Mixup-Based Knowledge Distillation with Causal Intervention for Multi-Task Speech Classification

Kwangje Baeg, Hyeopwoo Lee, Yeomin Yoon, Jongmo Kim

Samsung Financial Networks

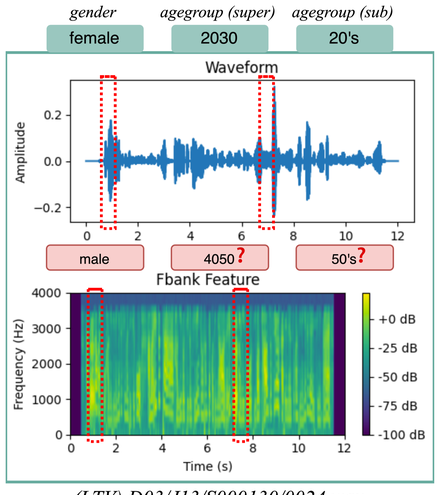
NEURAL INFORMATION PROCESSING SYSTEMS

ABSTRACT

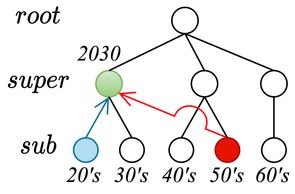
Speech classification is an essential yet challenging subtask of multitask classification, which determines the gender and age groups of speakers. Existing methods face challenges while extracting the correct features indicative of some age groups that have several ambiguities of age perception in speech. Furthermore, the methods cannot fully understand the causal inferences between speech representation and multi-label spaces. In this study, the causes of ambiguous age group boundaries are attributed to the considerable variability in speech, even within the same age group. Additionally, features that indicate speech from the 20's can be shared by some age groups in their 30's. Therefore, a two-step approach to (1) mixup-based knowledge distillation to remove biased knowledge with causal intervention and (2) hierarchical multi-task learning with causal inference for the age group hierarchy to utilize the shared information of label dependencies is proposed. Empirical experiments on Korean open-set speech corpora demonstrate that the proposed methods yield a significant performance boost in multitask speech classification.

PROBLEM STATEMENTS

- Noisy speech datasets: reverberation, multi-talker bubble, background noise, music interference, …
- **Label noise**: difficult to label, leading to disagreements among labelers, …
- **Ambiguous boundaries of age groups**: difficulties in predicting a speaker's age from speech
- Recklessly learning all the correlated features in training data (including the spurious features)



(LTV) D03/J13/S000130/0024.wav



Dependency punishment

- forcing the model to learn hierarchical information when conflicting the agegroup category from the hierarchy

PROPOSED METHODS

- The training objective for Multi-task Speech Classification Model can be derived as:

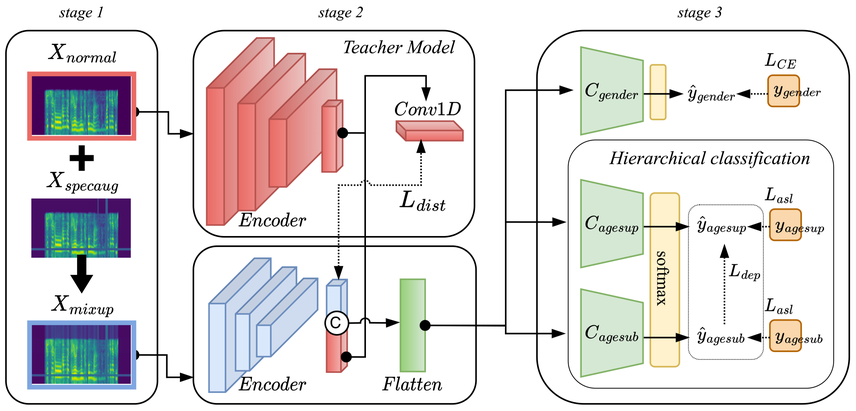
$$\mathcal{L}_{total} = \lambda \mathcal{L}_{dist}(S,T) + \mathcal{L}_{CE}(H_{gender}, Y_{gender}) + \mathcal{L}_{asl}(H_{agesup}, Y_{agesup}) + \mathcal{L}_{asl}(H_{agesub}, y_{agesub}) + \mathcal{L}_{dep}(H_{agesup}, y_{agesub})$$

Learning the Multi-task Speech Classification Model

- (stage 1) To imitate noisy label environment, the input is generated by **temporal frequency mixup** operation :

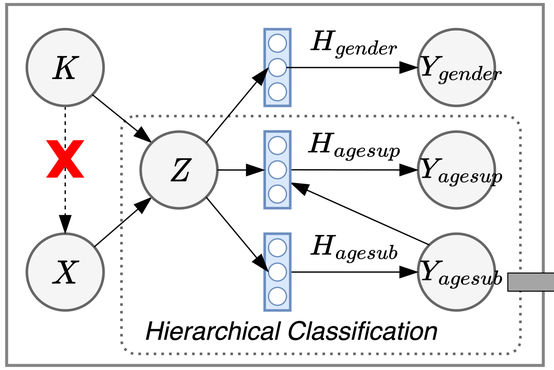
$$x^i_{mixup} = \lambda x^i_{normal} + (1-\lambda)\frac{1}{T}\sum_{j=i-\frac{T}{2}}^{i+\frac{T}{2}} (x^j_{specaug})$$

