INTERSPEECH 2020
(Speaker recognition)

Overview

- Attentive pooling and aggregation occupy the most.
  - Dual, Vector-based, Hierarchical, Bidirectional, Positive-wise, Character-level, mask-pooling, etc.
  - Segment, Multi-scale aggregation

- A few scraps focus on training objective …
  - Dynamic margin softmax, Angular margin centroid

Overview

- New directions
  - Audio-visual speaker recognition
  - SSL (APC)
  - Speech enhancement / quality estimation
  - Interactive training with reinforcement learning

Why Did the x-Vector System Miss a Target Speaker? Impact of Acoustic Mismatch Upon Target Score on VoxCeleb Data

Rosa Gonzalez Hautamaki and Tomi Kinnunen

Problem

- Why does a given ASV system miss (reject) a target speaker?

- Analyses
  - Predictions on average.
  - Open-world setting
  - Unlimited unknown variations.

Idea

- To model the dependency of ASV detection score upon acoustic mismatch of the enrollment and test utterances.

Methods

2.2. Mixed effects model

In LME models [10], predictors that are common to all observations are known as *fixed effects*. They are represented by means of contrast. In our model, these are the acoustic distances for each single target trial. Factors that are considered as a sample of a population, in turn, are known as *random effects*. The random effects in our model are the speakers. The model reflects variations associated with the speakers, as a variable with zero mean and unknown variance.

To be more specific, our model is defined as:

$$y_{ij} = \boldsymbol{\beta}^t \boldsymbol{x}_{ij} + b_i + \varepsilon_{ij}, \tag{2}$$

where y_{ij} is the LLR score for the jth trial of target speaker i, $\boldsymbol{\beta}^t \boldsymbol{x}_{ij}$ is the fixed effect part (acoustic distances and their weights), b_i is the per-speaker *random effect* and ε_{ij} is the residual. The assumption for a random speaker effect and the residual error is that they are independent of each other and follow a normal distribution: $b_i \sim \mathcal{N}(0, {\sigma_b}^2)$ and $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$.

Let $\mathcal{U} = (\mathcal{U}_\text{e}, \mathcal{U}_\text{t})$ denote a pair of enrollment and test utterances. An ASV system produces a *log-likelihood ratio* (LLR) score (*dependent variable*, y) between the two utterances as,

$$y = \log \frac{p(\mathcal{U}|H_0, \boldsymbol{\theta}_\text{asv})}{p(\mathcal{U}|H_1, \boldsymbol{\theta}_\text{asv})}, \tag{1}$$

where H_0 and H_1 represent the target (same-speaker) and non-target (different-speaker) hypotheses, respectively, and $\boldsymbol{\theta}_\text{asv}$ encapsulates all the ASV parameters. In our case, (1) represents LLR score from a *probabilistic linear discriminant analysis* (PLDA) back-end classifier [3], while the two utterances are represented using their x-vector [2] speaker embeddings. The

While y serves as the response variable, our predictor variables, x, are formed by *acoustic distances* of the form $x = |\varphi(f(\mathcal{U}_\text{e})) - \varphi(f(\mathcal{U}_\text{t}))|$. Here $f(\cdot)$ is a short-term (frame-level) feature extractor that converts a speech utterance into a sequence of scalar features, and $\varphi(\cdot)$ is a fixed summary statistics operator. By including different features and summary operators, we come up with a vector of D acoustical predictors, $\boldsymbol{x} = (x_1, \ldots, x_D)$ for any utterance pair $(\mathcal{U}_\text{e}, \mathcal{U}_\text{t})$. In this work, $\varphi \in \{\texttt{mean}, \texttt{std}\}$ consists of mean and standard deviation while the features include various standard speech features (see

Acoustic features

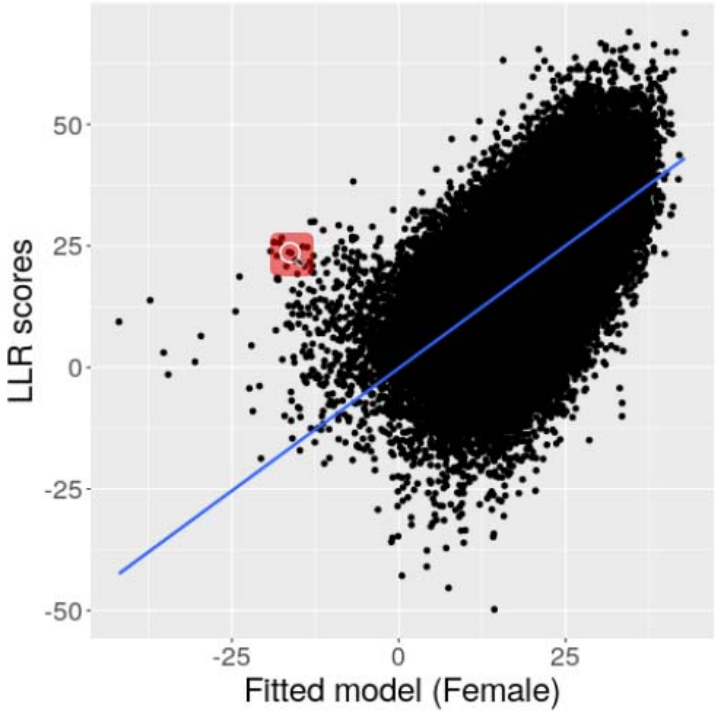
Table 1: *The mixed effect model uses a total of 23 predictor features, formed from the following combinations of features and their long-term statistical summary measures.*

