# Personalized Contrastive Learning for Dysarthric Keyword Recognition

**James Dana**

## Abstract

Dysarthric speech poses a persistent challenge for Automatic Speech Recognition (ASR). This study explores contrastive learning with targeted triplet sampling and curriculum strategies for keyword identification across healthy and dysarthric speech, balancing generalization and personalization. Contrastive models consistently outperformed non-contrastive baselines, with specific curricula proving effective for each task. However, performance was sensitive to random seed initialization, and the study was limited by dataset size. These results highlight the promise of contrastive learning for dysarthric speech while underscoring the need for larger datasets and more robust training protocols.

## 1 Introduction

Automatic Speech Recognition (ASR) systems often struggle to transcribe dysarthric speech accurately due to its high acoustic variability and reduced intelligibility. Unlike typical speech, dysarthric speech exhibits speaker-specific patterns shaped by neuromuscular impairments, which limits the effectiveness of standard ASR models. While commercial ASR systems perform well on typical datasets, they can reach word error rates as high as 60–90% on dysarthric speech, revealing a major gap in accessibility and robustness (Harrell, 2020).

A central challenge in speech recognition is the trade-off between generalization and personalization: should a model be optimized to perform well across many speakers, or tailored to individual users? Models that generalize well often struggle with speech that deviates from the training distribution, such as dysarthric or accented speech, while personalized models may fail to adapt to new voices. This tension underlies many of the difficulties commercial ASR systems face with atypical speech.

This project applies contrastive learning to keyword identification in dysarthric and non-dysarthric speech, building on prior work in representation learning for speech. The primary contribution is a novel triplet sampling strategy designed to reflect the generalization–personalization trade-off, encouraging models to learn both speaker-specific and speaker-independent representations. In addition, I explore curriculum learning as a means of structuring training to support this dual objective.

## 2 Related Work

Contrastive learning has been used in ASR to improve keyword detection, especially in settings with limited labeled data. Zhang et al. (2023) introduced Contrastive Augmentation-Based Keyword Spotting (CAB-KWS), a framework that integrates contrastive learning with data augmentation to boost keyword recognition performance (Zhang et al., 2023). By encouraging the model to distinguish between different instances of the same keyword, CAB-KWS improves generalization across acoustic variations.

Li et al. (2024) proposed Phoneme-Level Contrastive Learning (PLCL), a method that operates at the phoneme level to better distinguish acoustically similar keywords (Li et al., 2024). By leveraging phoneme-level representations, the model learns finer-grained distinctions between confusable words, which is critical for improving recognition accuracy in keyword spotting tasks.

In 2025, Lee et al. proposed DyPCL, a framework tailored for dysarthric speech recognition (Lee et al., 2025). A key innovation of DyPCL is its dynamic curriculum learning strategy, which progressively transitions from easy to difficult negative samples based on phonetic similarity. This progressive difficulty allows the model to first anchor basic phonetic distinctions before confronting more subtle variations—an approach well-suited to the high variability and reduced intelligibility of

dysarthric speech.

## 3 Dataset and Problem Setup

To balance generalization and personalization in dysarthric speech classification, I designed a structured set of seven triplet types based on alignment across speaker, word, and speech status.

This framework introduced constraints on keyword selection: each word needed sufficient representation across multiple speakers (ideally 3 or more instances per speaker) to support robust triplet generation and enable speaker-level evaluation. Due to limited data availability, I selected eight keywords that met these criteria and filtered participants accordingly.

The selected keywords (air, knew, leak, no, sigh, sip, slip, yes) are all short, high-frequency English words. Many share phonetic similarities, introducing meaningful acoustic confusability that makes them well-suited for evaluating contrastive learning's ability to distinguish fine-grained speech patterns. This setup yielded a Training dataset (n=314, 11 speakers, 8 keywords each) a Generalization Holdout (n=170, 4 speakers; all keywords except leak for dysarthric speakers) and a Personalization Holdout (n=292, 11 speakers; variable extra utterances). The task is to classify short speech segments into these eight keyword classes, based on audio embeddings derived from Wav2Vec2.

The selected keywords (air, knew, leak, no, sigh, sip, slip, yes) are short, high-frequency English words. Many share phonetic similarities, introducing acoustic confusability that makes them well-suited for evaluating contrastive learning's ability to distinguish fine-grained speech patterns. This setup yielded a Training set (n = 314, 11 speakers, 8 keywords each), a Generalization Holdout (n = 170, 4 speakers; all keywords except leak for dysarthric speakers), and a Personalization Holdout (n = 292, 11 speakers; variable extra utterances). The classification task is to assign short speech segments to one of the eight keyword classes using audio embeddings derived from Wav2Vec2.

Although the training set is balanced, the holdout sets exhibit class and group imbalance, particularly in the dysarthric subset. This imbalance is visualized in the accompanying heatmap and informed the choice to use weighted evaluation metrics.

