

Performance Fairness for Speech and Speaker Recognition & Graph Label Propagation for Cross-Utterance Rescoring

Andreas Stolcke

December 21, 2023



Overview

Part 1: Group fairness for speech recognition and speaker recognition

Quick tour of techniques and applications to ASR and SpkrRec

Deep Dive 1: Geographic fairness for ASR

Deep Dive 2: Group-adapted fusion for SpkrRec

Part 2: Graph label propagation for SpkrRec and ASR

Introduction to graph-LP, with applications to SpkrRec

Deep Dive 3: cross-utterance ASR rescoring



Part 1

Group fairness for speech recognition and speaker recognition

What is Group Fairness?

- Bing Chat (GPT-4) says:
 - “Group fairness is a concept in machine learning that measures how a group of individuals with certain **protected attributes** (like gender or race) is impacted differently from other groups. It aims to achieve the same outcomes across different demographics or a set of protected population classes”
- But what about “non-protected” groups/attributes?
 - For example: age, regional accent, tenure with a voice assistant
 - Goal is to make speech-enabled AI systems perform about equally well for all speakers/attributes

$$P(f(x) \geq \theta \mid A(x)) \approx P(f(x) \geq \theta \mid \neg A(x)),$$

for a performance metric $f(x)$, threshold θ , and all attributes $A(x)$ we care about.

In this talk

- Focus on algorithmic approaches that reduce disparities in performance
 - For speech recognition (ASR)
 - For speaker recognition (specifically, speaker verification - SV)
- Metrics will be depend on task
 - Word error rate (WER) for ASR
 - Equal error rate (EER) for speaker verification
 - Absolute or relative differences in metric between groups

Fairness and representation

- Empirically, group underperformance in ML systems is typically associated with underrepresentation in the training set
 - Training objective is to minimize loss over the entire dataset
 - It “pays” more to minimize loss for the majority
- Example:

If nonnative speakers are a minority in the data, we expect ASR models to perform poorly for them
- Remedy:

Increase the underrepresented group’s aggregate contribution to the loss function

How to define/identify groups?

- Several approaches:
 - By pre-existing categories, e.g., demographic labels, metadata, ...
 - Proxy labels (e.g., ZIP codes for demographics)
 - By automatic discovery / clustering
 - By an adversarial model (implicitly)
- Getting labels is a challenge in itself
 - Especially for protected / demographic attributes
 - Use proxy attributes (e.g., ZIP code associated with demographics)
- Methods that require no group labels in training are preferred, other things being equal
 - We then only need labels on the test data for evaluation

Mitigation: Improving representation

- “Target group” = group that is underrepresented / found to have sub-par performance
- How to increase representation of the target group in training loss?
 - Oversample the target group
 - Give extra weight to target group samples in the loss computation
 - Adapt / fine-tune model on the target group
 - Use group-specific models (and combine them)
 - Use modified loss function that penalizes disparities (adversarial weighting)
 - Fabricate data for the target group

Example 1: ASR group performance fairness

- Target group: cohorts of similar speakers with empirically low ASR accuracy
- Identify by:
 - ZIP codes as a proxy for minority demographics
 - Unsupervised speaker embedding clustering
- Mitigate by:
 - Oversampling the target group with semi-supervised training data
 - Adding cohort embedding as input to ASR model
- Results:
 - Reduce WER gap between top and bottom cohort from 56% to 39%
 - Three-way human labelers disagree 33% more on hardest 10% of ASR data
- More info at [Pranav Dheram, Murugesan Ramakrishnan, A. Raju, I-F. Chen, B. King, K. Powell, M. Saboowala, K. Shetty, A. Stolcke, Toward Fairness in Speech Recognition: Discovery and mitigation of performance disparities, *Proc. Interspeech*, 2022](#)

Example 2: Speaker verification fairness with adversarial reweighting

- Target group: any cohort characterizable by input features
- Identify by:
 - Implicitly; adversarial component of the loss function penalizes any disparities in performance that can be predicted from the input features
- Mitigate by:
 - Adversary *maximizes* loss by assigning higher weight to harder-to-classify inputs
 - Regular training loss *minimized* subject to adversarial weights, and iterate
- Results:
 - Formulated adversarial reweighting for two-input metric learning problems
 - Gap b/w speaker genders reduced from 0.70 to 0.58 abs.; overall EER down 8%
- More info at [Minho Jin, C. J.-T. Ju, Z. Chen, Y.-C. Liu, J. Droppo, A. Stolcke, *Adversarial Reweighting for Speaker Verification Fairness*, *Proc. Interspeech*, 2022](#)

Example 3: ASR robustness for stuttered speech

- Target group: speakers with stutter (severe lack of training data)
- Identify by:
 - Human labeling on seed data
- Mitigate by:
 - Generate synthetic stuttered data using a modified TTS system; fine-tune ASR
- Results:
 - Effective stutter synthesis controlled by tags embedded in input text
 - WER for stuttered speech reduced by 6% relative

- More info at

[Xin Zhang, I. Vallés-Pérez, A. Stolcke, C. Yu, J. Droppo, O. Shonibare, R. Barra-Chicote, V. Ravichandran, Stutter-TTS: Controlled Synthesis and Improved Recognition of Stuttered Speech, *Proc. NeurIPS Workshop on Synthetic Data for Machine Learning*, 2022](#)

Deep Dive 1: Geographic fairness for ASR

- Target group: speakers in geographic regions with poor ASR performance
- Identify by:
 - Geo-coordinates, with automatic induction of regions
- Mitigate by:
 - Oversampling training data
 - Elastic weight consolidation (EWC) in training loss
- Results:
 - Reduces WER in highest-error regions by 3.2% relative
 - Overall WER reduced by 1.3% relative
- More info at

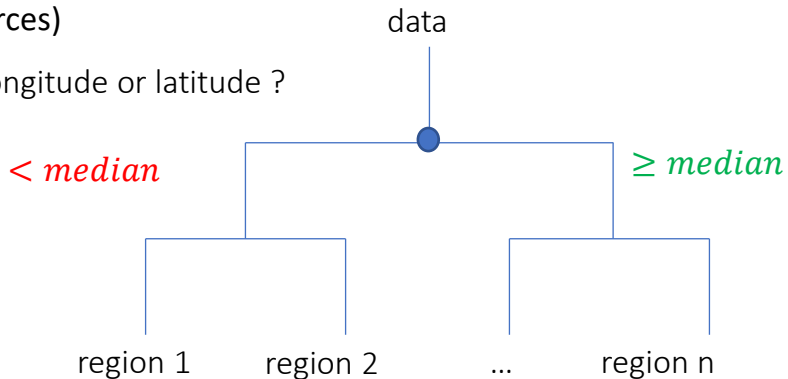
[Viet Anh Trinh, P. Ghahremani, B. King, J. Droppo, A. Stolcke, R. Maas, *Reducing Geographic Disparities in Automatic Speech Recognition via Elastic Weight Consolidation*, *Proc. Interspeech*, 2022](#)

