R-APM: Retrieval-Augmented Pragmatic Mapper for Cross-Lingual Prosody Transfer

Interspeech 2026 TOPI Challenge – System Description

Xiaoyang Luo[1,*], Siyuan Jiang[1], Shuya Yang[1]
Dengfeng Ke[1]*, Yanlu Xie[1]*, Jinsong Zhang[1]*

[1] Speech Acquisition and Intelligent Technology Laboratory (SAIT LAB)
Beijing Language and Culture University, Beijing, China

Abstract

This paper presents R-APM (Retrieval-Augmented Pragmatic Mapper), a hybrid system developed for the Interspeech 2026 TOPI Challenge. We investigate whether non-parametric retrieval can capture the nuanced relationship between pragmatic intent and prosody better than parametric models. Our experiments reveal a significant generalization challenge. While our best configuration (Config B: 103-dim subspace retrieval) achieves a cosine similarity of 0.8331 on the official blind test set, it lags behind the official parametric MLP baseline (0.8574). However, our analysis uncovers a critical "identity overfitting" phenomenon: high-dimensional retrieval (1024-dim) degrades significantly on unseen speakers due to timbre interference, whereas retrieving within a pragmatically-salient subspace (103-dim) improves generalization (+0.0032 for pure retrieval). Furthermore, we analyze the "Small Data Trap," showing that hybrid fusion networks struggle to improve upon retrieval priors when data density is low ($N < 3k$).

1 Introduction

Cross-lingual prosody transfer aims to map the pragmatic intent of a source language (English) to the prosodic realization of a target language (Spanish). The core challenge lies in learning a mapping that is invariant to speaker identity but sensitive to pragmatic context.

This work describes our submission to the Interspeech 2026 TOPI Challenge [1]. The official baseline employs a Multi-Layer Perceptron (MLP) [4] to learn this mapping directly. While parametric models are efficient, they may struggle to capture the full diversity of prosodic expression from limited data. We propose **R-APM**, a retrieval-augmented approach, to test if explicit memory of training examples can offer better guidance.

Our contribution is primarily analytical. We demonstrate that: 1) Direct high-dimensional retrieval suffers from severe identity overfitting on unseen speakers. 2) Reducing the search space to pragmatically relevant features improves generalization but still falls short of the parametric baseline. 3) Fusion networks face a "Small Data Trap," where they overfit the residual noise in the training set rather than learning systematic corrections.

Beyond performance, our retrieval-based approach offers transparency: for any prediction, we can directly inspect which training utterances informed the output, enabling error analysis and system debugging.

2 System Description

2.1 Task and Features

The task involves mapping a 1024-dimensional English HuBERT [7] feature vector ($\mathbf{x}_{en}$) to a 101-dimensional Spanish HuBERT feature vector (the target prosodic representation). We utilize the official dataset [2, 3] consisting of 2,893 aligned utterance pairs.

2.2 The Retrieval Module

Our approach is inspired by nearest-neighbor language models [8] (see Figure 1). We construct a key-value datastore $\mathcal{D} = \{(\mathbf{k}_i, \mathbf{v}_i)\}_{i=1}^{N}$ from the training set. Given a query $\mathbf{q}$, we compute cosine similarity, select the top-K neighbors, and aggregate using a temperature-scaled softmax.

$$\hat{\mathbf{y}}_{ret} = \sum_{j \in \text{Top-}K} \frac{\exp(s_j/\tau)}{\sum_{m \in \text{Top-}K} \exp(s_m/\tau)} \mathbf{v}_j \qquad (1)$$

2.3 Retrieval Configurations

We investigated two distinct retrieval workflows to isolate the effect of feature dimensionality on speaker generalization.

2.3.1 Config A: High-Res (1024-dim)

This configuration uses the full semantic capacity of the model.

[*] Corresponding authors: Dengfeng Ke (dengfeng.ke@blcu.edu.cn), Yanlu Xie (xieyanlu@blcu.edu.cn), Jinsong Zhang (jinsong.zhang@blcu.edu.cn).