Acoustic features, f		
F0	Fundamental frequency	F0
VQ	Loudness	
	Jitter	
	Shimmer	
	log Harmonic-to-noise-Ratio	HNR
	Spectral tilt	H1 – H2
		H1 – A3
Formant	Formant frequencies,	F1 to F4
	formant bandwidths,	B1 to B4
	formant amplitudes	A1 to A4
Spectral f.	Spectral flux	
Temporal	Voiced segments per second	
	Voiced segments length	
	Unvoiced segments length	

Male speakers				
Fixed effects:				
	Estimate	Std. error	t-value	r
β_0: Intercept	28.36	0.19	149.3	
β_1: F0	−1.02	0.02	−47.81	0.54
β_2: VQ	−0.36	0.006	−56.72	0.54
β_3: Formant 1	−0.20	0.01	−15.41	0.52
β_4: Formant 2	−0.15	0.01	−10.04	0.51
β_5: Formant 3	−0.16	0.01	−11.11	0.51
β_6: Formant 4	−0.31	0.01	−30.60	0.50
β_7: Temporal	−0.01	0.009	−1.56	0.48
β_8: Spectral flux	−0.29	0.007	−38.43	0.47
Random effects:				
	Variance			
Speaker: σ_b^2	4.67^2			
Residual:σ^2	9.01^2			

Female speakers				
Fixed effects:				
	Estimate	Std. error	t-value	r
β_0: Intercept	32.60	0.21	155.30	
β_1: F0	−1.01	0.03	−38.03	0.52
β_2: Formant 3	−0.37	0.02	−22.31	0.52
β_3: VQ	−0.25	0.008	−32.56	0.52
β_4: Formant 2	−0.41	0.02	−23.96	0.52
β_5: Formant 1	−0.29	0.01	−19.55	0.51
β_6: Formant 4	−0.32	0.01	−26.43	0.51
β_8: Spectral flux	−0.29	0.009	−32.51	0.47
β_7: Temporal	−0.13	0.01	−12.36	0.47
Random effects:				
	Variance			
Speaker: σ_b^2	4.5^2			
Residual:σ^2	9.3^2			

Conclusions



Overall, the acoustic variation impacts strongly the score of the ASV system. We found correlations up to ≈ 0.6 of the fitted model and the LLR score. Interestingly, our analysis confirms an important finding noted in [5] for a completely different corpus (but the same, Kaldi x-vector system): F0 mismatch plays a key role. Unsurprisingly, differences in formants and voice quality parameters contribute to degraded score, too.

Thinking

- How to *interpret* the ASV score especially predicted through deep neural networks (DNNs) ?
- How to build an explainable ASV system ?

- From speech signal analysis
  - Which acoustic feature makes the score strange ?
- From speech information analysis
  - Which information factor makes the score strange ?

Intra-class variation reduction of speaker representation in disentanglement framework

Yoohwan Kwon, Soo-Whan Chung and Hong-Goo

Motivation

- Current speaker embedding still include speaker-unrelated information.