Motivation: Geographical Fairness for ASR

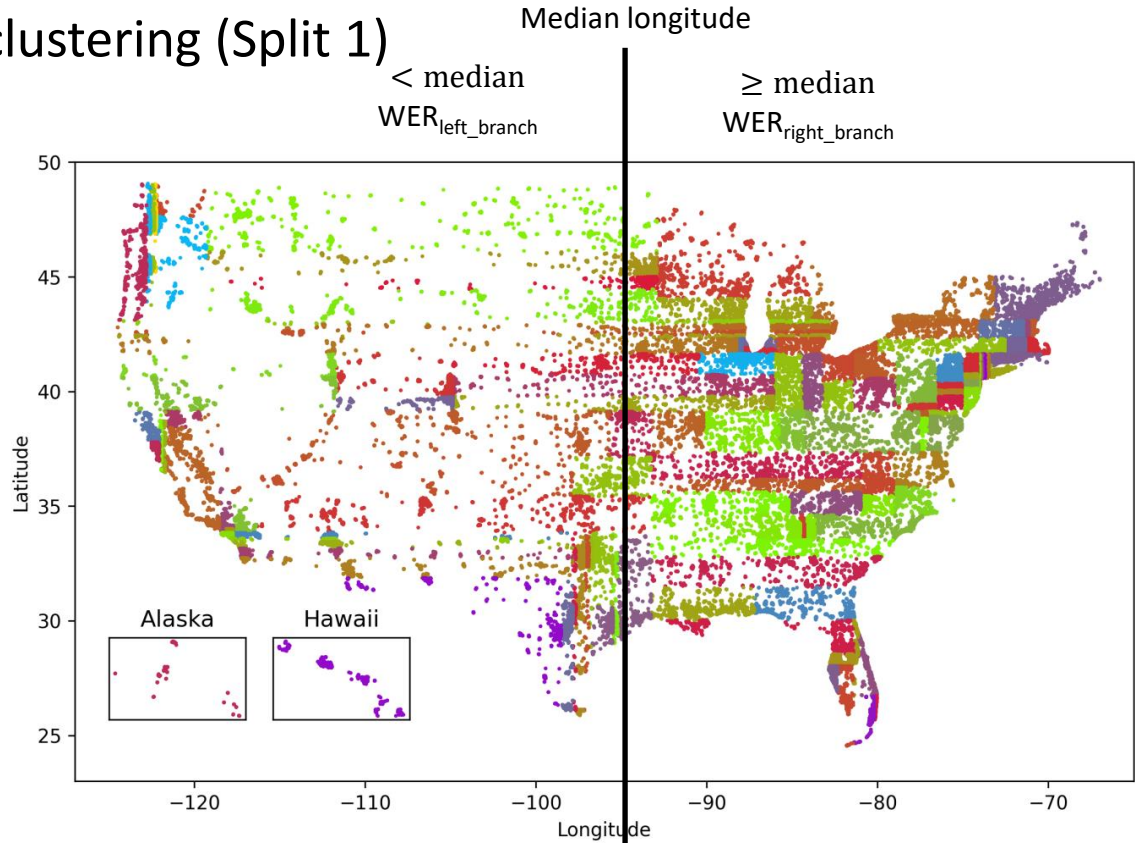
- ASR performance is affected by geography (e.g., regional accent, socio-economic differences)
- Geographic fairness is easy to explain and motivate
- Instead of ZIP codes and human population attributes, use geolocation directly for grouping
- Given a pretrained ASR model, cluster speakers by geolocation to identify areas of high error rate
- Mitigation:
 - Adapt ASR model to reduce the performance gap against these high error regions
 - Without degrading average performance for all regions
 - Without access to the data of the pretraining stage

Geographical Clustering by ASR Accuracy

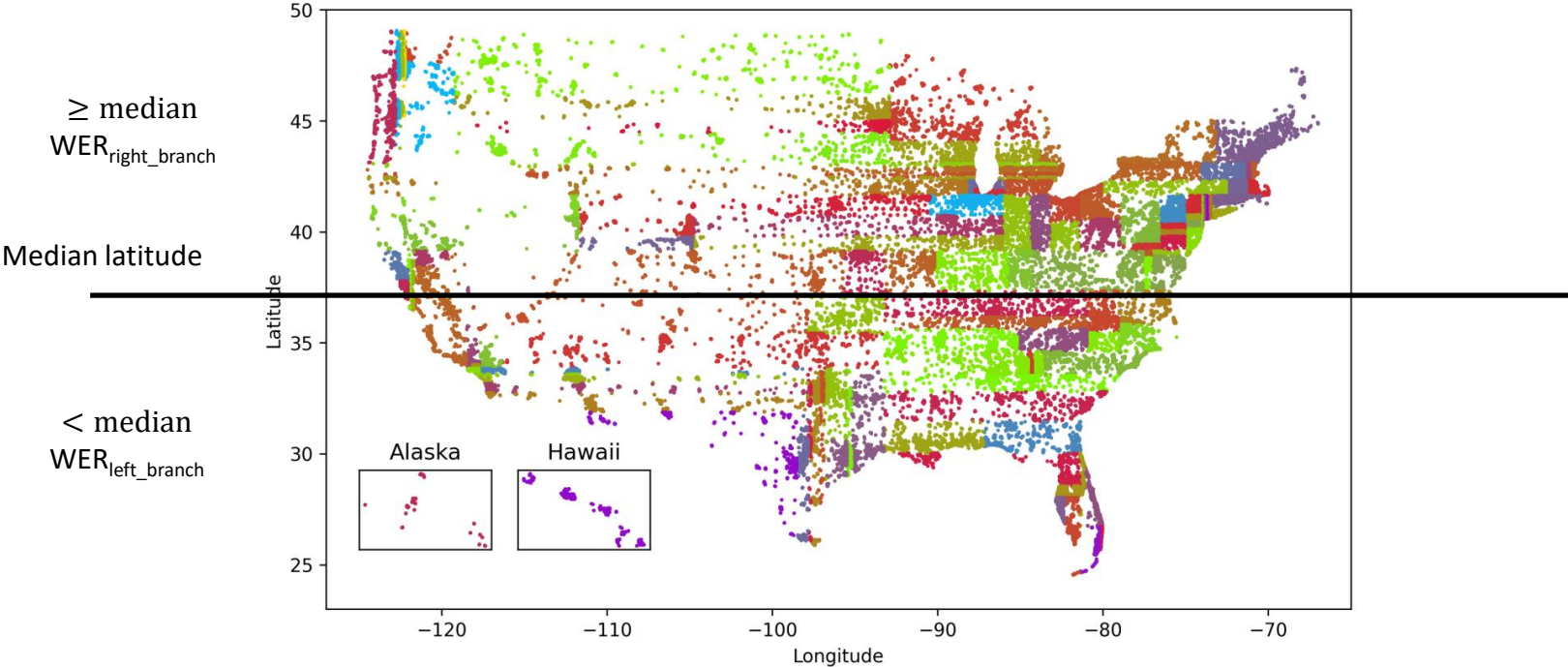
- Use clustering tree to split US data into regions while maximizing word error rate (WER) differences between regions
- $WER_{diff} = (WER_{left-branch} - WER_{right-branch})^2$
- Split the data by longitude if $WER_{diff-by-longitude} > WER_{diff-by-latitude}$, otherwise by latitude
- Repeat while the number of data sources (speakers, devices) in each leaf \geq threshold x (to ensure each region has at least x sources)



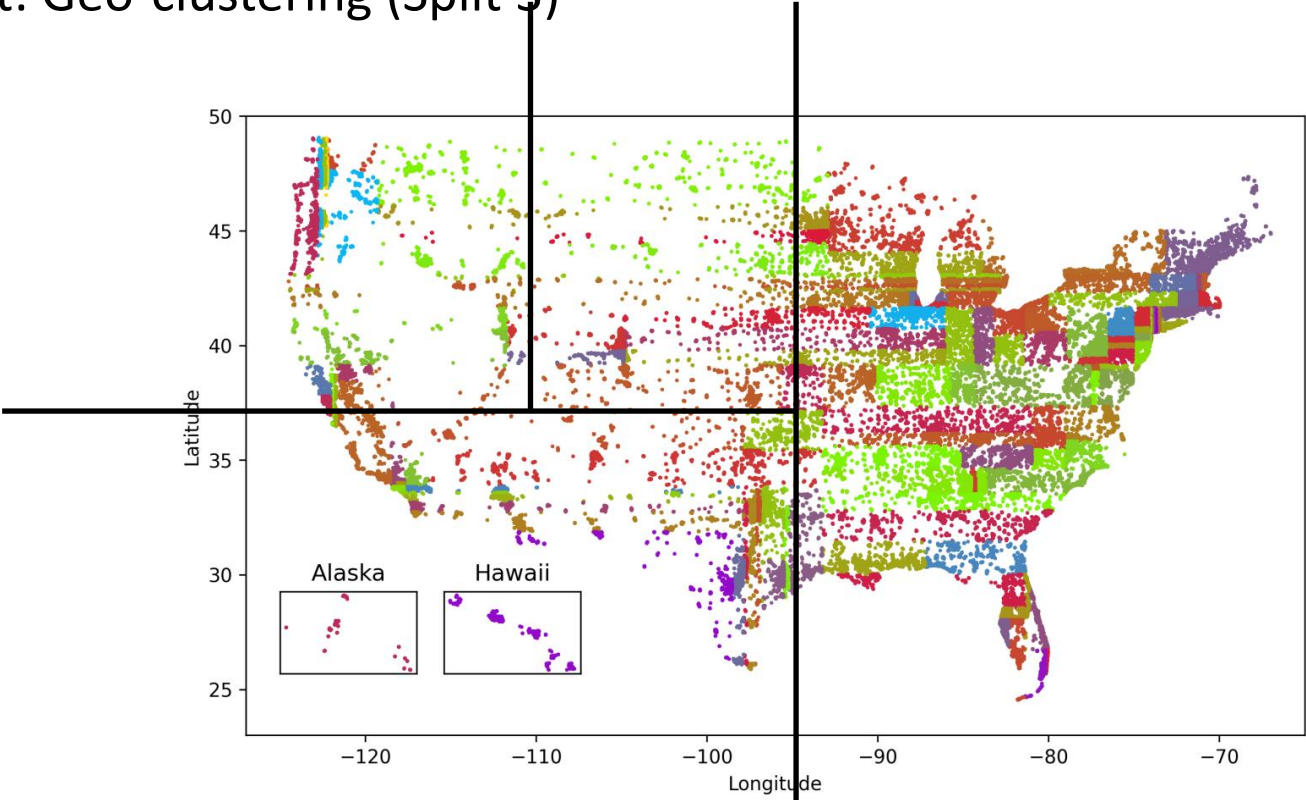
Result: Geo-clustering (Split 1)