[*] Contact for questions: 202211590399@stu.blcu.edu.cn.

- **Keys & Values:** Both keys $\mathbf{k}$ and values $\mathbf{v}$ are full 1024-dim vectors.
- **Output:** The aggregated result $\hat{\mathbf{y}}_{ret}^{(1024)}$ is projected to 101 dimensions ($\hat{\mathbf{y}}_{ret}^{(101)}$) using the official `spanish_winners` indices [5, 6] only at the final step.

2.3.2 Config B: Pragmatic-Subspace (103-dim)

This configuration aims to filter non-prosodic interference (e.g., speaker timbre).

- **Keys:** The input query $\mathbf{x}_{en}$ is reduced to 103 dimensions using `english_winners` indices [5, 6]. The datastore keys $\mathbf{k}$ are also 103-dim.
- **Values:** Crucially, we still retrieve the full 1024-dim values $\mathbf{v}$ associated with the matched keys.
- **Output:** This produces a high-resolution prior $\hat{\mathbf{y}}_{ret}^{(1024)}$ based on a low-dimensional semantic match.

2.4 The Fusion Network

The fusion network learns a residual correction, as illustrated in Figure 2. To preserve maximum context, the network input is the concatenation of the **full 1024-dim English input** $\mathbf{x}_{en}$ and the **101-dim retrieved prior** $\hat{\mathbf{y}}_{ret}^{(101)}$.

$$\mathbf{y}_{pred} = \hat{\mathbf{y}}_{ret}^{(101)} + \mathcal{F}_\theta(\text{concat}[\mathbf{x}_{en}, \hat{\mathbf{y}}_{ret}^{(101)}]) \tag{2}$$

Note that while Config B uses 103-dim queries for retrieval, the Fusion Network receives the original 1024-dim $\mathbf{x}_{en}$ to preserve full context. The MLP consists of layers [1125 → 256 → 128 → 101] with LayerNorm and GELU activation.

3 Experimental Setup

3.1 Data Configuration

We utilized the standard train/test split provided in the official baseline repository. The training set consists of utterances listed in the official `train_filelist`.

- **Internal Split (Seen):** The standard baseline test split. Note that this split contains speakers that are also present in the training set (Speaker Overlap).
- **Official Challenge Set (Unseen):** The blind test set provided by organizers. Critically, 4 out of 5 speakers in this set were not present in the DRAL training corpus, representing a severe domain shift that particularly challenges retrieval-based methods.

Table 1: **Main Performance Comparison.** We compare the High-Res (1024-dim) and Pragmatic-Subspace (103-dim) configurations against the Official Baseline. 'Gain' denotes the absolute improvement added by the Fusion Network.

System	**Ret.**	**Internal (Seen)**		**Official (Unseen)**	
	Dim	Cosine	Gain	Cosine	Gain
Baseline MLP	-	0.8732	-	**0.8574**	-
Config A: High-Res					
Pure Ret	1024	0.8722	-	0.8286	-
Ret + Fusion	1024	**0.8742**	+0.0020	0.8290	+0.0004
Config B: Subspace					
Pure Ret	103	0.8730	-	0.8318	-
Ret + Fusion	103	0.8741	+0.0011	0.8331	+0.0013

3.2 Training Implementation

We set the random seed to 42 for reproducibility. The Fusion Network was trained for 100 epochs using the AdamW optimizer with a learning rate of $1e^{-3}$ and a weight decay of $1e^{-4}$, using a batch size of 32. We employed an early stopping strategy based on validation loss to prevent overfitting. Retrieval operations were accelerated on GPU using ChromaDB for efficient approximate nearest neighbor search. The loss function is Cosine Embedding Loss.

4 Results

4.1 Main Performance Comparison

Table 1 presents the comprehensive performance. The metric reported is Cosine Similarity.

The Generalization Gap: On the Internal Split (Seen), our Config A (0.8742) performs comparably to the Baseline. However, on the Official Set (Unseen), all retrieval-based methods underperform the Baseline (0.8574).

