. For the total objective function, the weighting hyperparameters λ_1 and λ_2 were both set to 1.

To enhance the density of positive and negative pairs within each training iteration, we adopted a batch-duplication strategy where each mini-batch was concatenated with itself. This approach effectively

Table 7: Detailed architecture configurations and hyperparameters for all implemented models. All baselines were trained end-to-end, whereas the proposed method employed a fixed feature extractor followed by separate classifiers.

Model	Component	Configuration Details
Proposed (Dual-Glob)	Encoder	6-layer Conv1D (Kernels=[16, 12, 9, 6, 6, 6], Channels=[16, 32, 64, 128, 256, D_{emb}], Strides=[1, 2, 2, 1, 1, 1]) + Masked GAP
	Projector	2-layer MLP (Hidden=64, Output=64, ReLU)
	Predictor	2-layer MLP (Hidden=64, Output=64, ReLU)
	Classifier	Logistic Regression, Random Forest ($n_{est} = 200$), LightGBM ($n_{est} = 200$)
CNN	Backbone	5 Conv1D blocks (Channels=[64, 128, 256, 1024], Kernel=[16, 12, 9, 9]) + MaxPool(2)
	Head	AdaptiveAvgPool1d(1) $\rightarrow$ Linear(1024 $\rightarrow N_{classes}$)
LSTM	Backbone	Bidirectional LSTM (Hidden=64, Layers=2, BatchFirst=True)
	Head	Concatenated last hidden states (128 dim) $\rightarrow$ Linear(128 $\rightarrow N_{classes}$)
InceptionTime	Backbone	6 Inception Blocks (Depth=6, Hidden=128) + Residual Connections
	Head	AdaptiveAvgPool1d(1) $\rightarrow$ Flatten $\rightarrow$ Linear(512 $\rightarrow N_{classes}$)
Transformer	Backbone	2 Encoder Layers ($d_{model} = 256, n_{head} = 4, d_{ff} = 256$) + Positional Encoding
	Head	Global Average Pooling $\rightarrow$ Linear(256 $\rightarrow N_{classes}$)
MiniRocket	Transform	50,000 random convolutional kernels (Fixed weights)
	Classifier	Ridge Classifier (Linear, $\alpha \in [10^{-3}, 10^{3}]$)
TimesNet	Architecture	2 Layers, $d_{model} = 32$, Top-k=1, Multi-periodicity analysis
DLinear	Architecture	Single Linear Layer decomposition (SeqLen=200)

increases the number of contrastive relations, contributing to more stable and robust gradient estimation without requiring additional unique data samples.

For baseline comparisons, we deployed standard implementations of CNN, LSTM, InceptionTime, Transformer, DLinear, and TimesNet, along with MiniRocket. Deep learning baselines were trained end-to-end, whereas MiniRocket was evaluated using a Ridge Classifier.

Training Configuration. We optimized the models using **RAdam** combined with the **Lookahead** mechanism ($k = 5, \alpha = 0.9$) to stabilize convergence. The learning rate was set to 1×10^{-2} for self-supervised pre-training and 3×10^{-3} for supervised baselines, with a weight decay of 1×10^{-4}. The batch size was fixed at 64.

The training duration varied by model, ranging from 50 to 100 epochs depending on convergence speed. To ensure reliable performance estimation and account for training fluctuations, we calculated the final metrics by averaging the results of the last 5 training epochs across all 5 folds, rather than relying on a single best-epoch checkpoint.

Table 7 summarizes the specific architecture configurations and hyperparameter settings for all implemented models.

Appendix C: Effect of Data Augmentation

For contrastive learning, we employed a stochastic data augmentation strategy to learn noise-invariant representations. For each training instance, we **randomly selected 3 transformations** from the following pool of five techniques:

- **Random Jittering (J):** Adding Gaussian noise ($\sigma = 0.02$) to simulate recording noise.
- **Scaling (S):** Multiplying the amplitude by a random factor ($0.8 \sim 1.2$) to handle intensity variations.
- **Masking (M):** Randomly zeroing out a portion of the sequence (ratio=0.2) to encourage robustness against missing data.
- **Magnitude Shift:** Adding a random constant bias to the Log-F_0 contour, simulating global pitch level shifts.
- **Time Warping:** Applying non-linear temporal deformation using random knots and linear interpolation to simulate local speaking rate variations.