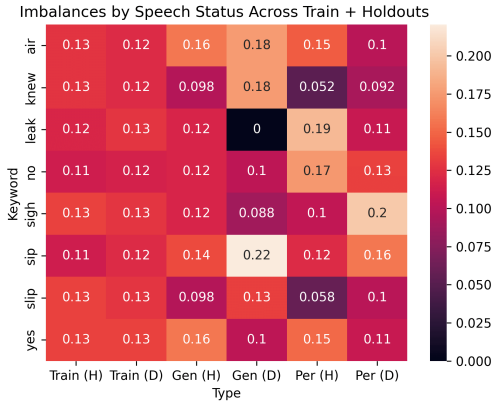


Figure 1: Keyword/Status Balance by Dataset

## 4 Methodology

### 4.1 Embedding and Baseline Classification

To establish a baseline for classification, I trained models using audio embeddings from a pretrained, unfinetuned Wav2Vec2 model, chosen for its robust performance on general speech tasks. Audio from the HuggingFace TORGO dataset, with no additional processing or augmentation, was padded or truncated to 60,000 samples, preserving over 96% of utterances and ensuring consistent input for Wav2Vec2.

Given the small training set (314 examples), I focused on lightweight architectures and regularization to mitigate overfitting. I experimented with both mean-pooled and unpooled embeddings. While a shallow feedforward network with pooled embeddings showed initial promise, a 1D convolutional network (Conv1D) on unpooled embeddings outperformed the feedforward approach, highlighting the importance of local temporal patterns for keyword classification.

The dense classifier used pooled embeddings passed through two ReLU-activated hidden layers (one variable sized and one 32 units), followed by batch normalization, dropout, and a softmax output layer. Hyperparameters were tuned using a small random search with n=3 cross-validation, focusing on the learning rate, L2 regularization, and the size of the first hidden layer. The Conv1D model utilized unpooled embeddings processed by two convolutional blocks (each with batch normalization, dropout, and pooling), followed by a global max pooling layer and a dense classifier. For both models, hyperparameters were selected based on n=3 cross-validation, with average validation accu-

racy used to choose the optimal parameters. Early stopping was applied based on validation accuracy, while the weighted F1-score was used for final performance evaluation, given its robustness in handling imbalanced data.

### 4.2 Triplet Construction and Contrastive Learning

Contrastive learning used four triplet types: Type 1 (same speaker, same word), Type 2 (same speaker, different word), Type 3 (different speaker, same word), and Type 4 (different speaker, different word). For triplet types involving multiple speakers (Types 1, 3, and 4), the construction considered whether the speakers had the same or different speech statuses (healthy or dysarthric), resulting in a total of 7 types.

Seven triplets were generated per audio event in the training set, using 176 unique audio events (314 files total, due to multi-microphone recordings), yielding 1,232 triplets. To mitigate sampling bias from unequal file counts across participants and microphones, a balanced sampling procedure was employed for both positive and negative selections, ensuring consistent representation across speakers, keywords, and speech conditions. After sampling, I verified that the resulting triplet set preserved the demographic and class distributions of the training data.

To train the embedding model, I used a margin-based triplet loss function with a margin parameter $\mu$ and $L^2$ norm distance. The loss encourages the model to pull the anchor closer to the positive sample (same class or speaker) than to the negative sample, by at least the margin. Formally, for each triplet $(a, p, n)$ of anchor, positive, and negative inputs, and embedding function $\theta$, the loss is defined as $L_{\text{triplet}}(s_a, s_p, s_n, \mu, \theta) = \max\left(|\theta(a) - \theta(p)|_2 - |\theta(a) - \theta(n)|_2 + \mu, 0\right)$.

I trained the plain Contrastive Learning model for 10 epochs using the Adam optimizer with a learning rate of $1 \times 10^{-5}$ and margin $\mu = 0.5$, selected from a small sweep over reasonable values. Due to limited data, I did not use a validation split during training; instead, I evaluated the learned embeddings post hoc using K-Nearest Neighbors classification on both the training and holdout sets, treating classification accuracy and F1-score as downstream indicators of embedding quality.

### 4.3 Curriculum Strategy

Three curricula were designed to gradually increase training complexity, based on triplet types and speaker status. These curricula aimed to expose the model to different levels of generalization and personalization, progressing from simpler to more complex patterns. The design was based on the assumption that triplet types represent varying levels of difficulty in generalization and personalization, enabling controlled exploration of these factors. By keeping hyperparameters consistent across all curricula, I aimed to isolate the effects of curriculum design on model performance.

The "H to All" curriculum consists of three 5-epoch stages, moving from healthy to dysarthric speakers and then a mixed dataset. Stage 1 (healthy only) focuses on general speech patterns. Stage 2 (dysarthric only) allows specialization in dysarthric features. Stage 3 (all speakers combined) aims to improve generalization across both groups. The intent is to observe model adaptation from general speech to dysarthric complexities and subsequent generalization.

The "1 to All" curriculum includes four 5-epoch stages, starting with Type 1 triplets (same word, same speaker) and gradually introducing more complex types. Stage 1 (Type 1 only) focuses on learning speaker-specific features. In Stage 2, Type 2 triplets (same speaker, different word) are added to encourage the model to distinguish between words spoken by the same person. Stage 3 introduces Type 3 triplets (different speaker, same word) to promote generalization across speakers. Stage 4 adds Type 4 triplets (different speaker, different word), which require the model to handle both speaker and word variability. This curriculum is designed to transition the model from learning speaker-specific features to generalizing across speakers and words.