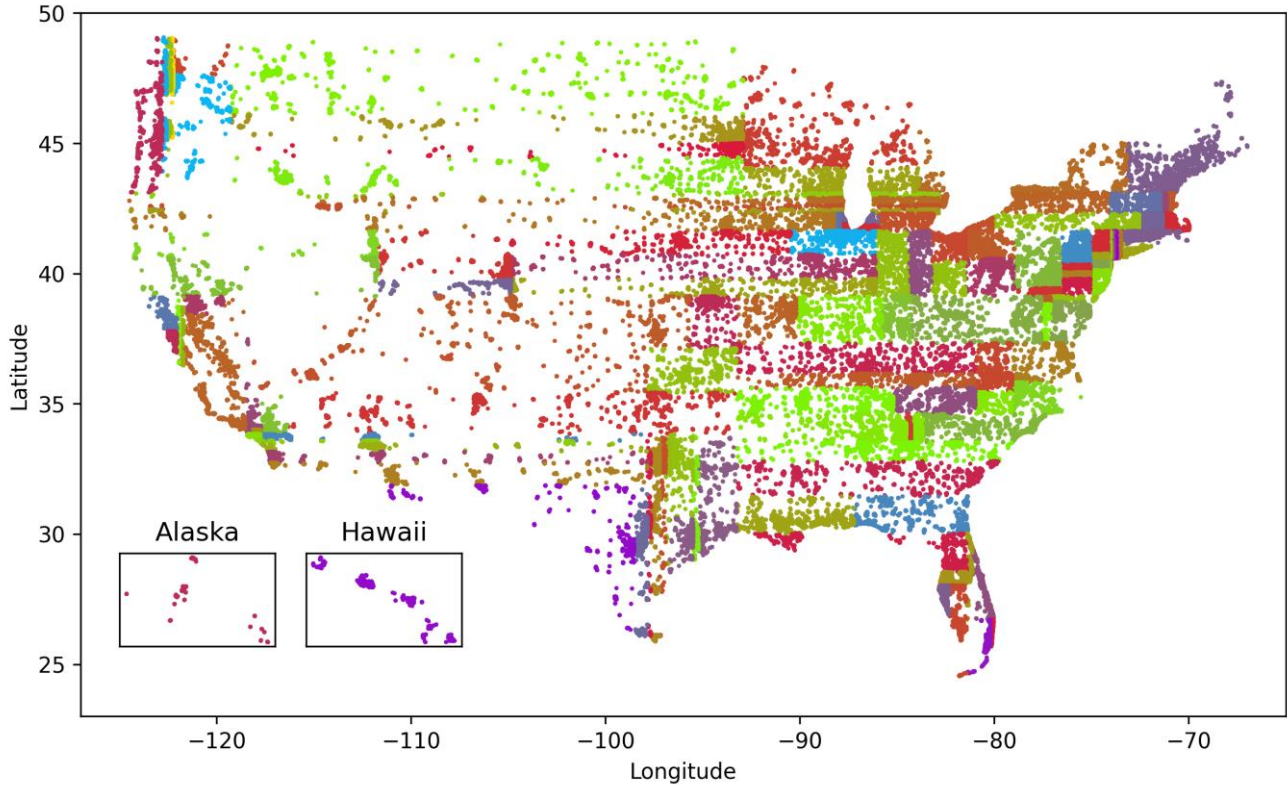
Result: Geo-clustering (Split 2)



Result: Geo-clustering (Split 3)



Final Geo-clustered Regions (same color = same WER)



Elastic Weight Consolidation for ASR Adaptation

- End-to-end ASR using RNN-T (5x1024 encoder layers, 2x1025 predictor, 1x1024 joint network)
- To prevent catastrophic forgetting, use EWC [1] loss in addition to standard RNN-T loss:

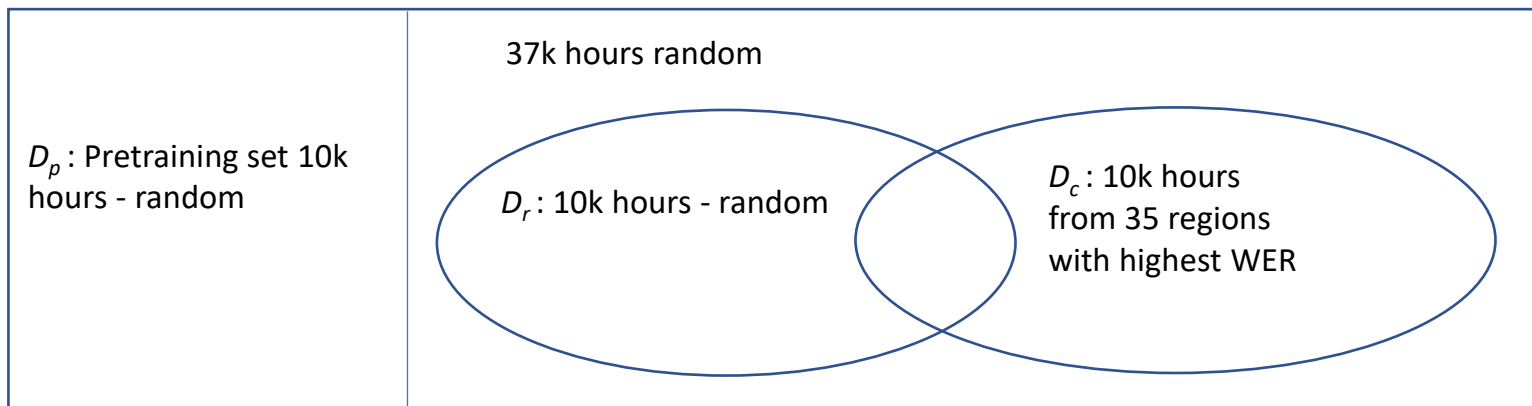
$$\mathcal{L}(\theta) = \mathcal{L}_{\text{ASR}}(\theta) + \frac{\lambda}{2} \sum_i F_i (\theta_i - \theta_{p,i}^*)^2$$

- Intuition behind EWC: force ASR parameters θ to **be close to the best parameters of the pretrained model θ_p^* , along the directions that are important to the pretrained task** (based on Fisher information)
- Alternative to heuristic approaches that freeze portions of the pretrained network (note: this was before adapters, LoRA, etc.)

[1] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, D. Hassabis, C. Clopath, D. Kumaran, and R. Hadsell, “Overcoming catastrophic forgetting in neural networks”, *Proc. National Academy of Sciences*, vol. 114, no. 13, pp. 3521–3526, 2017

Experiment data

- Drawn from de-identified user data from a commercial voice-enabled AI assistant
- Pretrain models on 10k hours, randomly sampled
- Rank geo-clustered data by region-averaged WER (min no. devices per region = 1500 → 126 regions)
- Sample 37k hours, disjoint with pretraining set
- Sample 10k adaptation data in order of high-to-low region-average WER; random 10k as control



Results

Description	Data	Region WERR				Overall WERR
		variance	mean	max	min	
Baseline	D_p	0	0	0	0	0
No freeze	D_c	-5.3	-0.9	-2.9	-4.6	-1.1
Freeze Encoder	D_c	-1.8	0.0	-1.4	-5.4	-0.1
Freeze Predictor	D_c	1.8	-0.3	-1.3	-8.5	-0.4
Freeze 3 lowest encoderlayers and 1 predictor layer	D_c	-0.9	-0.5	-2.5	-2.7	-0.4
Proposed method	D_c	-7.9	-1.1	-3.2	-5.8	-1.3
Empirical bound	$D_p + D_c$	-5.3	-1.2	-2.3	0.2	-1.0
	$D_p + D_r$	-12.3	-2.3	-0.9	-7.3	-2.1