To evaluate the impact of different augmentation selection strategies on model performance, we conducted comparative experiments across six distinct configurations (D1–D6). These configurations vary in the number of transformations selected and the diversity of the augmentation pool.

As shown in Table 8 the D4 strategy (2∼3 random selection) consistently achieves the highest overall performance across all classifiers. It reaches a peak of 77.75% Acc and 51.54% F1 with the LR model and maintains the highest average performance (77.53% Acc / 50.97% F1). These results demonstrate that the dynamic combination of 2 to 3 augmentations is the most effective approach for learning robust and discriminative pitch accent representations.

Table 8: Comparative results of 5-fold cross validation across different adapter selection strategies (Encoder Dimension: 1024). Acc and F1 are reported in percentages (%).

Dataset	**Selection Strategy**	**RF (Acc/F1)**	**LR (Acc/F1)**	**LGBM (Acc/F1)**	**Avg. (Acc/F1)**
D1	1 random (Full)	77.25 / 50.12	77.49 / 51.25	77.48 / 50.39	77.41 / 50.59
D2	2 random (Full)	77.39 / 49.96	77.64 / 51.25	77.30 / 50.21	77.44 / 50.47
D3	3 random (Full)	77.23 / 49.74	77.61 / 50.35	77.24 / 50.78	77.36 / 50.29
D4 (Ours)	**2∼3 random (Full)**	**77.40 / 50.51**	**77.75 / 51.54**	**77.43 / 50.86**	**77.53 / 50.97**
D5	1 random {J, S, M}	77.09 / 49.53	77.47 / 50.87	77.23 / 50.04	77.26 / 50.15
D6	2 random {J, S, M}	77.30 / 49.99	77.57 / 50.81	77.60 / 51.18	77.48 / 50.68

To further evaluate the effectiveness of the proposed approach, we investigated the impact of data augmentation (D4) on various baseline models. For these baseline experiments, we trained the models by concatenating the original clean data with the augmented data, effectively doubling the training size to ensure a fair comparison. As summarized in Table 9, the application of data augmentation generally led to performance gains across most architectures. Specifically, models such as BiLSTM, InceptionTime, and Transformer exhibited notable increases in both Acc and F1. These results suggest that our data augmentation strategy contributes to better generalization by allowing models to learn from a more diverse set of pitch contour variations, even when applied to conventional supervised learning frameworks.

Table 9: Performance comparison of baseline models with (w/) and without (w/o) data augmentation. The w/ Augmentation setting uses a concatenation of clean and augmented data. Results represent the average of 5-fold cross-validation (Mean ± SD).

Model	**w/o Augmentation**		**w/ Augmentation**	
	Acc	**F1**	**Acc**	**F1**
1D-CNN	0.7291 ± 0.0087	**0.5011** ± 0.0157	**0.7410** ± 0.0104	0.4930 ± 0.0134
BiLSTM	0.7440 ± 0.0087	0.4782 ± 0.0190	**0.7568** ± 0.0156	**0.4915** ± 0.0186
InceptionTime	0.7327 ± 0.0071	0.4915 ± 0.0196	**0.7426** ± 0.0106	**0.5043** ± 0.0147
DLinear	0.6237 ± 0.0070	0.3875 ± 0.0139	**0.6461** ± 0.0078	**0.3892** ± 0.0242
MiniRocket	0.7287 ± 0.0159	**0.4372** ± 0.0236	**0.7303** ± 0.0152	0.4322 ± 0.0179
Transformer	0.6943 ± 0.0770	0.4434 ± 0.0133	**0.7177** ± 0.0107	**0.4680** ± 0.0248

Appendix D: Detailed Error Analysis

To quantitatively evaluate the classification behavior, we analyze the confusion matrix shown in Figure 7. The high concentration of values along the diagonal indicates that the model successfully distinguishes distinct tonal classes in most cases. However, off-diagonal elements reveal that misclassifications are not random but primarily occur between similar patterns (e.g., HL vs. HHLL). These confusions suggest that while the global F_0 contour shapes are well-captured, subtle ambiguities in duration and syllable boundaries remain challenging.