The "4 to All" curriculum also includes four 5-epoch stages, beginning with the most complex Type 4 triplets (different speaker, different word). Stage 1 focuses on learning generalization across both speaker and word variability. Stage 2 introduces Type 3 triplets (different speaker, same word) to reduce complexity by emphasizing generalization across speakers for a consistent word. Stage 3 adds Type 2 triplets (same speaker, different word), allowing the model to learn to distinguish between different words from the same speaker. Stage 4 includes Type 1 triplets (same word, same

speaker), focusing on personalization by learning speaker-specific characteristics. This curriculum tests whether starting with complex generalization enhances the model's ability to handle personalization later in training.

### 4.4 Evaluation Strategy

The evaluation was conducted separately for the baseline and contrastive learning models, using classification strategies suited to each approach.

For the baseline models, training was performed using frozen Wav2Vec2 embeddings. These embeddings, extracted from a pre-trained model, remained fixed during training. A supervised classifier was then trained on top of these embeddings to predict word class labels. This setup follows a standard supervised learning pipeline, where the classifier learns directly from labeled examples without modifying the underlying audio representations.

In contrast, the contrastive learning models fine-tuned a Wav2Vec2 model using a margin-based triplet loss, which encourages the model to pull similar samples closer and push dissimilar ones further apart. After fine-tuning, new embeddings were generated for all audio files, and a K-Nearest Neighbors (KNN) classifier was applied to evaluate these embeddings. The number of neighbors $K$ was set to match the number of word classes.

KNN was chosen because it is a simple, non-parametric method that directly probes the embedding space. Its decision-making process mirrors the contrastive learning objective by relying on local neighborhood structure, making it an interpretable choice for evaluating the quality of embeddings without adding complexity.

Each model was trained with three random seeds (100, 200, 300). For each run, I recorded weighted F1 scores across three datasets (Train, Generalization Holdout, and Personalization Holdout) and three speaker subsets (All Speakers, Healthy Speakers, and Dysarthric Speakers). Model performance was reported based on the best seed for each model, prioritized by the F1 score for dysarthric speakers on the relevant holdout set, emphasizing performance for the primary group of interest.

## 5 Results and Discussion

### 5.1 Generalization Task

Table 1 highlights that contrastive learning (CL) models consistently outperformed the non-contrastive baseline models in the generalization task across all speaker groups. The Dense model showed poor performance overall (F1: 0.5345), particularly struggling with dysarthric speakers (F1: 0.4696). The Conv1D model performed better with an overall F1 score of 0.8122, achieving strong results for healthy speakers (F1: 0.9227). However, its generalization to dysarthric speakers was weaker, with an F1 score of 0.6810.

| Model | Seed | All Speakers | Healthy Speakers | Dysarthric Speakers |
|---|---|---|---|---|
| Dense | 300 | 0.5345 | 0.5848 | 0.4696 |
| Conv1D | 100 | 0.8122 | 0.9227 | 0.681 |
| CL No Curr | 300 | **0.9249** | 0.9508 | **0.8907** |
| CL "H to All" | 300 | 0.8248 | 0.8873 | 0.7416 |
| CL "1 to All" | 300 | 0.8770 | **0.9707** | 0.7339 |
| CL "4 to All" | 300 | 0.9229 | 0.9499 | 0.8834 |

Table 1: Best Seeds per Model, Selected on Generalization performance. Weighted F1 scores are shown for each speaker group (All, Healthy, Dysarthric). **Bolded** values mark best performance per group.

Among the CL models, CL No Curr achieved the highest overall performance (F1: 0.9249) and excelled in generalization to dysarthric speakers (F1: 0.8907). CL "4 to All" also demonstrated strong generalization (F1: 0.9229), particularly for dysarthric speakers (F1: 0.8834), and was competitive with CL No Curr in this regard. In contrast, the curricula models, CL "H to All" (F1: 0.7416) and CL "1 to All" (F1: 0.7339), showed lower performance on dysarthric speech, with CL "1 to All" performing best on healthy speakers (F1: 0.9707) but falling behind on dysarthric ones.

Overall, the CL "4 to All" curriculum achieved the most balanced performance, particularly excelling with dysarthric speakers, suggesting that starting with more complex generalization tasks and progressing towards personalization is effective for generalizing to unseen dysarthric speakers. However, CL No Curr outperformed the curricula models in terms of overall and dysarthric speaker generalization, highlighting the importance of fine-tuning without curriculum constraints for this task.

### 5.2 Personalization Task

Table 2 demonstrates that contrastive learning (CL) models again outperform the non-contrastive baselines in the personalization task, with the CL "1 to All" model achieving the highest overall performance (F1: 0.9458) and excelling across all speaker groups, particularly dysarthric speakers (F1: 0.9096).