- EWC beats standard fine-tuning, both in overall WER and variance across regions
- Better than heuristic freezing of network portions
- Better than training on combined pretraining and adaptation data
- Training on all-random data best overall; not much help to highest-WER regions

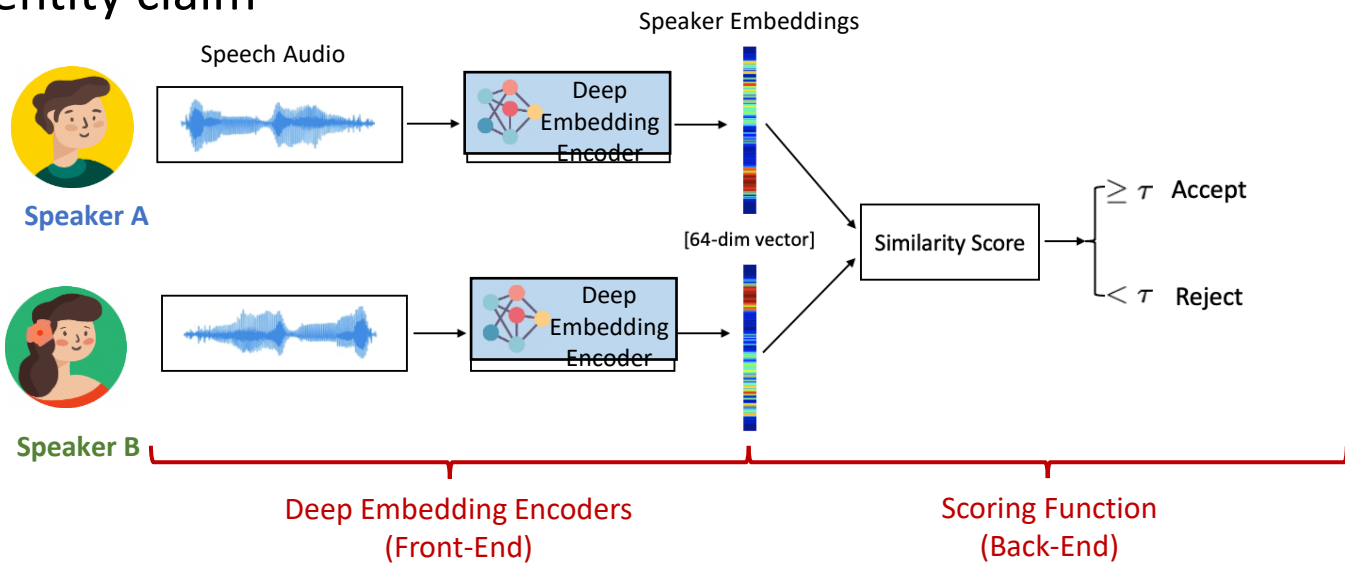
Deep Dive 2: Group-adapted fusion for Speaker Verification

- Target group: any group or attribute with imbalanced representation
 - Study example: artificially manipulated gender imbalance
- Identify by:
 - Metadata or automatic labeling (e.g., clustering)
- Mitigate by:
 - Training group-specific submodels
 - Merge submodel predictions in equal proportion
- Results:
 - Minority group EER reduced up to 18.6% relative (more when more imbalance)
 - Overall EER reduced up to 29.9% relative (more when more imbalance)
- More info at

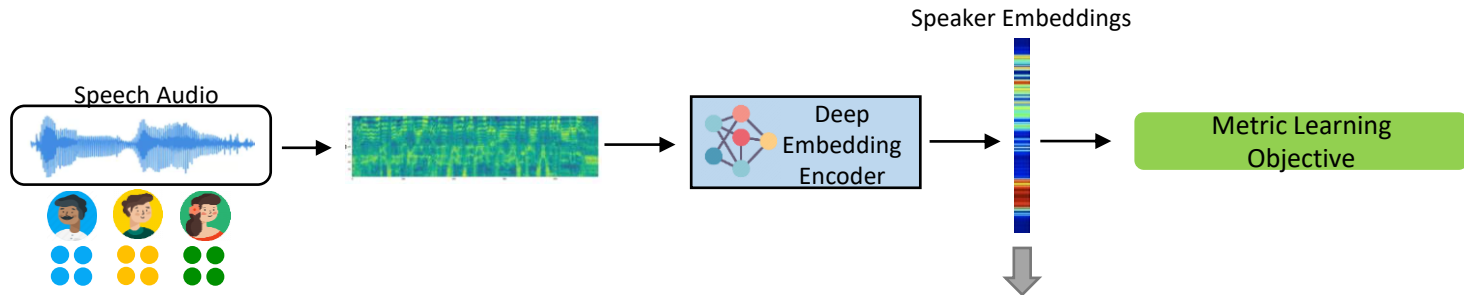
[Hua Shen, Y. Yang, G. Sun, R. Langman, E. Han, J. Droppo, A. Stolcke, Improving Fairness in Speaker Verification via Group-Adapted Fusion Network, *Proc. ICASSP, 2022*](#)

Deep Learning Speaker Verification

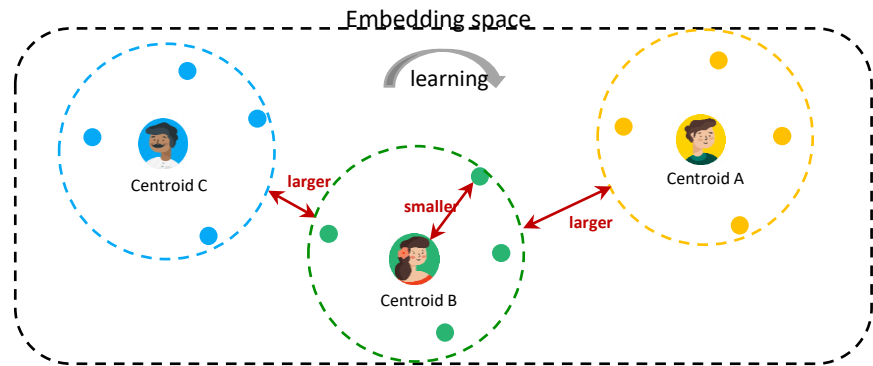
- An encoder network maps audio stream to fixed-size embeddings
- A scoring backend compares embeddings to accept/reject speaker identity claim



Speaker Verification - Training



- Encoder network is trained either
 - for classification (1 of N training speakers)
 - via metric learning (triplet loss, GE2E, angular prototypical loss, ...)
- Metric learning is easier to scale

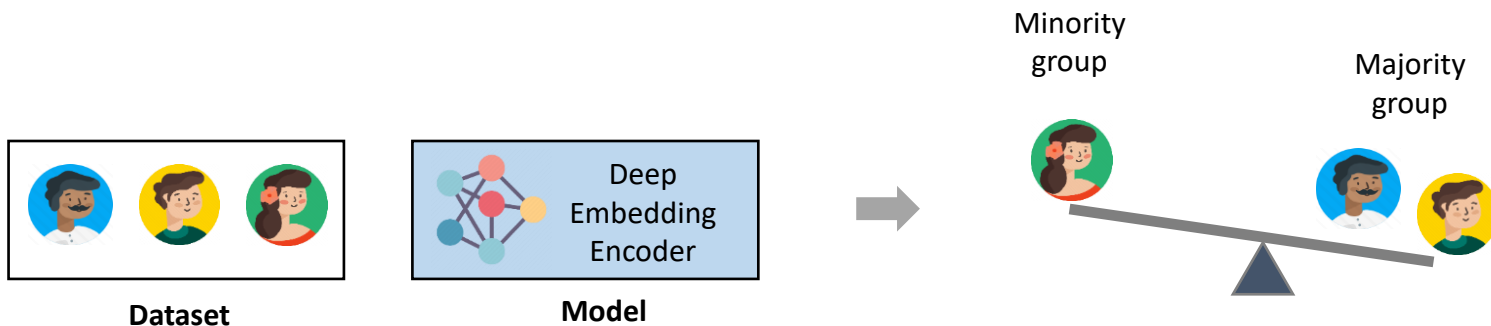


Learn to optimize the embedding to get:

- **smaller** distance between **same** speakers
- **larger** distance with **different** speakers.