- (stage 2) The **feature-based KD methods** are used to measure the discrepancy between the teacher and the student: the MSE and Cosine embedding loss
- (stage 3) Causal approaches to **hierarchical multi-task learning** is used to learn robust representations by enforcing unseen causalities between the representation and the target



CAUSAL PERSPECTIVE

- $X \rightarrow Z \leftarrow K$: concatenating X and K
- $Z \rightarrow H \rightarrow K$: learning the causal representation Z, ensuring **invariant causal mechanisms** between the causal representation and the task labels Y
- $Y \rightarrow H \rightarrow Y$: conveying the subclass information to the superclass using **hierarchical dependency loss**



- X : representation of student model extracted from input data
- K : prior knowledge; representation of teacher model extracted from high-quality speech
- Z : shared representation of the teacher and student models
- H : hidden representation for each task
- Y : target label (e.g., gender/agesup/agesub)

$$\mathcal{L}_{+} = (1-p)^{\gamma_+} \log p$$
$$\mathcal{L}_{-} = (p_m)^{\gamma_-} \log(1-p_m)$$

Tackling high negative–positive imbalance for age group label and ground-truth mislabeling issues in classification

notated as $\mathcal{L}_{agesub}$ & $\mathcal{L}_{agesup}$

$$\mathcal{L}_{dep} = -(\mathcal{L}_{agesub})^{D_{agesup} I_{agesub}} \cdot (\mathcal{L}_{agesup})^{D_{agesup} I_{agesup}}$$

***Dependency loss**, being hierarchy-related, acted as a penalty when predictions were misaligned with a higher hierarchy, specifically agesup.*

EXPERIMENT RESULTS

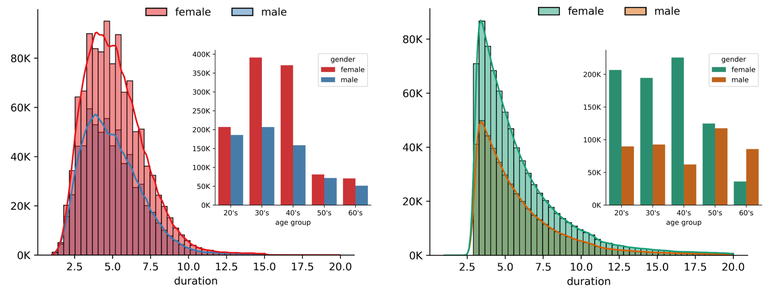


Table 1: Experimental results on FCVG, FCVE, LTV, in-house dataset

Target label			Gender			Agegroup (super class)		
Models	Dataset	Methods	P	R	F1	P	R	F1
ECAPA ResNet	FCVG† (pre-training)	base	99.85	99.83	99.84	96.67	96.90	96.78
		base	99.86	99.86	99.86	94.13	94.12	94.11
ECAPA	LTV† +FCVE†	base	97.13	97.14	97.13	79.80	68.07	70.01
		+ MKD	97.03	96.76	96.89	81.35	66.14	67.43
		(cos↑ mse↓)	(-0.22)	(-0.21)	(-0.22)	(+0.01)	(-0.05)	(-0.68)
		+ CH	97.35	97.35	97.35	83.45	83.13	81.82
		+ MKD + CH	**98.42**	**98.42**	**98.42**	**86.63**	**86.13**	**85.09**
ResNet	LTV† +FCVE†	base	95.34	95.26	95.26	73.36	73.42	73.00
		+ MKD (cos)	96.11	96.07	96.07	74.78	70.46	73.05
		+ CH	97.55	97.56	97.55	82.38	81.15	77.66
		+ MKD + CH	**98.95**	**98.96**	**98.96**	**88.40**	**87.69**	**87.83**
ECAPA	in-house call center dataset ‡	base (F/T)	96.83	96.82	96.81	62.30	62.88	60.01
		+ CH	98.32	98.29	98.29	75.13	**70.68**	70.15
		+ MKD + CH	**98.59**	**98.59**	**98.59**	**76.04**	70.65	**70.48**

† This paper used datasets from 'The Open AI Dataset Project (AI-Hub, S. Korea)'. All data information can be accessed through 'AI-Hub'. (`www.aihub.or.kr`)
– FCVG; Free Conversation Voice (General men and women); 자유대화 음성(일반남여)
– FCVE; Free Conversation Voice (Elderly men and women); 자유대화 음성(노인남여)
– LTV; Low-quality Telephone network Voice recognition data; 저음질 전화망 음성인식 데이터
‡ This is a operational data harvested from the company's in-house call center services.