Identity Overfitting: Comparing Config A and B on the Official Set reveals that reducing dimensions from 1024 to 103 improves pure retrieval from 0.8286 to 0.8318. This suggests that the 1024-dim space is dominated by speaker identity; when speakers are unseen, high-dimensional retrieval finds false neighbors (timbre matches) rather than prosodic matches.

Domain Shift Severity: The performance gap is exacerbated by the composition of the official test set, where 4 out of 5 speakers were entirely absent from the DRAL corpus. This extreme speaker mismatch particularly disadvantages retrieval-based methods, which rely on finding similar examples in the datastore.

4.2 Ablation Study: Retrieval Design

Table 2 details the design choices made during the development phase (Internal Split). We explicitly compared the official winners indices against PCA-based dimensionality reduction to

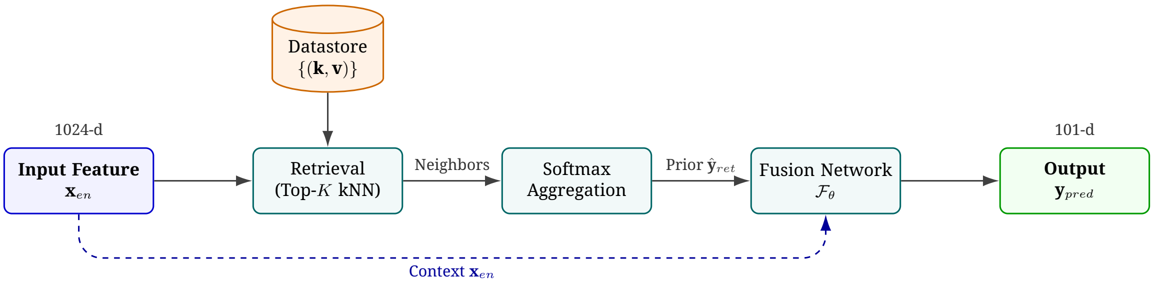

Figure 1: R-APM System Architecture. Config A retrieves in full 1024-dim space; Config B projects queries to 103-dim first (see Section 2.3).

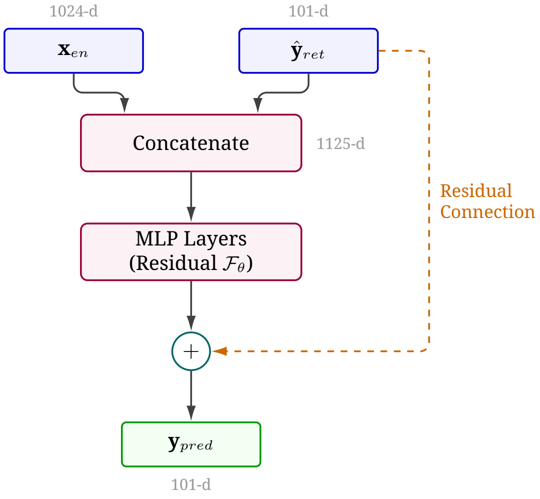

Figure 2: Fusion Network Design: A residual correction block where the network learns to correct the retrieved prior.

validate the feature space selection.

Table 2: **Internal Ablation Study (Seen Speakers).** Comparing aggregation strategies and feature spaces. PCA fails to capture semantic similarity compared to the official 'winner' indices.

Retrieval Setting	Cosine	Δ
Aggregation Strategy		
Hard Top-1	0.7859	-0.0883
Top-Truncated ($K = 50$)	0.8734	-0.0008
Softmax ($K = 70$, $\tau = 0.04$)	**0.8742**	Ref.
Feature Space		
PCA (101-dim)	0.4456	-0.4286
Official Winners (103-dim)	0.8741	-0.0001

The catastrophic failure of PCA (Cosine = 0.4456) confirms that variance-based dimensionality reduction destroys the pragmatic structure essential for prosody transfer. In contrast, retrieving within the subspace defined by the official `english_winners` (103-dim) preserves performance (0.8741) while filtering noise, which proves critical for generalization as shown in Table 1.