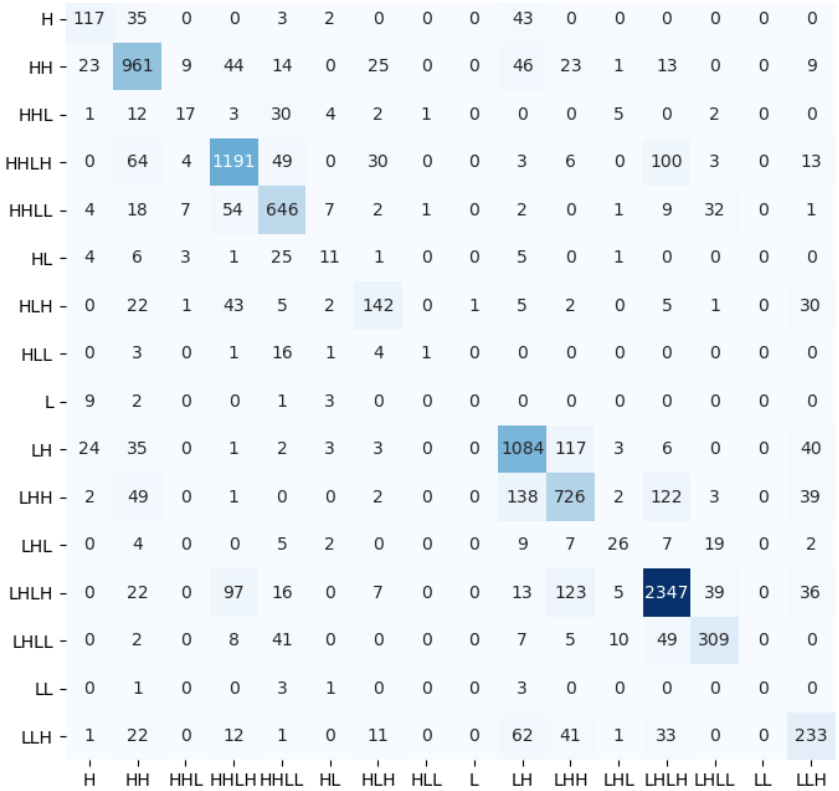


Figure 7: Confusion matrix of the proposed Dual-Glob method with LR. The strong diagonal density confirms robust classification accuracy across all tonal classes.

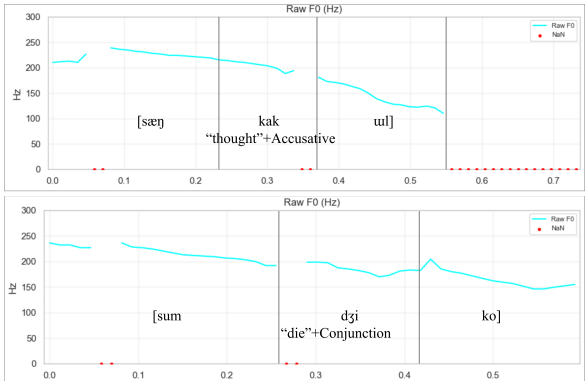
(a) Analysis of Error Patterns

We qualitatively analyzed failure cases where the model misclassifies tonal patterns. In each figure below, the upper and lower panels show representative examples of the error patterns. These error patterns are largely coming from the ambiguous situations when interpreting sustained tonal contours or lengthened syllables without segmenting syllable boundaries.

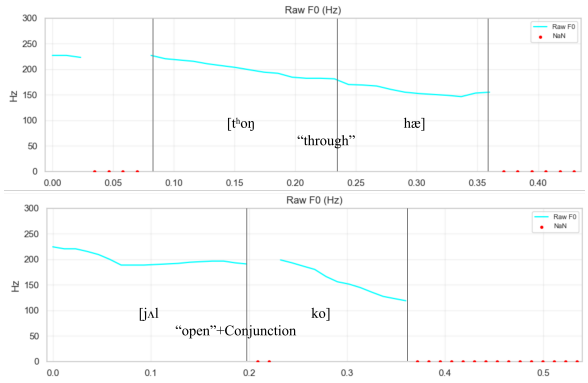
(1) HHL → HHLL (Figure 8a): This category shows APs with three syllables, which is misclassified as four tones. In the example on the top, [sæŋ.kak + ɯl] ("thought"+Accusative), the final syllable [ɯl] is lengthened. Similarly, in the example at the bottom, [sum.dʒi.ko] ("die"+Conjunction), the final syllable [ko] shows a sustained contour. Due to these reasons, the model incorrectly splits these final segments into two low tones (LL).