Among the non-contrastive models, the Conv1D

| Model | Seed | All Speakers | Healthy Speakers | Dysarthric Speakers |
|---|---|---|---|---|
| Dense | 200 | 0.570 | 0.6063 | 0.5098 |
| Conv1D | 100 | 0.7288 | 0.7774 | 0.6594 |
| CL No Curr | 300 | 0.9132 | 0.9412 | 0.8715 |
| CL "H to All" | 200 | 0.9239 | 0.9471 | 0.8913 |
| CL "1 to All" | 200 | **0.9458** | **0.971** | **0.9096** |
| CL "4 to All" | 100 | 0.9311 | 0.9589 | 0.8919 |

Table 2: Best Seeds per Model, Selected on Personalization performance. Weighted F1 scores are shown for each speaker group (All, Healthy, Dysarthria). **Bolded** values mark best performance per group.

model (F1: 0.7288) showed better results than the Dense model (F1: 0.570), especially for healthy speakers (F1: 0.7774 vs. 0.6063). However, the contrastive models significantly outperformed these, with CL No Curr and CL "H to All" showing strong performance for all groups, especially for dysarthric speakers. The CL "4 to All" curriculum also showed competitive results (F1: 0.9311), though it trailed behind CL "1 to All" in both overall and dysarthric speaker performance.

In summary, CL "1 to All" performed the best overall and for dysarthric speakers, suggesting that gradually progressing from speaker-specific to more complex generalization tasks helps improve personalization performance.

### 5.3 Comparison of Generalization and Personalization

Across both the generalization and personalization tasks, contrastive learning (CL) models consistently outperformed the non-contrastive baselines. The largest improvements were seen in the personalization task, where the model focused on learning individual speaker characteristics. As expected, personalization was less challenging, allowing the CL models—especially those trained with the "1 to All" curriculum—to achieve very high F1 scores, ranging from the mid- to high-0.9s for both healthy and dysarthric speakers.

In contrast, the generalization task, which focused on performance with unseen speakers, showed a different trend among the CL approaches. While CL "1 to All" performed well in personalization, its generalization performance lagged behind that of CL No Curr and CL "4 to All". Specifically, for all speakers, CL No Curr and CL "4 to All" reached F1 scores of 0.9249 and 0.9229, respectively, while CL "1 to All" scored 0.8770. A similar pattern emerged for dysarthric speakers, where CL No Curr (F1: 0.8907) and CL "4 to All" (F1: 0.8834) outperformed CL "1 to All" (F1: 0.7339).

Conversely, the generalization task (focused on performance on unseen speakers) revealed a different performance pattern among the curriculum learning approaches. While CL "1 to All" excelled in personalization, its generalization performance was notably lower compared to "CL No Curr" and "CL '4 to All'". Specifically, for all speakers, "CL No Curr" and "CL '4 to All'" attained F1 scores of 0.9249 and 0.9229, respectively, while "CL '1 to All'" achieved 0.8770. A similar trend was evident for dysarthric speakers, where "CL No Curr" (F1: 0.8907) and "CL '4 to All'" (F1: 0.8834) outperformed "CL '1 to All'" (F1: 0.7339).

This divergence in performance suggests that while a curriculum emphasizing initial personalization ("1 to All") is highly effective for adapting to individual speakers, CL models trained without a specific curriculum ("No Curr") or with a curriculum that progresses from general patterns to personalization ("4 to All") appear to achieve a more favorable balance between capturing overarching speech structures and ensuring robust generalization to unseen speakers.

### 5.4 Keyword Confusion Patterns

Analysis of classification performance on individual keywords (see Appendix C for full confusion matrices) revealed consistent confusion patterns across the contrastive learning models. As expected, acoustically similar words were often misclassified, such as 'sip' vs. 'slip', 'knew' vs. 'no', and 'slip' vs. 'leak'. Some less intuitive confusions also emerged, including between 'air' and 'yes'. To address potential class imbalance in the keyword distribution, the weighted F1 score was used as the primary evaluation metric, ensuring a balanced performance assessment across all keywords.

It is important to note a limitation in the dataset: the keyword 'leak' was not included in the generalization holdout set for dysarthric speakers. This omission likely influenced the model's performance on this keyword for that group during the generalization task, impacting the reported F1 scores, even with the use of the weighted metric.

### 5.5 Impact of Seed Initialization

The results presented in Tables 1 and 2 highlight the best performance achieved by each model based on seed selection. However, examination of the full seed performance data (See Appendix B) reveals a significant impact of the initial random seed

[100, 200, 300] on the training and subsequent generalization and personalization capabilities of all models. As illustrated in those tables, substantial performance variations were observed across different seeds for the same model and task. For instance, the Dense model's Generalization F1 for dysarthric speakers ranged from 0.3704 to 0.4696, while for CL '4 to All', the Train F1 score itself varied considerably (e.g., from 0.6879 to 0.9554 for all speakers).

### 5.6 Impact of Seed Initialization

The results in Tables 1 and 2 reflect the best performance achieved for each model based on seed selection. However, a review of the full results across all seeds (see Appendix B) reveals that random seed initialization (100, 200, 300) had a substantial effect on model performance across both generalization and personalization tasks. For example, the Dense model's generalization F1 score for dysarthric speakers ranged from 0.3704 to 0.4696 depending on the seed. Similarly, CL "4 to All" showed considerable variation in training performance across seeds (e.g., F1 scores from 0.6879 to 0.9554 for all speakers).