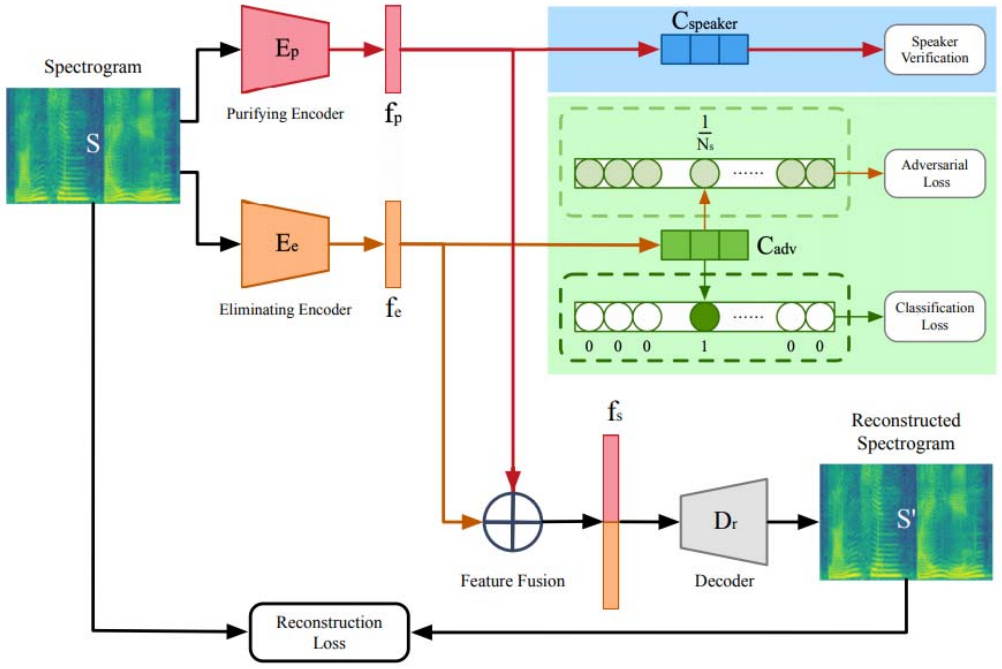
- To disentangle the embeddings with the use of relevant and irrelevant speaker information

Disentangled feature learning

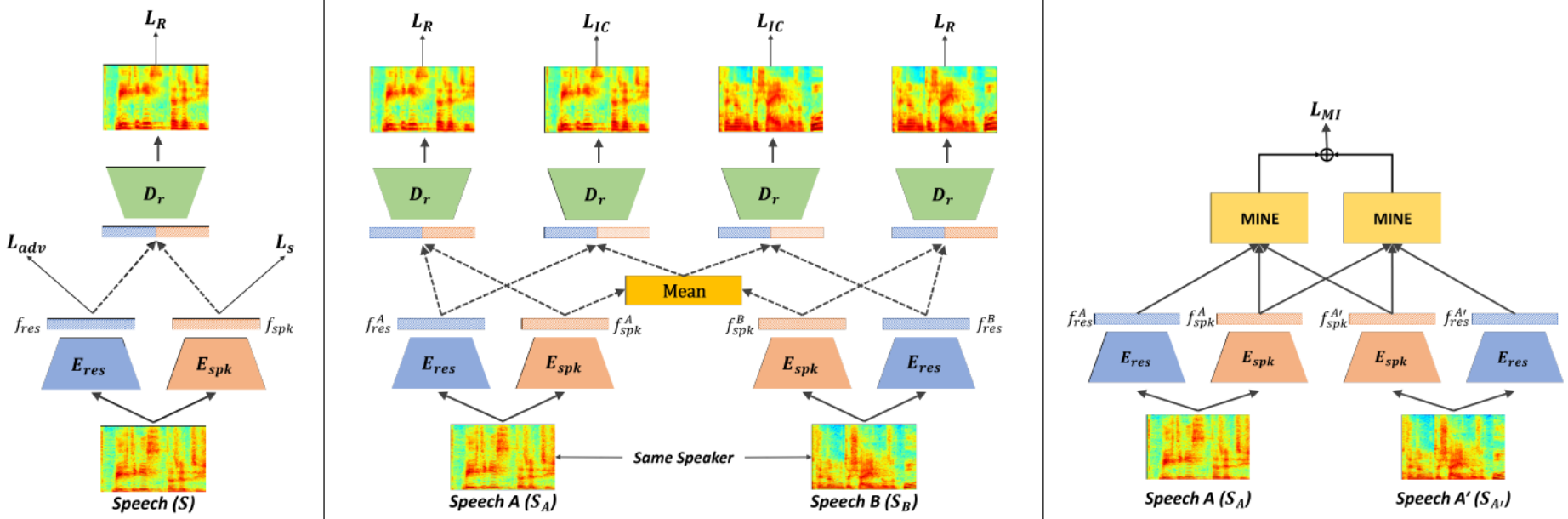
2.2. Disentangled feature learning

Disentanglement is a learning technique that represents the input signal's characteristics through multiple separated dimensions or embeddings. Therefore, it is beneficial for obtaining representations that contain certain attributes or for extracting discriminative features. Adversarial training [19–23] and reconstruction based training [24–28] are widely used to obtain disentangled representations.

