**Research 1: Multimodal, Multi-Class Bias Mitigation for Predicting Speaker Confidence**

**🎯 Purpose**

* **In contexts like interviews or presentations, speaker confidence is a key factor.**
* **Automatic systems that predict confidence can be biased against:**
  + **Gender (e.g., rating women lower than men).**
  + **Accents (non-native speakers).**
  + **Age groups.**
* **Such bias leads to unfair feedback and unequal opportunities.**
* **This study develops a fair multimodal model to predict speaker confidence while reducing bias.**

**📂 Dataset**

* **Collected data from 260 speakers via Amazon Mechanical Turk (U.S.-based).**
* **Each participant recorded a 2-minute video using camera + microphone.**
* **Total dataset: 1,891 videos.**
* **For experiments: 233 videos selected.**
* **Videos annotated into three confidence levels:**
  + **Low**
  + **Medium**
  + **High**

**🧾 Annotation Process**

* **Partnered with Scale.ai: each video received 10 human ratings.**
* **Used a structured rubric with:**
  + **Eye gaze: looking at camera or not.**
  + **Gestures: natural vs. chaotic.**
  + **Posture: upright vs. tense.**
  + **Voice variation: monotone vs. confident.**
  + **Facial expressions: relaxed vs. anxious.**
  + **Pace: rushed vs. steady.**

**👥 Perceived Demographics**

* **Instead of self-reported data, raters estimated:**
  + **Gender: Male / Female.**
  + **Race: White / Non-White.**
  + **Age: <35 / ≥35.**
  + **Accent: Native / Non-Native.**
* **Purpose: test whether the system shows demographic bias (e.g., systematically rating females or non-native speakers lower).**

**🛠️ Multimodal Features Extracted**

* **Vision:**
  + **Eye movement anomalies.**
* **Speech:**
  + **Response length.**
  + **Number of pauses.**
  + **Hesitation rate.**
  + **Number of fillers (*uh, um*).**
  + **Speaking rate.**

**🤖 Models**

* **Used XGBoost classifier for prediction.**
* **Two versions tested:**
  1. **Unmitigated: baseline model (no fairness constraints).**
  2. **Mitigated: fairness-aware model using FairLearn library.**
* **FairLearn applies algorithms that adjust predictions to reduce bias across sensitive demographic groups.**

**📊 Results**

* **Without mitigation: noticeable demographic bias.**
  + **Example: females received systematically lower accuracy scores than males.**
* **With mitigation: performance improved, predictions became more balanced.**
  + **Non-White females:**
    - **Accuracy increased from 0.608 → 0.650.**
    - **F1-score increased from 0.548 → 0.631.**

**📌 Conclusion**

* **Standard AI systems for predicting confidence can be biased against gender, race, age, or accent.**
* **This research shows that:**
  + **Multimodal features (vision + speech) improve prediction.**
  + **Bias mitigation (FairLearn) significantly reduces unfairness.**
* **The result is a fairer system for assessing confidence, ensuring equal feedback and opportunities across diverse speaker groups.**

**Research 2: Phonological-Level Mispronunciation Detection and Diagnosis**

**🔍 Purpose of the Research**

**Most systems for Computer-Assisted Pronunciation Learning (CAPL) detect errors at the phoneme level.**

**But these approaches have major problems:**

* **They only detect errors if enough training data exists for that phoneme.**
* **They cannot capture *out-of-inventory* errors (sounds not present in English).**
* **They provide weak feedback: the learner only knows *which phoneme was wrong*, not *how* it was wrong.**

**✅ Proposed solution: detect errors at the phonological feature level (voicing, nasality, manner/place of articulation, etc.).  
This allows the system to say *how exactly* the learner went wrong.**

**📊 Datasets**

**They used 3 English corpora:**

* **LibriSpeech (LS): Native English audiobooks.**
* **TIMIT: Native English benchmark dataset.**
* **L2-ARCTIC (L2): Non-native English from 6 L1 backgrounds (Arabic, Hindi, Korean, Chinese, Spanish, Vietnamese).**
* **Total: 27 hours of audio, with 3.5 hours manually phoneme-annotated.**
* **Speakers: 18 for training, 6 for testing.**

**⚙️ Models**

**1. Phoneme-Level MDD (Baseline)**

* **Based on Wav2Vec2 → maps speech to phoneme sequences.**
* **Compares with reference → detects phoneme-level errors.**