(2) HL → HHLL (Figure 8b): These are APs with two syllables, but misclassifed as four tones. In [t^hoŋ.hæ] ("through") (Top), the final [hæ] is lengthened. In [jʌl.ko] ("open"+Conjunction)(Bottom), the initial syllable [jʌl] is also long, which results in a lengthy flat trajectory. The model misinterprets these longer durations as multiple tonal targets, predicting HHLL for both.

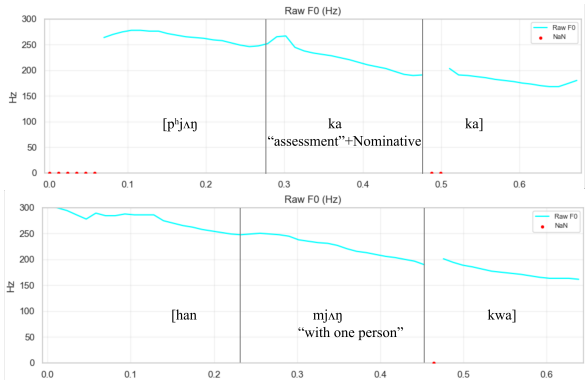
(3) HLL → HHLL (Figure 8c): Both [p^hjʌŋ.ka.ka] ("assessment"+Nominative) (Top) and [han.mjʌŋ.kwa] ("with one person") are APs with three syllables ending in two L tones. The continuous low-pitch part for the last two syllables (e.g., [ka.ka]' or '[mjʌŋ.kwa]) does not show a clear acoustic



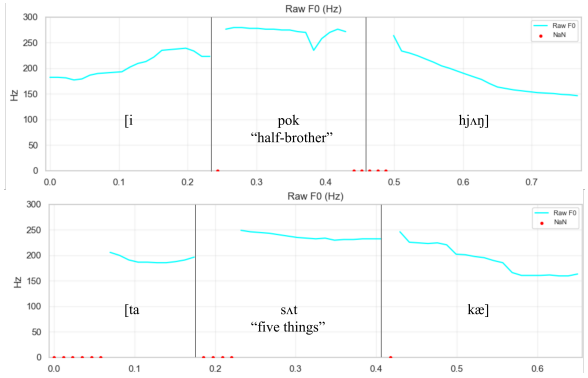
(a) HHL → HHLL.
Top: [sæŋ.kak + ɯl] ("thought"+Accusative)
Bottom: [sum.dʒi.ko] ("die"+Conjunction)



(b) HL → HHLL.
Top: [t^hoŋ.hæ] ("through")
Bottom: [jʌl.ko] ("open"+Conjunction)



(c) HLL → HHLL.
Top: [p^hjʌŋ.ka.ka] ("assessment"+Nominative)
Bottom: [han.mjʌŋ.kwa] ("with one person")



(d) LHL → LHLL.
Top: [i.pok.hjʌŋ] ("half-brother")
Bottom: [ta.sʌt.kæ] ("five things")

Figure 8: Visualization of common misclassification patterns. Each subfigure displays two examples (Top/Bottom) showing the same type of prediction error. The vertical lines indicate the ground-truth syllable boundaries, which are hidden from the model.

dip. Without syllable boundaries, the model fails to match the F_0 contour with each L tone, leading to an HHLL pattern.

(4) LHL → LHLL (Figure 8d): This error occurs in LHL patterns where the final Low is misrecognized. In [i.pok.hjʌŋ] ("half-brother") (Top) and [ta.sʌt.kæ] ("five things") (Bottom), the final syllables ([hjʌŋ] and kæ]) exhibit a falling or sustained L tone. The model correctly identifies the initial LH rise but misinterpret the final L into LL, resulting in an LHLL prediction.

(b) Per-Class Metrics and Error Analysis

Table 10 provides a detailed breakdown of the classification performance for each pitch accent class. While the macro-averaged F1 of the proposed model is approximately 0.51, which may appear relatively modest, a closer inspection of the per-class metrics reveals that this is primarily driven by the extreme data sparsity in minority categories.