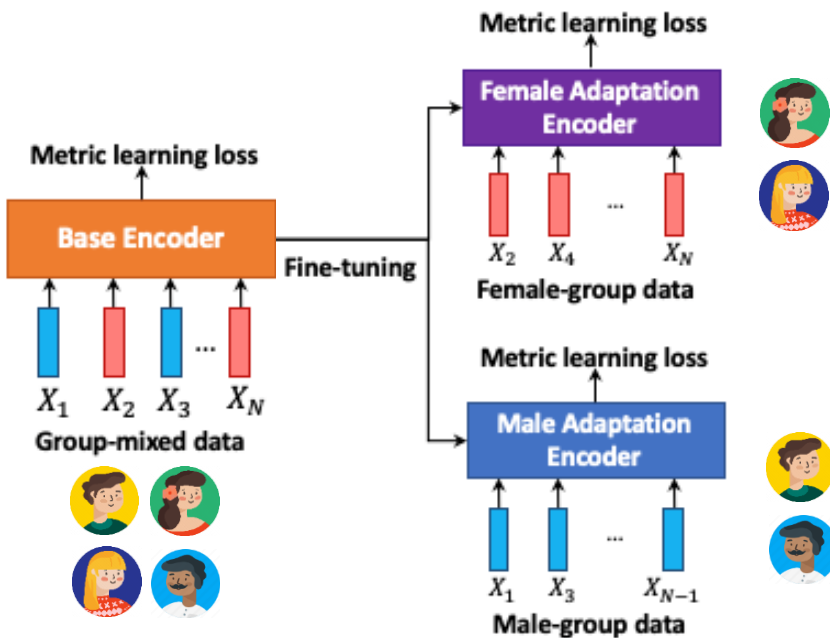
Training Data Imbalance and SV Performance

- First systematic study of effect of speaker attribute imbalance on SV
- Manipulate the training set to control the ratio of males to females



- Measure effect on
 - group-wise EER
 - overall EER
 - DiparityScore (DS) = $|EER[F] - EER[M]|$

Imbalance Mitigation: Group-adapted Fusion Networks (GFN)



Group Embedding Adaptation

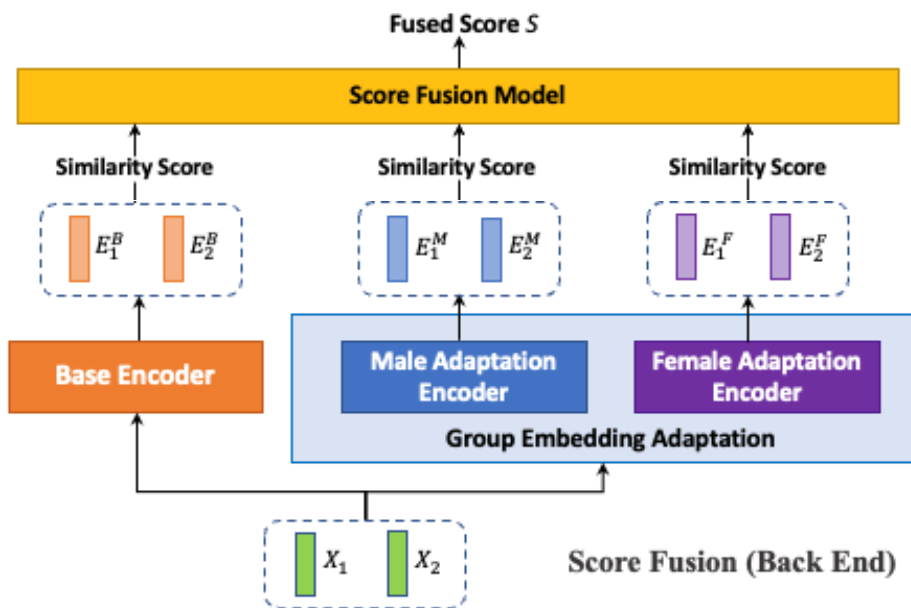
$$\mathbf{E}_i^B = \text{BaseEncoder}(\mathbf{X}_i), i = 1, 2$$

$$\mathbf{E}_i^F = \text{FemaleAdaptationEncoder}(\mathbf{X}_i), i = 1, 2$$

$$\mathbf{E}_i^M = \text{MaleAdaptationEncoder}(\mathbf{X}_i), i = 1, 2$$

1. Train all-data base models
2. Adapt model to each group
3. Generate multiple embeddings for each utterance

Score fusion Back-end



Score fusion model

$$S^B = \text{CosineSimilarity}(E_1^B, E_2^B),$$

$$S^F = \text{CosineSimilarity}(E_1^F, E_2^F),$$

$$S^M = \text{CosineSimilarity}(E_1^M, E_2^M)$$

$$S = \text{Sigmoid}(f([S^B, S^F, S^M]; W)). \quad \leftarrow \text{Neural Network}$$

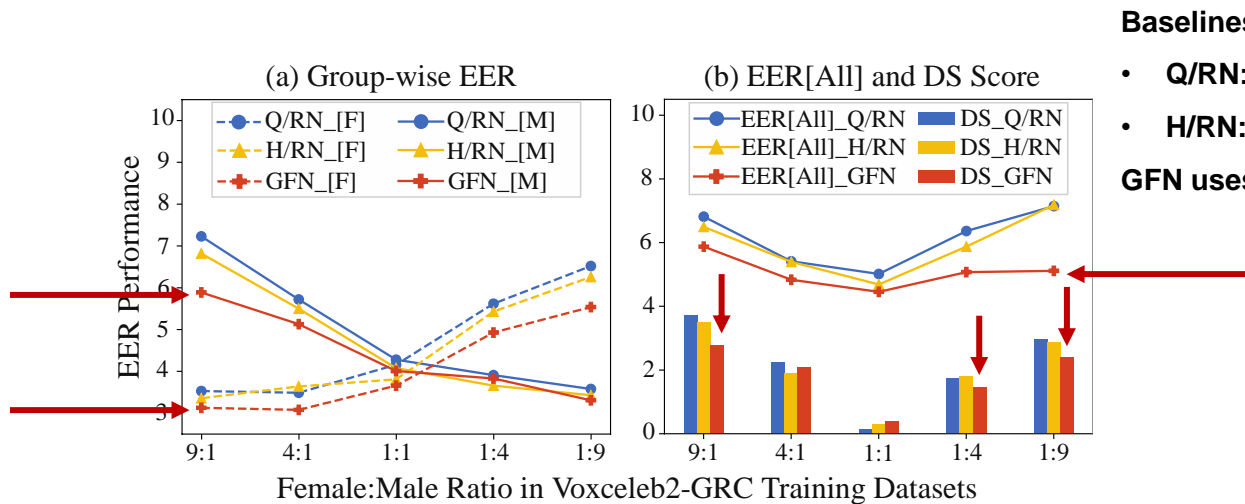
The back-end score fusion model combines all scores for speaker verification.

Training objective

Binary cross-entropy loss with positive and negative training pairs

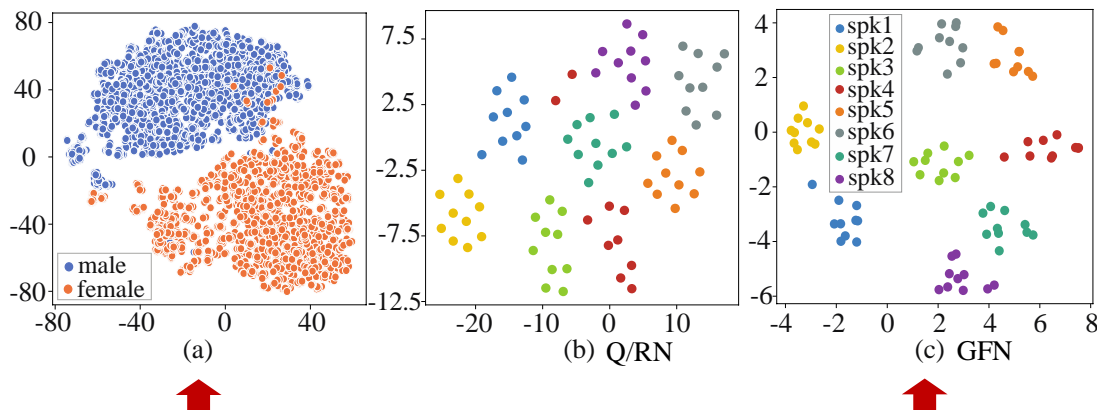
$$L = -\frac{1}{M} \left(\sum_{n \in \mathcal{P}} y_n \log S_n + \sum_{n \in \mathcal{N}} (1 - y_n) \log(1 - S_n) \right)$$

Imbalance mitigation with GFN: Results



- GFN achieves better group-wise and overall EERs than baselines, for both genders and all imbalance ratios
- GFN reduces the disparity score for most (9:1, 1:4, 1:9) gender ratios