**2. Phonological-Level MDD (Proposed)**

* **Also based on Wav2Vec2 + Linear Layer.**
* **Predicts 35 phonological features (binary labels: + / –).**
* **Uses SCTC-SB Loss (multi-label version of CTC loss).**
* **Each utterance is represented as sequences of features, not just phonemes.**

**🏆 Results**

* **Phonological-Level model outperforms Phoneme-Level:**
  + **FAR (False Acceptance Rate): ↓ from ~63% → <30%.**
  + **DER (Diagnostic Error Rate): ↓ from ~31% → <10%.**
* **Adding non-native data (L2-ARCTIC) further improved performance:**
  + **Training on TIMIT only: FAR = 18%.**
  + **Training on TIMIT + L2: FAR ↓ to 7%, DER ↓ to 15%.**

**✅ Key advantage: phonological features generalize well, even with limited non-native data.**

**📌 Example**

**Learner says “bat” instead of “pat”:**

* **Phoneme-Level system → “You mispronounced /p/ as /b/.”**
* **Phonological-Level system → “The problem is *voicing*: you made the sound voiced instead of voiceless.”**

**The second feedback is more actionable for learners.**

**🧩 Conclusion**

**This study addresses the limitation of phoneme-level detection, which lacks useful feedback.**

* **Method: Wav2Vec2 + SCTC-SB Loss at phonological feature level (35 features).**
* **Datasets: LibriSpeech, TIMIT, L2-ARCTIC.**
* **Results: much lower FAR and DER, with clearer diagnostic feedback.**

**📖 Applications: language learning and speech therapy, since learners can understand *how to fix* their errors, not just *which phoneme* they got wrong.**

**dataset:**

https://psi.engr.tamu.edu/l2-arctic-corpus/

https://www.openslr.org/12

**Research 3: Evaluating ASR Robustness to Spontaneous Speech Errors (WhisperX + SFUSED)**

**🎯 Purpose**

* **Investigate how WhisperX handles spontaneous speech errors in natural conversation.**
* **Speech errors include:**
  + **Substitution, deletion, or addition of sounds.**
  + **Using the wrong word (e.g., saying *username* instead of *password*).**
  + **Stopping mid-word.**
  + **Repeating words incorrectly.**
* **Previous ASR research often assumed error-free speech, but natural conversation contains 1–2 errors per minute (~1% of speech).**
* **Training/evaluating ASR on error-containing speech can double the Word Error Rate (WER).**

**📂 Dataset: SFUSED English (Simon Fraser University Speech Error Database)**

* **360 hours of natural podcasts (e.g., Astronomy Podcast, Rooster Teeth).**
* **10,000 annotated speech errors.**
* **Each error labeled with:**
  + **Error type: sound-level or word-level.**
  + **Whether the speaker self-corrected.**
  + **Whether the word was complete or incomplete.**
  + **Whether the error was contextual or independent.**
  + **Position of sound error (onset, nucleus, coda).**
* **Includes both the spoken error and the intended word.**

**🤖 Model**

* **WhisperX (built on Whisper-large-v2 by OpenAI).**
* **Modifications:**
  + **Handles long audio recordings.**
  + **Includes Voice Activity Detection + Speaker Diarization.**
  + **Uses wav2vec2.0 for alignment with timestamps.**
  + **Beam search decoding (beam size = 5).**

**🧪 Evaluation Method**

* **Sample: 5,300 annotated errors.**
* **WhisperX outputs compared against annotations.**
* **Categories of ASR output:**
  1. **Corrected → model outputs the intended word.**
  2. **Faithful → model outputs the spoken error.**
  3. **Incorrect → neither correct nor faithful.**
* **Both Corrected + Faithful counted as accurate.**

**📊 Results**

* **Sound-level errors:**
  + **Accuracy = 83%**
  + **(81% Corrected, 2% Faithful, 17% Incorrect)**
* **Word-level errors:**
  + **Accuracy = 74%**
  + **(43% Corrected, 31% Faithful, 26% Incorrect)**
* **Performance by conditions:**
  + **Self-corrections: helpful for sound errors, but harmful for word errors (two competing words).**
  + **Context-related errors: little effect on sound errors, increased mistakes on word errors.**
  + **Incomplete words: improved ASR performance (model relies on context).**
  + **Error position (onset/nucleus/coda): no major effect, but medial errors were more easily corrected.**

**📌 Conclusion**

* **Natural speech contains frequent errors, so ASR systems must be evaluated on this basis.**
* **WhisperX is relatively strong with sound errors but struggles more with word errors.**
* **Incomplete words may actually aid recognition due to context.**

**dataset:** <https://osf.io/8c9rg/>