As illustrated in Table 10, classes with a small number of samples, such as HHL, HL, HLL, LHL, L, and LL show significantly lower performance, which heavily penalizes the macro-average. In contrast, dominant labels with over 1,000 samples, including HH, HHLH, LH, LHH, and LHLH, consistently achieve F1-scores ranging from 0.68 to as high as 0.87. These results indicate that the model achieves robust performance on dominant prosodic patterns, while the lower macro-average is primarily a reflection of the inherent data imbalance.

For future research, we aim to construct a more extensive and balanced dataset to mitigate these issues. Additionally, we plan to incorporate advanced training strategies, such as class-balanced loss or

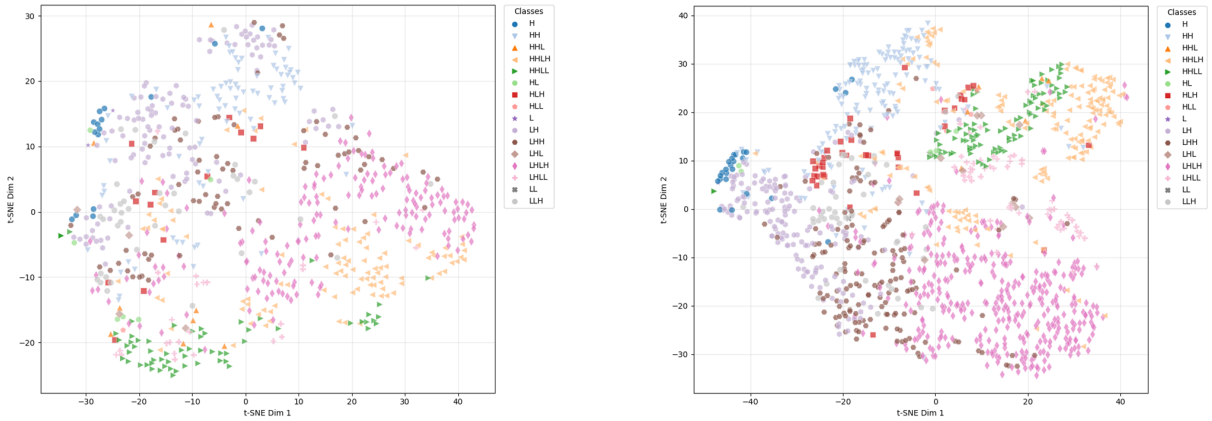
Table 10: Detailed classification performance for each pitch accent class in Seoul Korean. The results are evaluated via 5-fold cross-validation.

Tone Class	**Precision**	**Recall**	**F1**	**Support**
H	0.63	0.58	0.61	200
HH	0.76	0.82	0.79	1,168
HHL	0.41	0.22	0.29	77
HHLH	0.82	0.81	0.82	1,463
HHLL	0.75	0.82	0.79	784
HL	0.31	0.19	0.24	57
HLH	0.62	0.55	0.58	259
HLL	0.33	0.04	0.07	26
L	0.00	0.00	0.00	15
LH	0.76	0.82	0.79	1,318
LHH	0.69	0.67	0.68	1,084
LHL	0.47	0.32	0.38	81
LHLH	0.87	0.87	0.87	2,705
LHLL	0.76	0.72	0.74	431
LL	0.00	0.00	0.00	8
LLH	0.58	0.56	0.57	417
Accuracy		**0.77**		10,093
Macro Avg.	0.55	0.50	0.51	10,093
Weighted Avg.	0.77	0.77	0.77	10,093

cost-sensitive learning, to further improve the model's sensitivity to uncommon prosodic patterns.

Appendix E: Latent Space Visualization by Gender

We visualized learned representations via t-SNE (Figure 9). Figure 9b shows the female model forms well-separated clusters, indicating wider pitch ranges provide clear cues. In contrast, Figure 9a reveals that the male space overlaps due to narrower pitch excursions, creating ambiguous boundaries between similar contours.



(a) Male Speakers

(b) Female Speakers

Figure 9: t-SNE visualization of the feature space learned by the proposed model. Female speakers (b) show distinct class separation, whereas male speakers (a) exhibit significant overlap and scattering.

