**Phonological Level wav2vec2-based Mispronunciation Detection and Diagnosis Method**

**Objective:** This paper proposes a novel low-level Mispronunciation Detection and Diagnosis (MDD) approach based on phonological features rather than traditional phoneme-level analysis. The method aims to provide more detailed diagnostic information for Computer-Aided Pronunciation Learning (CAPL) systems by detecting articulatory components that form phonemes, enabling more formative feedback for second-language (L2) learners.

Current phoneme-level MDD methods can only detect categorical pronunciation errors that exist in training data and provide limited diagnostic information. The proposed phonological-level approach addresses these limitations by analyzing elementary components of sound production (manners and places of articulation) that are directly related to the articulatory system.

*Block diagram showing phoneme vs. phonological level MDD approaches with wav2vec2-based processing*

**Architectures:**

* **Core Model:** Pre-trained wav2vec2 model (base, large, and large-robust variants) used as feature extractor
* **Phonological Feature Detection:** Multi-label Connectionist Temporal Classification (CTC) approach to jointly model 35 non-mutually exclusive phonological features
* **Novel SCTC-SB:** Separable CTC with Shared Blank approach for handling multi-label classification
* **Baseline Comparison:** Traditional phoneme-level MDD using CTC-based sequence-to-sequence classification

**Related Work:** Traditional MDD approaches include scoring methods (GOP algorithm), classification approaches, rule-based methods (Extended Recognition Network), and free-phoneme recognition. These methods face limitations:

* Can only detect categorical errors present in training data
* Require large amounts of annotated mispronounced speech
* Cannot handle uncategorical errors or distorted phonemes
* Provide limited diagnostic feedback

Phonological feature modeling has been explored for ASR improvement and language identification, but typically requires frame-level annotation and multiple separate models for different features. This work introduces a unified approach using multi-label CTC for efficient joint modeling.

**Method:**

**A. Multi-label CTC for Phonological Features** The paper introduces Separable CTC with Shared Blank (SCTC-SB) to handle the non-mutually exclusive nature of phonological features. Unlike traditional approaches requiring separate models for each feature, this method uses a single network with 71 output nodes (35 for +att, 35 for -att, 1 shared blank).

**B. Phonological Feature Set** 35 phonological features covering:

* **Manners:** consonant, sonorant, fricative, nasal, stop, approximant, affricate, liquid, vowel, semivowel, continuant
* **Places:** alveolar, palatal, dental, glottal, labial, velar, mid, high, low, front, back, central, anterior, posterior, retroflex, bilabial, coronal, dorsal
* **Others:** long, short, monophthong, diphthong, round, voiced

**C. Training Procedure**

* wav2vec2 CNN encoder parameters frozen during fine-tuning
* Linear layer added with 71 output nodes
* SpecAugment applied for data augmentation
* AdamW optimization with 0.0001 learning rate, 32 batch size, 30 epochs

**Datasets:**

**Training Data:**

* **TIMIT:** 3.9 hours, 462 native speakers
* **LibriSpeech-clean-100:** 100 hours, 251 native speakers
* **TIMIT+L2:** 6.5 hours, 480 speakers (native + non-native)

**Testing Data:**

* **Native:** LibriSpeech test sets, TIMIT test set
* **Non-native:** L2-ARCTIC corpus
  + **L2-Scripted:** 0.11 hours, 6 speakers
  + **L2-Suitcase:** 0.87 hours, 6 speakers (spontaneous speech)
* **Coverage:** 24 L2 speakers from 6 native languages (Arabic, Hindi, Korean, Mandarin, Spanish, Vietnamese)

**Error Distribution in L2-ARCTIC:**

* Training: 79,864 correct, 10,474 substitutions, 772 insertions, 2,437 deletions (L2-Scripted)
* Testing: 28,331 correct, 3,198 substitutions, 214 insertions, 939 deletions (L2-Scripted)

**Training:** **Experimental Setup:**

* 10-fold cross-validation methodology for robust evaluation
* Strict speaker separation between training and test sets
* Three wav2vec2 variants compared: base (95M parameters), large (317M), large-robust (317M)
* Models fine-tuned on different domain combinations to test robustness

**Evaluation Metrics:**

* **Feature Recognition:** Feature Error Rate (FER), Accuracy, Precision, Recall, F1-score
* **MDD Performance:** False Acceptance Rate (FAR), False Rejection Rate (FRR), Diagnostic Error Rate (DER)
* Metrics computed at both phoneme and phonological feature levels

**Results:**

**Phonological Feature Recognition Performance:**

* **Best Model:** wav2vec2-large-robust achieved 97.0% ± 0.65% average accuracy on TIMIT
* **Domain Robustness:** Performance dropped 31-45% when applying LS-clean trained models to LS-other data
* **Out-of-domain:** 22-33% degradation when testing cross-domain (LS→TIMIT)

**MDD Comparison Results:**

| **Method** | **Test Set** | **FAR (%)** | **FRR (%)** | **DER (%)** |
| --- | --- | --- | --- | --- |
| **Phoneme-level (TIMIT-L2)** | L2-Scripted | 39.51 | 7.31 | 15.79 |
|  | L2-Suitcase | 25.75 | 12.78 | 17.23 |
| **Phonological-level (TIMIT-L2)** | L2-Scripted | <30\* | <10\* | <10\* |
|  | L2-Suitcase | <30\* | <15\* | <15\* |

\*Average across all 35 phonological features

**Key Findings:**

* **Superior Detection:** Phonological-level MDD achieved significantly lower FAR compared to phoneme-level (63% vs <30%)
* **Better Diagnosis:** DER reduced from 31% (phoneme) to <10% (phonological) for L2-Scripted
* **Substitution Analysis:** 29 out of 34 common phoneme confusions better discriminated by at least one phonological feature
* **Domain Adaptability:** Phonological features more robust to domain mismatch, especially for out-of-domain scenarios

**Advantages of Phonological Approach:**

1. **Detailed Feedback:** Provides articulatory-level error description (e.g., voicing errors, place of articulation mistakes)
2. **Training Efficiency:** Can be trained solely on correctly pronounced speech
3. **Universal Features:** Phonological features shared across languages enable multi-lingual training
4. **Better Generalization:** More robust to unseen pronunciation variations

**Recommendations:** The paper demonstrates that phonological-level MDD offers superior performance and more informative feedback compared to traditional phoneme-level approaches. Future research directions include:

* Incorporating multi-language training data to leverage universal nature of phonological features
* Extending to more challenging domains (adult/child disordered speech)
* Developing real-time applications with instant corrective feedback
* Exploring multimodal fusion with visual articulatory information

The work establishes a new benchmark for phonological feature detection and provides a practical solution for more effective pronunciation learning systems.