Embedding visualization



t-SNE projections

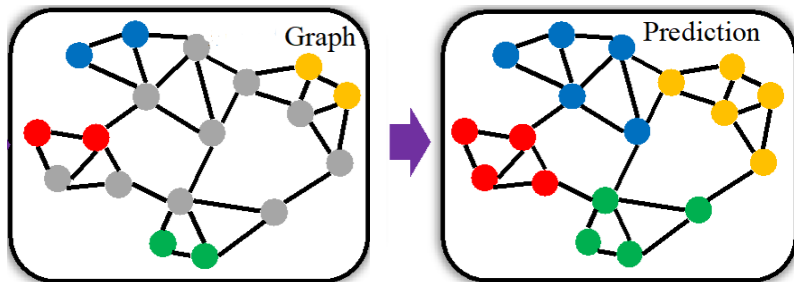
- Genders tend to aggregate in different regions of the embedding space
- GFN encoder generates higher quality embeddings relative to Q/RN baseline (more compact for the same speakers; more separate for different speakers)

Part 2

Graph label propagation for speaker recognition and ASR

Graph Label Propagation for unsupervised learning

- Represent labeled and unlabeled instances as graph nodes
- Encode sample similarity as edge weights
- Propagate labels so as to
 - Stay close to the original labels (supervision)
 - Minimize discrepancies between similar instances



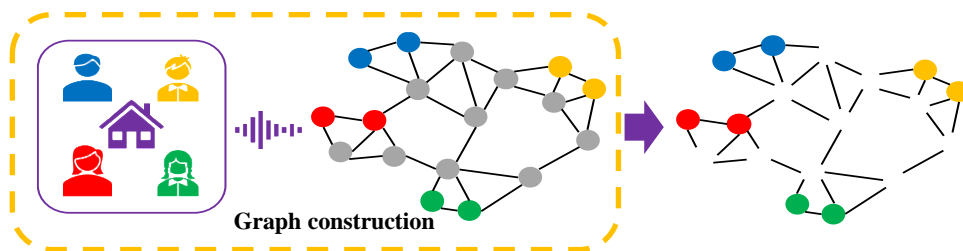
Graph-LP for Household Speaker ID

- Task: ID speakers in household settings
 - Small number of labeled samples (enrollment utterances)
 - Large number of unlabeled samples (runtime utterances)
- How to leverage unlabeled samples for classifying subsequent utterances?
- Apply graph-LP:
 - Utterances are graph nodes
 - Edge weights given by distance in speaker embedding space
 - Enrollment utterances anchor the label propagation
- More info at

Long Chen, V. Ravichandran, A. Stolcke, [Graph-based Label Propagation for Semi-Supervised Speaker Identification, Proc. Interspeech, 2021](#)

Long Chen, Yixiong Meng, V. Ravichandran, A. Stolcke, [Graph-based Multi-View Fusion and Local Adaptation: Mitigating Within-Household Confusability for Speaker Identification, Proc. Interspeech, 2022](#)

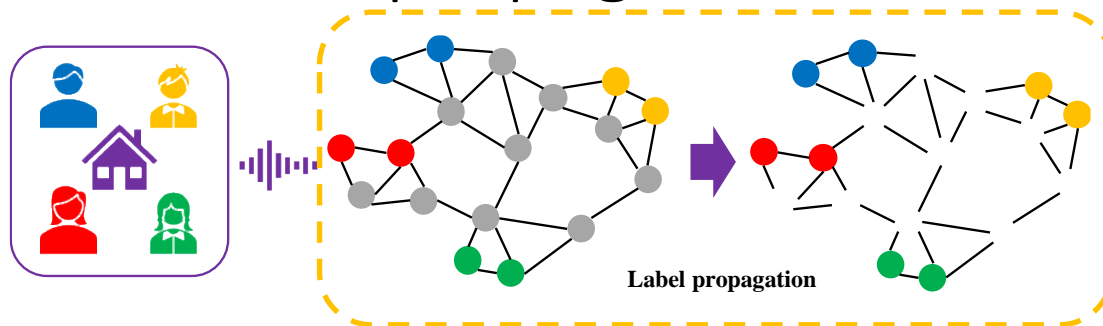
Household SID: utterance graph construction



- Nodes = utterances (l labeled, u unlabeled)
- Label set = speaker IDs (1 ... C)
- Edge weights = pairwise utterance embedding similarity scores
- W matrix of edge weights, σ is a temperature-like hyperparameter

$$W_{ij} = \exp\left(-\frac{\sum_{d=1}^D (x_i^d - x_j^d)^2}{\sigma^2}\right)$$

Household SID: label propagation



Label propagation

- Objective function:
 - a) supervised loss over the labeled instances
 - b) a graph-based regularization term to ensure labeling is smooth over the graph

$$\operatorname{argmin}_{\mathbf{f}} \|\mathbf{f} - \mathbf{Y}\|_2^2 + \lambda \mathbf{f}^T L_{\text{sym}} \mathbf{f}$$

$$\text{where } L_{\text{sym}} = I - D^{-1/2} W D^{-1/2}, D_{ii} = \sum_{j=1}^{l+u} W_{ij}.$$

- Solution:
 - iterative algorithm* to spread samples' label information through the graph until achieving global convergence.
 - class normalization** in order to minimize the influence of imbalance in the labels/pseudo-labels.

Algorithm 1: Label Propagation with Normalization

Compute the affinity matrix W as eq. 1 if $i \neq j$ & $W_{ii} = 0$

Compute matrix $S = D^{-1/2} W D^{-1/2}$

Initialize $\hat{Y}^{(0)}$: $\hat{Y}_{ij}^{(0)} = \begin{cases} 1 & (i \leq l, x_i \text{ is labeled as } j) \\ 0 & (\text{else}) \end{cases}$

Normalize $\hat{Y}^{(0)}$: $\hat{Y}_{ij}^{(0)} = \hat{Y}_{ij}^{(0)} / \sum_k \hat{Y}_{kj}^{(0)}$

Choose a parameter $\alpha \in (0,1)$

Iterate $\hat{Y}^{(t+1)} = \alpha S \hat{Y}^{(t)} + (1 - \alpha) \hat{Y}^{(0)}$ until convergence

Label each point x_i by $y_i = \operatorname{argmax}_{j \leq C} \hat{Y}_{ij}^{(\infty)}$

*D. Zhou, et al., "Learning with local and global consistency," in *Proceedings of NIPS*, Dec. 2003, pp. 321–328.

**F. Liu et al., "Normalized label propagation for imbalanced scenario classification," in *Foundations of Intelligent Systems*, vol. 277, Springer Verlag, 2014, pp. 901–909.

Graph-LP for Household Speaker ID: Experiments

- SID trained on VoxCeleb2:
 - GE2E: Generalized end-to-end loss
 - GE2E-Att: Generalized end-to-end loss with attention
- Simulated 4-speaker households drawn from VoxCeleb1
- Metric: Speaker Identification Error Rate
 - $\text{SIER} = 1 - (\text{accuracy of top-scoring predicted speaker})$
- Baselines: cosine scoring (CS) against all labeled samples, embedding averaging (CSEA), pseudo-labeling of unlabeled data (2-CS, 2-CSEA)
- Proposed: label propagation on all utterances (LP), pseudo-labeling with LP (2-LP), pseudo-labeling with embedding averaging (2-LPEA)

Graph-LP for Household Speaker ID: Results

Method	GE2E		GE2E-Att	
	$U=40$	$U=All$	$U=40$	$U=All$
CS	3.36	3.36	2.28	2.28
CSEA	3.06	3.06	2.08	2.08
2-CS	2.05	1.69	1.18	1.01
2-CSEA	1.93	1.39	1.15	0.87
LP	1.82	1.38	1.00	0.77
2-LP	1.73	1.25	0.94	0.69
2-LPEA	1.49	1.31	0.88	0.84

- LP-based methods outperform all baseline methods
- SIER reduced by 10%-23% lower
- 2-LPEA is production-friendly: graph-LP happens offline and generates a single speaker embedding, with standard scoring at runtime