4.3 Hyperparameter Sensitivity

We performed a grid search to optimize K and τ. As shown in Table 3, performance saturates at $K = 70$ with a sharp temperature $\tau = 0.04$. Note that while Optimal Transport weights were explored, simple Softmax proved sufficient and more stable.

Table 3: **Hyperparameter Grid Search.** Pure Retrieval performance on Internal Split.

Top-K	Cosine	Temp (τ)	Cosine
$K = 1$	0.786	0.01	0.844
$K = 10$	0.864	0.03	0.871
$K = 30$	0.871	**0.04**	**0.872**
K=70	**0.872**	0.07	0.872
$K = 90$	0.874	0.10	0.871

4.4 The "Small Data Trap"

A key observation is the inefficiency of the Fusion Network (Table 1).

- **Seen Speakers (Config A):** The Fusion network improves performance by +0.0020.
- **Unseen Speakers (Config A):** The gain collapses to +0.0004.
- **Unseen Speakers (Config B):** The gain recovers slightly to +0.0013.

We term this the "Small Data Trap." With only 2,893 samples, the datastore is too sparse to provide a dense enough neighborhood for retrieval to work perfectly. Simultaneously, the parametric Fusion network (1125-dim input) has too few examples to learn a generalized correction function, instead overfitting to the training noise (Training Cosine > 0.999). The Baseline MLP likely outperforms our hybrid system because it learns a smoother, lower-variance manifold directly from the data, without being misled by noisy retrieved neighbors.

5 Discussion & Conclusion

Our investigation into retrieval-augmented prosody transfer yields both insights and a cautionary tale. On the positive side, our approach offers full transparency: for any prediction, practitioners can inspect the retrieved neighbors to understand why a particular prosodic pattern was generated. This interpretability is valuable for error analysis and system debugging, which is difficult with black-box MLP models.

However, this transparency comes at a cost. Retrieval is highly sensitive to the **sparsity of the datastore** and **identity shifts** in the embedding space. The severe speaker mismatch in the official test set (4/5 unseen speakers) particularly disadvantages our method, as it relies on finding similar examples that simply do not exist for new speakers. The underperformance compared to the MLP baseline suggests that for small-scale datasets ($\approx 3k$ pairs) with significant speaker variation at test time, direct parametric modeling may be more robust. Nevertheless, the superior generalization of Config B over Config A confirms that explicit subspace selection is crucial for mitigating identity overfitting in cross-lingual tasks. Future work should explore speaker normalization techniques and larger, more diverse datastores.

6 Reproducibility

The code and checkpoints are available at https://github.com/TheGrSun/Interspeech2026-TOPI-RAPM.git. We utilized the PyTorch framework on a single NVIDIA RTX 3090 GPU.

References

[1] N. G. Ward *et al.*, “The Interspeech 2026 challenge on transfer of pragmatic intent in speech-to-speech translation,” in *Proc. Interspeech*, 2026.

[2] N. G. Ward, J. E. Avila, E. Rivas, and D. Marco, “Dialogs reenacted across languages, version 2,” University of Texas at El Paso, Tech. Rep. UTEP-CS-23-27, 2023.

[3] Linguistic Data Consortium, “Dialogs reenacted across languages,” 2024, LDC Catalog No. LDC2024S08.

[4] J. Vazquez, “HuBERT-based models and evaluation strategies for pragmatically-faithful speech-to-speech translation,” Master’s thesis, University of Texas at El Paso, 2025.

[5] N. G. Ward, A. Segura, A. Ceballos, and D. Marco, “Towards a general-purpose model of perceived pragmatic similarity,” in *Proc. Interspeech*, 2024.

[6] N. G. Ward and D. Marco, “A collection of pragmatic-similarity judgments over spoken dialog utterances,” in *Proc. LREC-COLING*, 2024.

[7] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, “HuBERT: Self-supervised speech representation learning by masked prediction of hidden units,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[8] U. Khandelwal, O. Levy, D. Jurafsky, L. Zettlemoyer, and M. Lewis, “Generalization through memorization: Nearest neighbor language models,” in *Proc. ICLR*, 2020.