Appendix F: Difficulties in Pitch Tracking

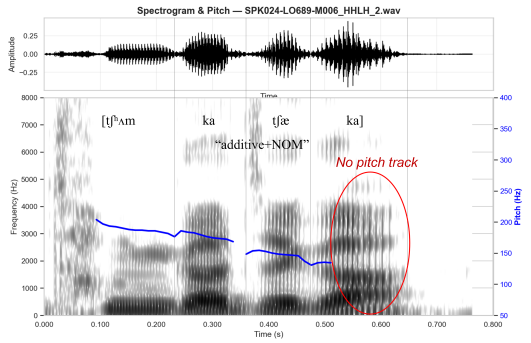
The performance of our model is closely related to the quality of the extracted F_0 contours. However, as shown in Figure 10, real-world speech data often includes pitch discontinuations or pitch tracking errors that make precise pitch tracking difficult, leading to potential misclassifications.

First, pitch tracking loss is quite frequent. The F_0 disappears entirely in many cases, such as during vowel devoicing following [s] or the affricate (Figure 10d and 10e) or at sentence-final positions (Figure 10a). These missing portions of the F_0 are problematic, as the model loses crucial information required to identify the entire tonal pattern.

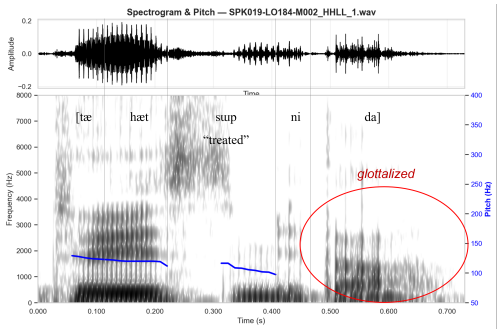
Also, pitch tracking errors such as pitch halving creates sudden drops in the F_0 contour (Figure 10f and 10h). When these errors occur, the model may misinterpret a H tone as a L tone.

Third, the realization of low tones is often accompanied by glottalization on the vowel, in which F_0 contours exhibit pitch tracking errors or signal loss. The pattern is particularly frequent in the phrase-final L tones as in LHL$\underline{\text{L}}$ or HHL$\underline{\text{L}}$ (Figure 10b and 10c).

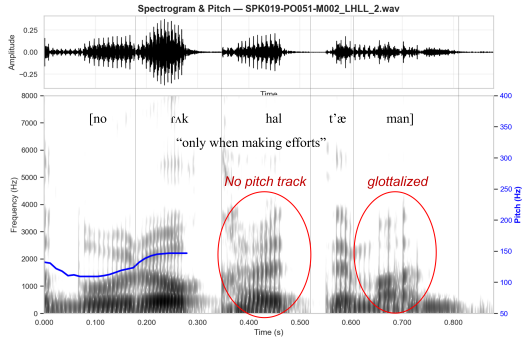
Lastly, local F_0 perturbations often lead to slight F_0 rising following the release of stops and affricates (Figure 10g and 10h).



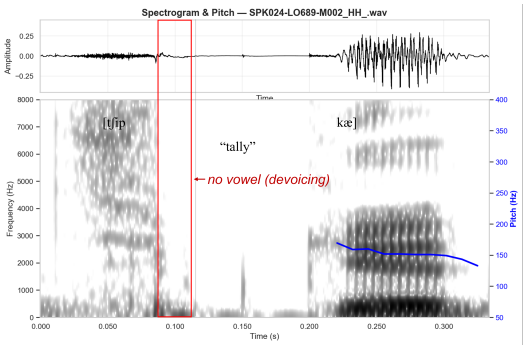
(a) Pitch track loss in the AP-final syllable.



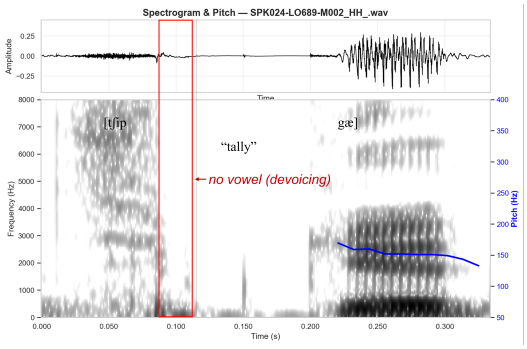
(b) Pitch track loss due to glottalization on the AP-final vowel.