We observed a consistent trend: seeds that yielded lower training performance typically also resulted in weaker generalization and personalization outcomes. This suggests that seed effects are not merely introducing variance in overfitting, but are instead influencing the model's ability to learn meaningful representations from the training data in the first place. This sensitivity is likely exacerbated by the relatively small size of our datasets (314 training samples per slice, 170 in the generalization holdout, and 292 in the personalization holdout) which makes the models more vulnerable to stochastic variation in early training dynamics.

## 6 Limitations and Future Work

Several limitations in this study may affect the interpretation and generalizability of the findings:

**Data Constraints**: The limited size of the training (n=314) and holdout datasets (generalization: n=170; personalization: n=292) constrained both model complexity and evaluation robustness. The training and personalization splits were also imbalanced in terms of dysarthric versus healthy speakers (6:5), which, despite the use of weighted F1 scores, may have biased model performance. The generalization holdout set was more balanced, but the overall number of participants was small (15 total; 4 held out), and no analysis of dysarthria severity was included—limiting the representativeness of the sample.

**Curriculum Learning Design and Tuning**: Due to time constraints, the curriculum learning framework was not extensively tuned. Fixed triplet sets were used throughout training, rather than dynamic triplet mining strategies used in the contrastive learning literature I explored. This may have limited the potential effectiveness of the curriculum structures tested.

**Model Sensitivity to Seed Initialization**: All models exhibited high sensitivity to random seed initialization, leading to considerable performance variance across runs. This raises concerns about stability and reproducibility, and suggests that the reported best-seed results may not reflect average-case performance.

Future work should focus on expanding and diversifying the dataset to include more participants with varied dysarthria severities and demographic backgrounds. Dynamic curriculum learning strategies and more comprehensive hyperparameter tuning should be explored to better capture the benefits of curriculum design. Additional studies should incorporate multiple random seed runs to assess performance stability, investigate the influence of dysarthria severity, evaluate alternative model architectures, and explore advanced data augmentation techniques to enhance generalization and robustness.

## 7 Conclusion

In summary, this study explored contrastive learning and curriculum strategies for classifying keywords in both healthy and dysarthric speech. Contrastive models consistently outperformed non-contrastive baselines across generalization and personalization tasks, with certain curricula showing notable promise. However, limitations such as small, imbalanced datasets, model sensitivity to seed initialization, and the use of fixed triplet sets point to key directions for future work. Expanding the dataset to include more diverse speakers and severity levels, and adopting more adaptive curriculum methods, will be critical for improving robustness and generalizability.

## References


Erica Harrell. 2020. Recognizing atypical speech is asr's achilles' heel. *Speech Technology Magazine*. Accessed: 2025-05-06.

Yeong-Hun Lee, Seong-Ho Kim, Youngik Lee, and Sang-Hoon Lee. 2025. Dypcl: Dynamic phoneme-level contrastive learning for dysarthric speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

Wei Li, Zixuan Zhao, Qing Liu, Hui Lin, Lei Yu, and Lei Zhang. 2024. Phoneme-level contrastive learning for low-resource keyword spotting. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1231–1235. IEEE.

Ruixiang Zhang, Chuan Dong, Lei Zhang, Jia Huang, Hui Lin, Lei Yu, and Jianshu Li. 2023. Cab-kws: Contrastive augmentation-based keyword spotting with limited data. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 122–126. ISCA.