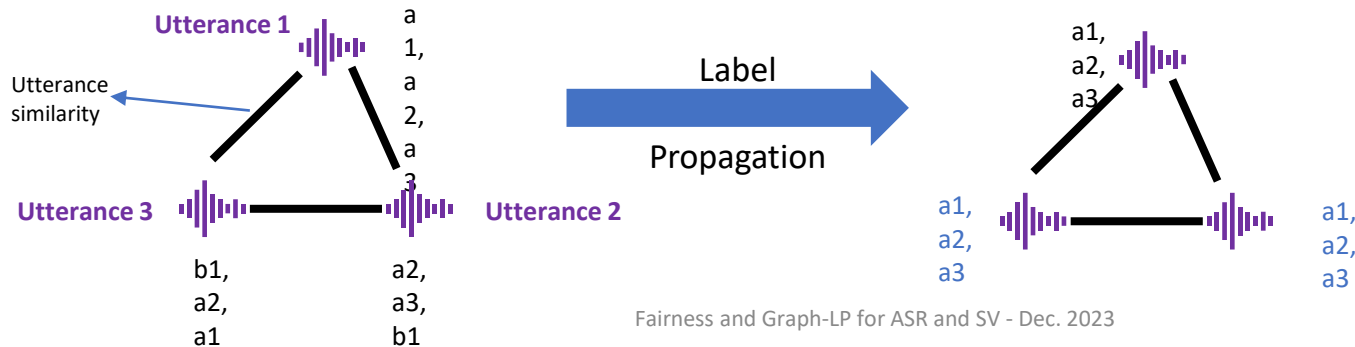
Deep Dive 3: Graph-LP for ASR rescoring

- Task: Offline ASR hypothesis rescoring
- How to leverage cross-utterance similarity?
 - **Similar-sounding utterances should have similar hypotheses**
- How to deal with infinite label set (all possible transcriptions)?
 - Use hypothesis clustering to group utterances for graph construction
 - Use the union of N-best hypotheses as labels
- Apply graph-LP:
 - Utterances are graph nodes
 - Edge weights given by *acoustic similarity scores*
 - All utterances have initial labels given by their 1-best ASR output
- More info at

[Srinath Tankasala, Long Chen, A. Stolcke, A. Raju, Q. Deng, C. Chandak, A. Khare, R. Maas, V. Ravichandran, Cross-Utterance ASR Rescoring with Graph-based Label Propagation, *Proc. ICASSP*, 2023](#)

Graph-LP for ASR rescoring: Motivation

- **ASR rescoring across utterances**
 - Most rescoring approaches (ex. LM rescoring) consider utterances one at a time, independently
 - Doesn't account for utterance-utterance similarity to make predictions
 - *Similar-sounding utterances should predict similar hypotheses*
- **Impact on ASR for underrepresented groups**
 - ASR accuracy degrades for groups not well-represented in the training data (e.g., regional accents)
 - Cross-utterance ASR can exploit local information that stays consistent across utterances:
 - acoustic conditions (noise environment, household, etc.)
 - Speaker and accent idiosyncrasies
- **New idea: map ASR rescoring problem to graph label propagation (*Graph LP*)**
 - Utterances → nodes of the graph
 - Acoustic similarity → edge weights between the nodes
 - The N-best hypotheses from ASR first pass → the node labels
 - Graph-LP algorithm harmonizes the node labels with the acoustic similarity between utterances

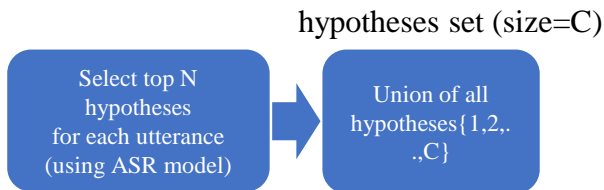


a1, a2, ... are *full* text transcriptions/hypotheses

Graph-LP for ASR rescoring: Details

➤ Node representation and graph edge function

- Utterances are represented using frame embeddings from an RNN-T audio encoder
- Compute utterance-utterance similarity only using audio (without performing full ASR) for graph edge weights
- Label propagation shares information between utterances, improving overall prediction accuracy
- Utterances with strong priors from local ASR will propagate better hypotheses to similar sounding utterances



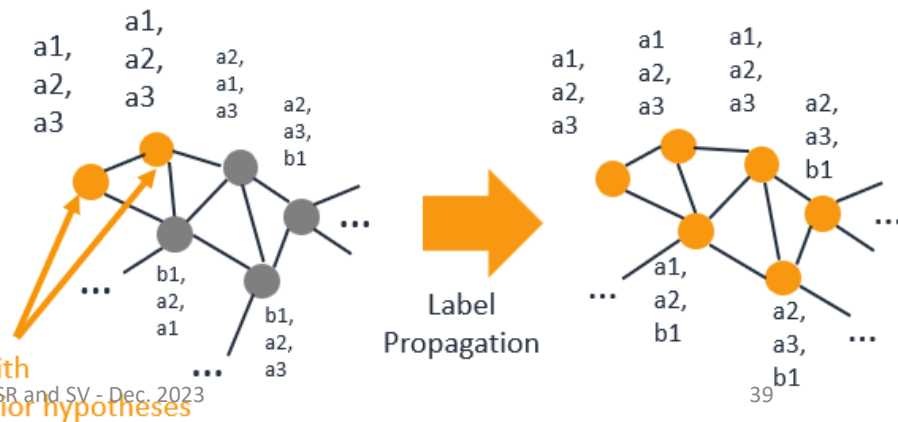
For N=3, let the grouped hypothesis set be {a1, a2, a3, b1, b2, c1, c2, c3, ...}
 Where a1, a2, ... are *full* text transcriptions

Label propagation

- Objective function:
 - a) supervised loss over the labeled nodes
 - b) a graph-based regularization term to ensure labeling is smooth over the graph

$$\operatorname{argmin}_{\mathbf{f}} (\|\mathbf{f} - \mathbf{Y}\|_2^2 + \lambda \operatorname{trace}(\mathbf{f}^T L_{sym} \mathbf{f}))$$

- Solution:
 - Iterative Laplacian algorithm to spread samples' label information through the graph until achieving global convergence



Graph and label set generation

- Perform first pass ASR on all utterances to generate initial hypotheses set for all test set utterances
- Cluster utterances in the hypotheses space based on their *tf-idf* distances
- Given a cluster of M utterances, we generate a graph with M nodes.
 - **Hypothesis sharing:** The label set for that graph is the union of all initial N-best hypotheses of the nodes
- Initial confidence/score of labels (hypotheses) for each node is calculated using normalized log-likelihoods of the hypotheses from the ASR model output

Table 3: Clustered utterances in test data derived from VCTK Corpus* (using *tf-idf* distance)

	#	#
Cluster size (n)	Clusters	Utterances
$n \leq 5$	1782	7112
$5 < n \leq 10$	1352	10008
$10 < n \leq 50$	837	14191
$n > 50$	37	4722
All clustered	4008	36033
All utterances	-	58098

Label propagation was performed independently for each of the 4008 graphs/clusters

Unclustered utterances are not affected by label propagation

* Veaux et al., "CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92)," 2019. [Online]

Experiments – Data and Metric

- Baseline Datasets
 - **Training:** Librispeech corpus*
 - **Test:** VCTK Common Corpus**
 - Has diverse regional accents with lots of overlapping utterances between them
 - Large variance in pre-trained RNN-T model performance across accents
 - Performance degradation for regional accents not well-representated in the training set
 - Baseline results on VCTK is worse for non-American accents (WER > 10%) as the ASR model is trained on Librispeech, biased towards North American accents (US, Canada).
- Evaluation metrics (micro-averages):
 - Word error rate (WER),
 - Sentence error rate (SER)

* Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). Librispeech: An asr corpus based on public domain audio books. In 2015 IEEE international conference on acoustics, speech and signal processing (pp. 5206–5210). IEEE.