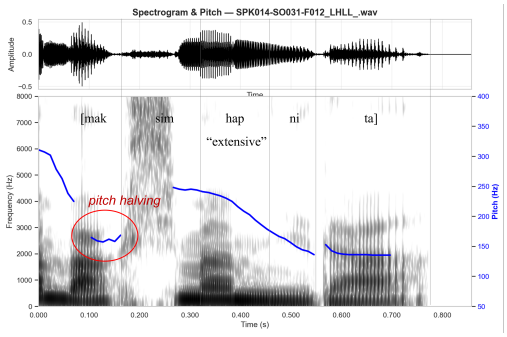
(c) Pitch track loss on the third syllable, as well as on the AP-final vowel due to glottalization.



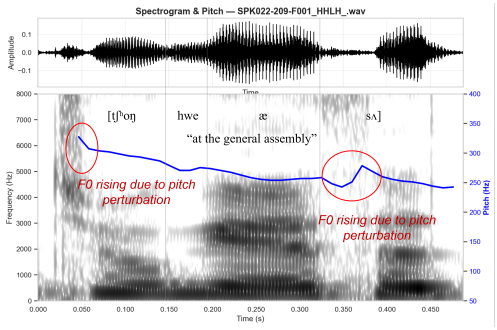
(d) Pitch track loss due to vowel devoicing after [s].



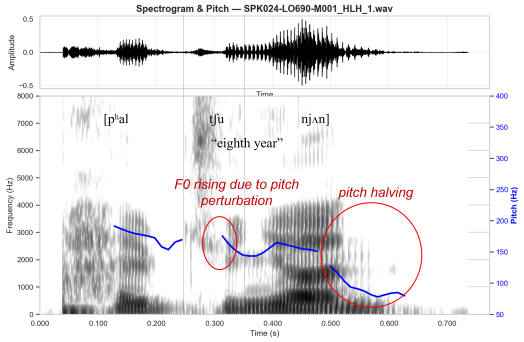
(e) Pitch track loss due to vowel devoicing after the affricate.



(f) Pitch track error due to pitch halving.



(g) F_0 rising due to local pitch perturbation.



(h) F_0 rising due to local pitch perturbation and pitch track error due to pitch halving.

Figure 10: Visual analysis of various F_0 discontiunations or pitch track errors in Seoul Korean speech data, including devoicing, pitch halving, glottalization, and F_0 perturbation.

Appendix G: Effect of Encoder Dimension

To investigate the impact of the representation capacity on downstream classification performance, we evaluated the extracted features using three classifiers—RF, LR, and LightGBM—across encoder dimensions ranging from 64 to 1024. The results are summarized in Table 11.

Overall, the classification performance exhibited a slight improvement as the encoder dimension increased. In particular, the LR classifier demonstrated the most notable enhancement, achieving the best overall performance of our proposed framework at a dimension of 1024 (Acc: 0.7775, F1: 0.5154). In contrast, while the RF and LightGBM models also showed marginal performance gains, the differences across dimensions were not highly significant. These findings suggest that expanding the encoder dimension generally yields minor performance benefits, but specifically enables the LR model to optimally leverage the high-dimensional latent representations.

Table 11: Performance comparison across different encoder dimensions. All results represent the average of 5-fold cross-validation (Mean ± SD).

Encoder Dimension	Metric	RF	LR	LightGBM
64	Acc	0.7710 ± 0.0057	0.7690 ± 0.0056	0.7694 ± 0.0062
	F1	0.4979 ± 0.0091	0.4701 ± 0.0148	0.4968 ± 0.0145
128	Acc	0.7728 ± 0.0082	0.7721 ± 0.0082	0.7686 ± 0.0087
	F1	0.4944 ± 0.0108	0.4880 ± 0.0106	0.4893 ± 0.0149
256	Acc	0.7720 ± 0.0046	0.7747 ± 0.0048	0.7731 ± 0.0065
	F1	0.5000 ± 0.0134	0.5029 ± 0.0110	0.5003 ± 0.0167
512	Acc	0.7710 ± 0.0070	0.7730 ± 0.0061	0.7689 ± 0.0083
	F1	0.4938 ± 0.0073	0.5004 ± 0.0133	0.4968 ± 0.0180
1024	Acc	0.7740 ± 0.0069	$\mathbf{0.7775 \pm 0.0064}$	0.7743 ± 0.0052
	F1	0.5051 ± 0.0061	$\mathbf{0.5154 \pm 0.0157}$	0.5086 ± 0.0064