## A Figures

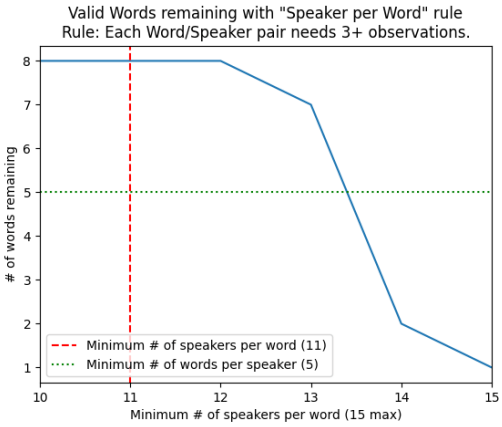


Figure 2: Tradeoffs in TORGO Keyword/Speaker Selection

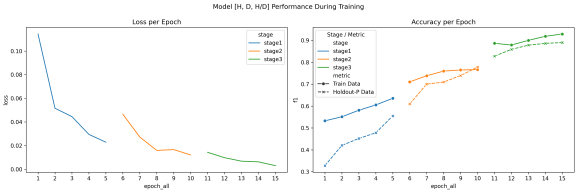

Figure 3: "H to All" Model: Best Seed (Gen) Learning over Epochs

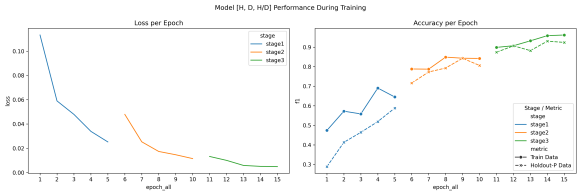

Figure 4: "H to All" Model: Best Seed (Per) Learning over Epochs

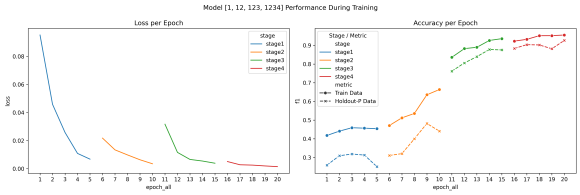

Figure 5: "1 to All" Model: Best Seed (Gen) Learning over Epochs

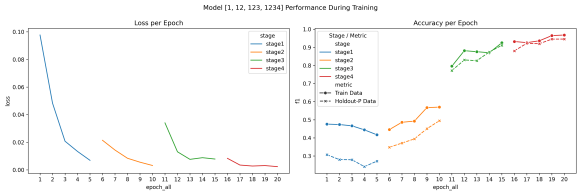

Figure 6: "1 to All" Model: Best Seed (Per) Learning over Epochs

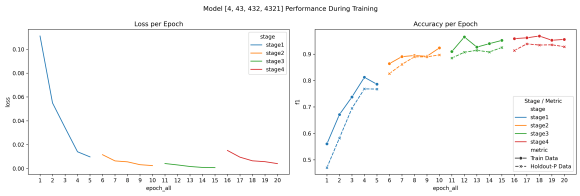

Figure 7: "4 to All" Model: Best Seed (Gen) Learning over Epochs

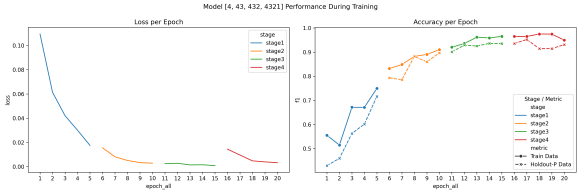

Figure 8: "4 to All" Model: Best Seed (Per) Learning over Epochs

## B Accuracy Tables

| Seed | Speaker | Train F1 | Gen F1 | Per F1 |
|---|---|---|---|---|
| 100 | All | 0.8387 | 0.4895 | 0.5553 |
| | Healthy | 0.7849 | 0.5741 | 0.5971 |
| | Dysarthria | 0.8809 | 0.3704 | 0.4844 |
| **200** | All | 0.8539 | 0.5433 | 0.57 |
| | Healthy | 0.8245 | 0.6536 | 0.6063 |
| | Dysarthria | 0.8771 | 0.3751 | **0.5098** |
| **300** | All | 0.8661 | 0.5345 | 0.5686 |
| | Healthy | 0.8662 | 0.5848 | 0.605 |
| | Dysarthria | 0.8661 | **0.4696** | 0.5091 |

Table 3: Dense baseline performance by seed. Weighted F1 scores are shown for each speaker group (All, Healthy, Dysarthria) on training, generalization (Gen), and personalization (Per) sets. **Bolded** values mark best Dysarthric performance in each holdout.

| Seed | Speaker | Train F1 | Gen F1 | Per F1 |
|---|---|---|---|---|
| **100** | All | 0.9302 | 0.8122 | 0.7288 |
| | Healthy | 0.9578 | 0.9227 | 0.7774 |
| | Dysarthria | 0.9069 | **0.681** | **0.6594** |
| 200 | All | 0.9077 | 0.6238 | 0.6721 |
| | Healthy | 0.9372 | 0.7283 | 0.6988 |
| | Dysarthria | 0.8835 | 0.4662 | 0.6284 |
| 300 | All | 0.8585 | 0.6096 | 0.637 |
| | Healthy | 0.9099 | 0.7348 | 0.6878 |
| | Dysarthria | 0.8159 | 0.4195 | 0.5631 |

Table 4: Conv1D baseline performance by seed. Weighted F1 scores are shown for each speaker group (All, Healthy, Dysarthric) on training, generalization (Gen), and personalization (Per) sets. **Bolded** values mark best Dysarthric performance in each holdout.

| Seed | Speaker | Train F1 | Gen F1 | Per F1 |
|---|---|---|---|---|
| 100 | All | 0.949 | 0.7972 | 0.8766 |
| | Healthy | 0.9577 | 0.8599 | 0.9137 |
| | Dysarthria | 0.9418 | 0.7048 | 0.8235 |
| 200 | All | 0.9519 | 0.9065 | 0.9283 |
| | Healthy | 0.9717 | 0.9702 | 0.9884 |
| | Dysarthria | 0.9356 | 0.8263 | 0.8441 |
| **300** | All | 0.9319 | 0.9249 | 0.9132 |
| | Healthy | 0.9436 | 0.9508 | 0.9412 |
| | Dysarthria | 0.9217 | **0.8907** | **0.8715** |

Table 5: Contrastive Learning performance by seed. Weighted F1 scores are shown for each speaker group (All, Healthy, Dysarthric) on training, generalization (Gen), and personalization (Per) sets. **Bolded** values mark best Dysarthric performance in each holdout.

| Seed | Speaker | Train F1 | Gen F1 | Per F1 |
|---|---|---|---|---|
| 100 | All | 0.8936 | 0.8318 | 0.8517 |
| | Healthy | 0.9282 | 0.9059 | 0.9064 |
| | Dysarthria | 0.8604 | 0.7336 | 0.7748 |
| **200** | All | 0.9618 | 0.8469 | 0.9239 |
| | Healthy | 0.9432 | 0.9079 | 0.9471 |
| | Dysarthria | 0.9767 | 0.7175 | **0.8913** |
| **300** | All | 0.9288 | 0.8248 | 0.8901 |
| | Healthy | 0.9713 | 0.8873 | 0.9356 |
| | Dysarthria | 0.8915 | **0.7416** | 0.8218 |