** Veaux et al., “CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92),” 2019. [Online]

Utterance distance from acoustic embeddings

- **Utterance embedding:** Last frame and all frame embeddings of audio encoder outputs of RNN-T model
- **Distance function:** d-DTW* distance function across all frame embeddings

Fig 1: TSNE plot of sample utterances based on normalized Euclidean distance between last frames

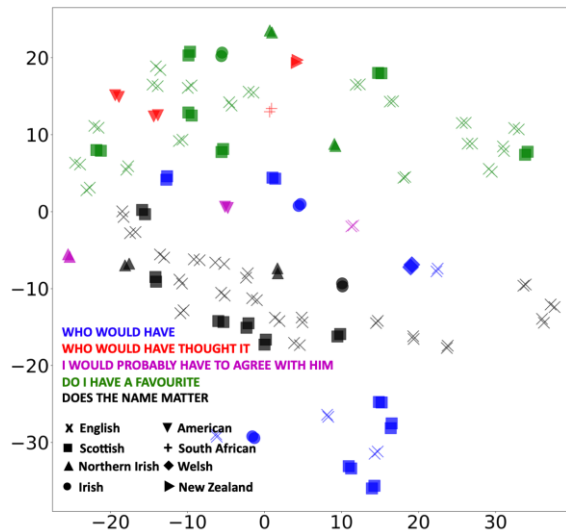
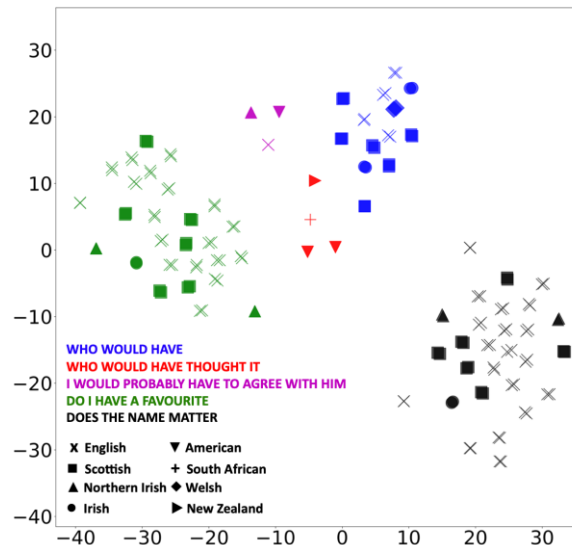


Fig 2: TSNE plot of sample utterances based on normalized d-DTW distance across all frames



* Shokoohi-Yekta et al., "Generalizing DTW to the multi-dimensional case requires an adaptive approach," Data mining and knowledge discovery, vol. 31, no. 1, pp. 1-31, 2017.

Results: Importance of hypothesis sharing

Two LP-based methods:

- **LP without label sharing:** re-ranking only initial hypothesis set for a given utterance/node
- **LP with label sharing:** all graph nodes use the **union** of initial hypotheses of all utterances/nodes

Table 4: Effect of label (hypothesis) sharing on graph-LP results.

Utterance set	Baseline		Without sharing		With sharing	
	WER	SER	WER	SER	WER	SER
All clustered	9.99	42.33	8.75	35.36	5.64	25.19
All utterances	13.97	50.31	13.17	45.98	11.14	39.67

- Label sharing **reduces WER by 35.5%, SER by 28.8%**
- Overall gains relative to baseline: WER 43.5%, SER 40.5% for clustered utterances
- Not shown here: Clusters of larger size seem to show a bigger performance gain

Results: Graph-LP on VCTK utterances

Table 5: Baseline and graph-LP results by regional accents. WER and SER are in %.

Accent	# Utterances	Baseline		Graph-LP	
		WER	SER	WER	SER
American	5807	6.67	30.15	5.02	23.23
Canadian	2151	6.77	30.96	5.01	22.55
English	12960	10.84	45.58	5.82	25.83
Scottish	7174	11.78	48.69	5.76	26.33
Irish	2899	10.46	45.50	5.68	26.04
Northern Irish	1657	9.94	40.13	5.82	22.75
South African	1164	7.82	34.62	4.68	21.82
Indian	937	13.26	48.88	7.94	31.06
Others	1284	10.43	46.11	5.99	25.62
Overall	36033	9.99	42.33	5.64	25.19

- WER improvement across all accents and ranges of cluster sizes
- Discrepancy in WER, SER reduces across different accents, more similar performance on underrepresented groups
- **Improvement is largest for the most difficult accents (improved ASR fairness)**

Graph-LP for ASR rescoring: Conclusions

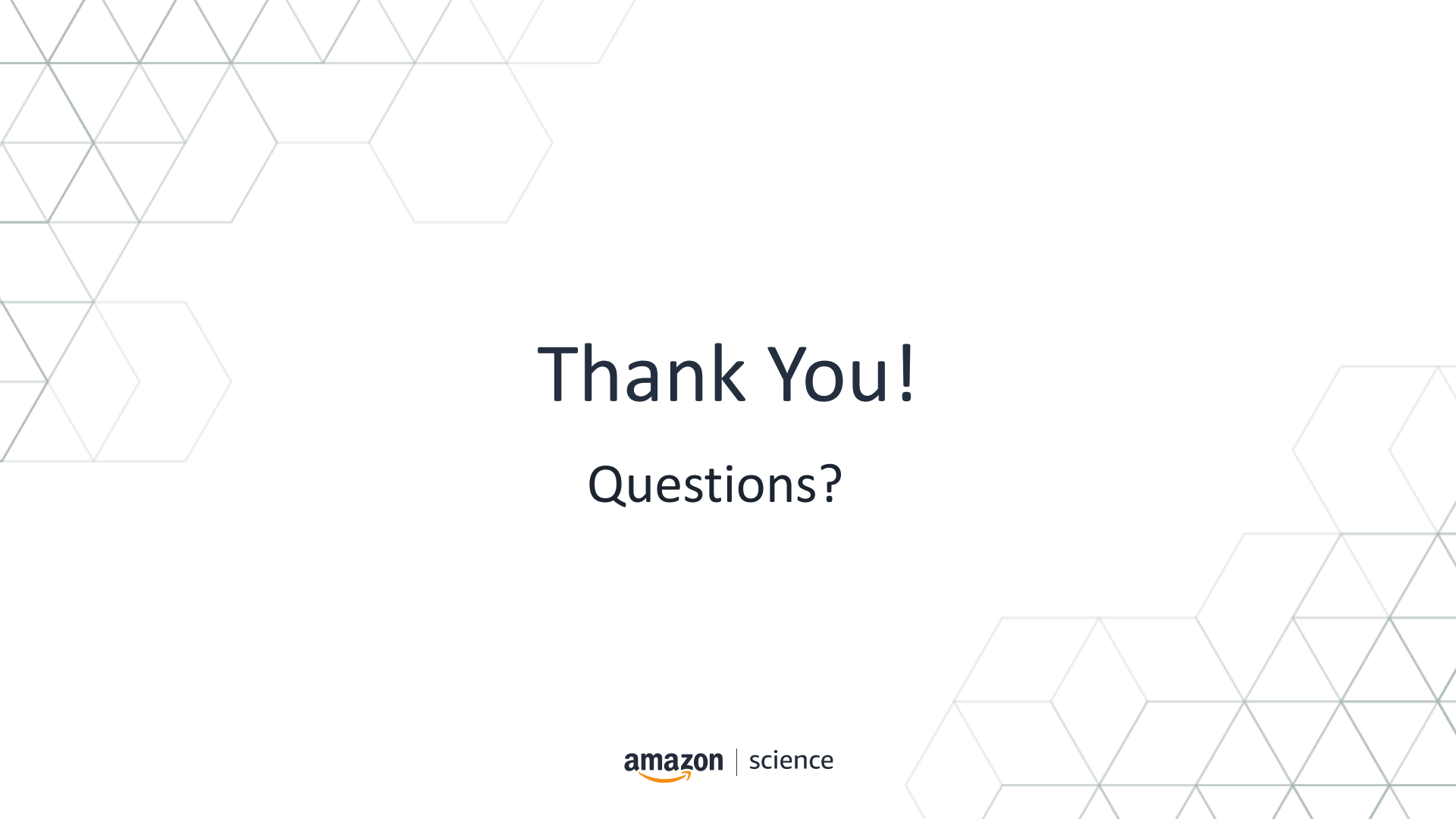
- We have proposed a graph-based approach to ASR rescoring *across utterances* that
 - only uses utterance-utterance similarity in the acoustic space, modeled by a DTW-based distance
 - is designed to help ASR adapt to idiosyncratic pronunciations, accents, or out-of-domain content
 - uses graph LP to ensure that similar-sounding utterances have similar hypotheses
- Experiments on a regional accents dataset demonstrate that our approach consistently reduces error rates, especially for accents underrepresented in the training set
- The method is well-suited to offline ASR (for example, teacher ASR for semi-supervised training)
- Does not require adaptation or fine-tuning of the baseline model
- Can be combined with (applied after) single-utterance based (e.g., LM) rescoring methods

Wrapping Up



Takeaways

- Part 1 (Fairness)
 - Both speaker recognition (ID, verification) and speech recognition systems suffer from unequal performance for different groups, due to underrepresentation in the training data
 - Mitigation by increasing representation of the targeted group in the overall training loss
 - Range of techniques (oversampling, loss function mods, data fabrication, model fusion)
- Part 2 (Graph-LP)
 - Graph-LP is a natural and effective for unsupervised learning in speaker recognition
 - Demonstrated a mapping to ASR rescoring that exploits cross-utterance similarities (relying less on training data)
 - Empirically, graph-LP also benefits ASR fairness



Thank You!
Questions?