htx-ds-task6

—PERSONAL NOTES, PROPOSAL STARTS LATER, FEEL FREE TO IGNORE THIS PART, IGNORED FROM WORD COUNT—

Recap: Paper Highlights

- SSL reduce WER of in-domain test set by 7%, compared to ALL other alternative pre-training strategies with out-domain data → In-domain test set highly relevant for enhancing accuracies, (makes sense since we are making it domain-specific)
- Fully utilise SSL for uncurated, wild audio data from YouTube, Xception-based AED boost performance by 2.5% for in-domain test set → Can adopt both open-source data collection plus using Xception-based AED
- Adam unstable for low learning rate, while AdamW can overcome instability and boost performance by >3% → OK can adopt new optimiser
- InfoNCE large biased estimation, newly developed flatNCE can overcome and boost performance by up to 2% → OK can adopt
- Multi-head multilingual SSL can outperform single-head multilingual SSL by 1.6% → OK can adopt multi-head

So What? Actionables and Learning Points

- Focus on gathering in-domain data to curate dataset
- Open-source data collection from “wild audio data”
- New technologies to consider adopting:
    - Multi-head multilingual SSL
    - newly developed flatNCE
    - AdamW optimiser

Understanding the Challenges: Dysarthric speech, known as slurred speech, presents unique challenges for ASR systems due to its distorted acoustic

features, reduced intelligibility, and variability across speakers and severity levels. Self-supervised learning (SSL) offers a promising approach by leveraging large amounts of unlabeled data to learn robust representations that can be fine-tuned for specific tasks.

——PROPOSAL START (WORD COUNT TOO)——

Proposed SSL Pipeline

1. **Data Collection and Preprocessing:**
    - **Data Collection:** Gather a diverse dataset of dysarthric speech recordings, considering various types of dysarthria, speaking styles, and recording environments. Notably, consider:
        - In-domain data → eg Relevant patients in Hospitals, elderly folks home
        - Open-source data → Existing public speech datasets, whether local or global; eg. data.gov.sg, UASpeech, Kaggle; potentially relatively cleaner speech since most speakers do not have dysarthric speech, so further data augmentation can be performed to introduce more noise to closer mimic patients speech patterns
    - **Data Augmentation:** Augment the dataset using techniques to increase data variability and improve model robustness. Notably, considering the AED model filters:
        - Speech filters (ignore moments of nothing)
        - Speech crop (split up audio to only include speech portion)
        - Random crop (split up long speeches)
2. **Self-Supervised Pre-training:**
    - **Model Development:**
        - **Model:** To follow paper and adopt Lfb2vec as candidate model
        - **Methodology:** Multi-head multilingual SSL over single-head variant, newly developed flatNCE as loss function, AdamW as optimiser
        - **Training:** Train model on unlabeled dysarthric speech data.

3. **Fine-tuning for Downstream Tasks:**
    - **Task-Specific Training Data:** Collect a smaller dataset of labeled dysarthric speech for specific tasks, such as speech recognition, speaker identification, or severity assessment.
        - Notably, focus on collecting in-domain data, eg patients from local hospitals or in rehabilitation, to curate test set
    - **Fine-tuning:** Then fine-tune the pre-trained model on the task-specific labeled data.
        - Can either be performed with the above in-domain test datasets, or possibly complementing with relevant open-source speech datasets

Continuous Learning

1. **Model Retraining: G**athering new data and periodic model re-training is essential to maintain model performance over time. Also, some key points:
    - **Monitoring and Evaluation:** Continuously monitor the model's performance on new data and evaluate the impact of retraining. This can be done by setting performance KPIs, eg WER metric
    - **Data Gathering:** Building on the multi-lingual point from the paper, Singapore is a multi-racial country with many language groups plus dialect communities for the older generations (e.g., Hokkien, Cantonese, Teochew)
        - As this may introduce additional complexity to the model’s learning process, depending on the target audience of the final tool, one idea is to expand the dataset and collect more data from speakers of these lesser spoken language groups
2. **Incremental Learning:** As new dysarthric speech data becomes available, incrementally train the model on the new data without forgetting previously learned information, using techniques like:
    - **Regularization:** Penalize changes to the model's parameters during incremental training.
    - **Knowledge Distillation:** Train a smaller, faster model to mimic the behavior of the larger, pre-trained model.

3. **Active Learning:** Select the most informative new data points to label and add to the training set, prioritizing data that is most likely to improve model performance.
    - **Dialect Diversity:** Active learning can help to efficiently incorporate data from speakers of different dialects and language groups; by strategically selecting data points from underrepresented dialects, the model can be trained to better understand and recognize the unique characteristics of each dialect.

(482 words)