Table 6: CSCL "H to All" performance by seed. Weighted F1 scores are shown for each speaker group (All, Healthy, Dysarthric) on training, generalization (Gen), and personalization (Per) sets. **Bolded** values mark best Dysarthric performance in each holdout.

| Seed | Speaker | Train F1 | Gen F1 | Per F1 |
|---|---|---|---|---|
| 100 | All | 0.958 | 0.8878 | 0.9249 |
| | Healthy | 0.9929 | 0.9902 | 0.9719 |
| | Dysarthria | 0.9275 | 0.7233 | 0.8578 |
| **200** | All | 0.9682 | 0.8682 | 0.9458 |
| | Healthy | 0.9792 | 0.9703 | 0.971 |
| | Dysarthria | 0.9592 | 0.7087 | **0.9096** |
| **300** | All | 0.9553 | 0.8770 | 0.9262 |
| | Healthy | 0.9508 | 0.9707 | 0.9664 |
| | Dysarthria | 0.959 | **0.7339** | 0.8697 |

Table 7: CSCL "1 to All" performance by seed. Weighted F1 scores are shown for each speaker group (All, Healthy, Dysarthric) on training, generalization (Gen), and personalization (Per) sets. **Bolded** values mark best Dysarthric performance in each holdout.

| Seed | Speaker | Train F1 | Gen F1 | Per F1 |
|---|---|---|---|---|
| **100** | All | 0.9495 | 0.8512 | 0.9311 |
| | Healthy | 0.972 | 0.8681 | 0.9589 |
| | Dysarthria | 0.9312 | 0.8272 | **0.8919** |
| 200 | All | 0.6879 | 0.6055 | 0.6546 |
| | Healthy | 0.7245 | 0.7143 | 0.7162 |
| | Dysarthria | 0.653 | 0.4828 | 0.5728 |
| **300** | All | 0.9554 | 0.9229 | 0.9281 |
| | Healthy | 0.9716 | 0.9499 | 0.9712 |
| | Dysarthria | 0.9425 | **0.8834** | 0.867 |

Table 8: CSCL "4 to All" performance by seed. Weighted F1 scores are shown for each speaker group (All, Healthy, Dysarthric) on training, generalization (Gen), and personalization (Per) sets. **Bolded** values mark best Dysarthric performance in each holdout.

# C Confusion Matrices

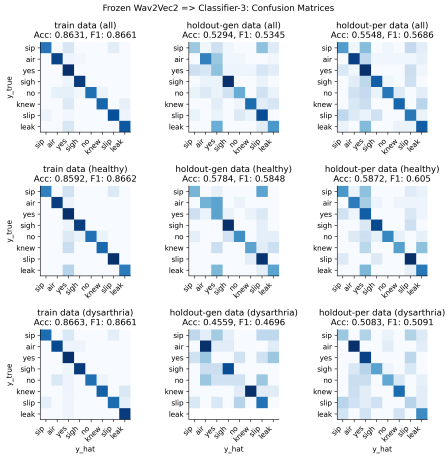


Figure 9: Dense Model: Best Seed Confusion Matrix (Gen)

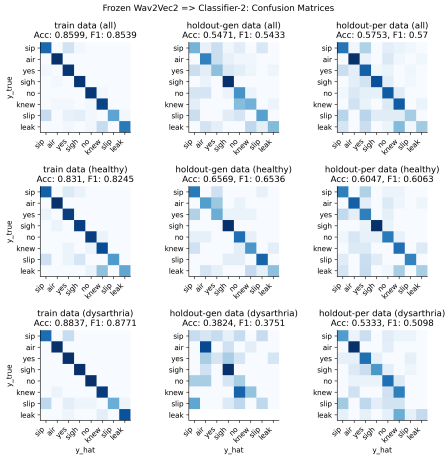


Figure 10: Dense Model: Best Seed Confusion Matrix (Per)

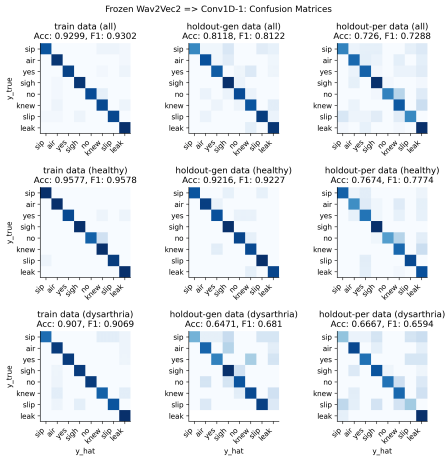


Figure 11: Conv1D Model: Best Seed Confusion Matrix (Gen+Per)

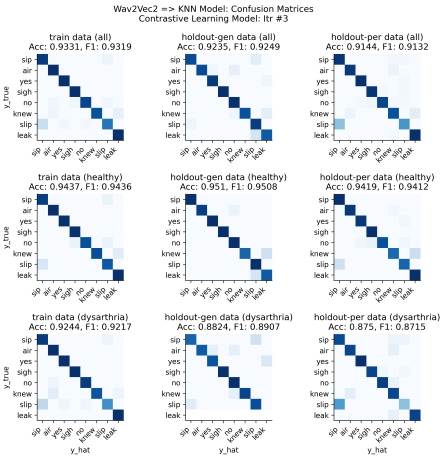


Figure 12: Contrastive Model: Best Seed Confusion Matrix (Gen+Per)

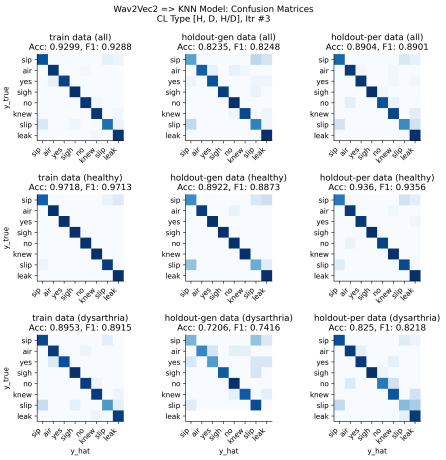


Figure 13: "H to All" Model: Best Seed Confusion Matrix (Gen)

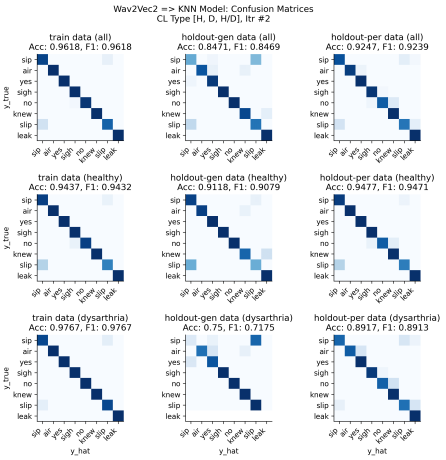


Figure 14: "H to All" Model: Best Seed Confusion Matrix (Per)

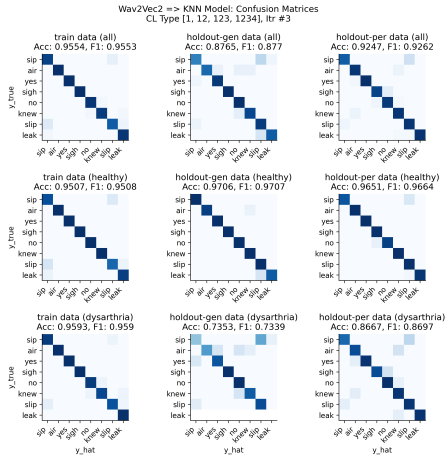


Figure 15: "1 to All" Model: Best Seed Confusion Matrix (Gen)

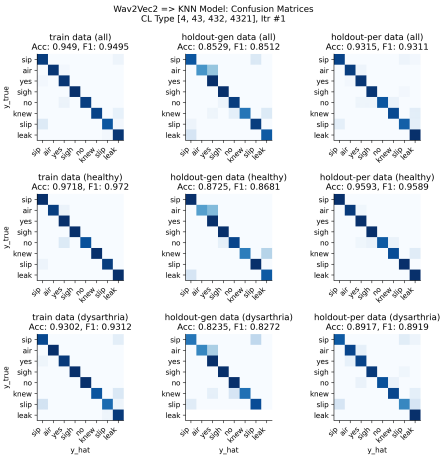


Figure 18: "4 to All" Model: Best Seed Confusion Matrix (Per)

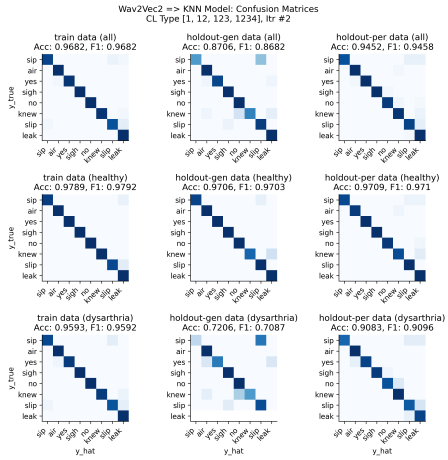


Figure 16: "1 to All" Model: Best Seed Confusion Matrix (Per)

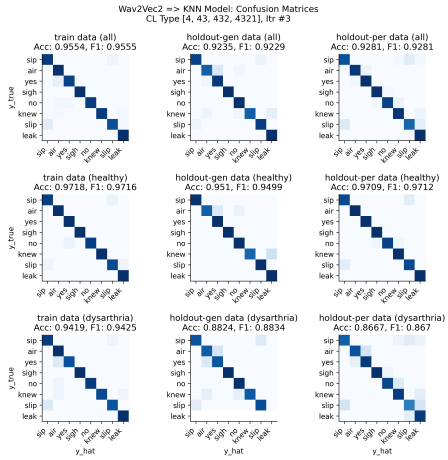


Figure 17: "4 to All" Model: Best Seed Confusion Matrix (Gen)