Tai at el. [5] proposed a disentanglement method for speaker recognition that is the baseline for our work. By constructing an identity-related and an identity-unrelated encoder, they trained each encoder to represent only speaker-related and -unrelated information using speaker identification loss and adversarial training loss. They also adopted an auto-encoder framework to maintain all input speech information within output embeddings. The information contained in the output embeddings is preserved using spectral reconstruction approaches.



Methods



(a) *Baseline loss* (b) *Identity change loss* (c) *Mutual information loss*

Figure 1: *Overview of proposed training criteria.* (a) *Training criteria based on [5]: speaker loss, disentanglement loss and reconstruction loss.* (b) *Identity change loss: switch the speaker embedding to mean of those.* (c) *Mutual information loss: estimate the mutual information from speaker and residual embeddings by MINE*

Training objective

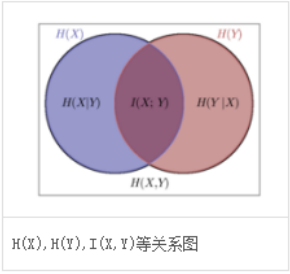
method with objective functions for training; speaker loss L_S, disentanglement loss L_{MI}, reconstruction loss L_R and identity change loss L_{IC} . The total objective function of the proposed method consists of four loss functions:

$$L_{total} = \lambda_1 L_S + \lambda_2 L_{MI} + \lambda_3 L_R + \lambda_4 L_{IC}. \quad (2)$$

The hyper-parameters are set based on experimental results, $[\lambda_1, ..., \lambda_4] = [1, 0.1, 0.1, 0.1]$.

设两个随机变量 (X, Y) 的联合分布为 $p(x, y)$ ，边缘分布分别为 $p(x), p(y)$ ，互信息 $I(X;Y)$ 是联合分布 $p(x, y)$ 与边缘分布 $p(x)p(y)$ 的相对熵， [2] 即

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)}$$



H(X)、H(Y)、I(X, Y)等关系图

$$L_S = -\sum_{i=1}^{C} t_i log(softmax(f_{spk})_i),$$

$$L_R = ||D_r(f_{spk}, f_{res}) - S_{mel}||^2,$$

$$L_{MI} = \mathbb{E}[T_\theta(f_{spk}^A, f_{spk}^{A'})] - log\left(\mathbb{E}\left[e^{T_\theta(f_{spk}^A, f_{res}^A)}\right]\right)$$
$$+\mathbb{E}[T_\theta(f_{spk}^{A'}, f_{spk}^A)] - log\left(\mathbb{E}\left[e^{T_\theta(f_{spk}^{A'}, f_{res}^{A'})}\right]\right),$$

$$L_{IC} = \|\hat{S}_A - S_A\|^2 + \|\hat{S}_B - S_B\|^2,$$
$$\hat{S}_A = D_r\left(\frac{f_{spk}^A + f_{spk}^B}{2}, f_{res}^A\right),$$
$$\hat{S}_B = D_r\left(\frac{f_{spk}^A + f_{spk}^B}{2}, f_{res}^B\right),$$

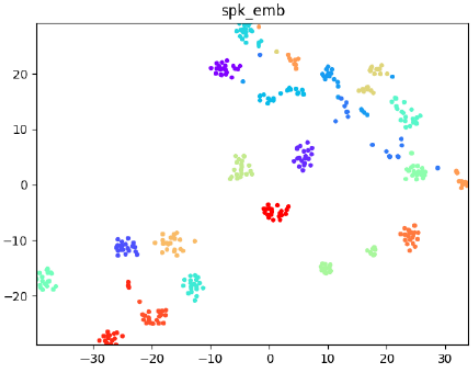
Experimental results

Table 2: *Ablation study of the proposed method*

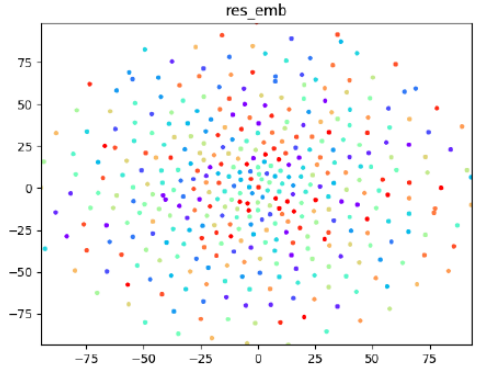
	L_s	L_r	L_{adv}	L_{mi}	L_{ic}	EER (%)
Baseline	✓	✓	✓	-	-	3.83%
Proposed	✓	✓	✓	✓	-	3.71%
	✓	✓	-	✓	-	3.81%
	✓	✓	✓	-	✓	3.59%
	✓	✓	-	✓	✓	**3.18%**

$$L_{adv} = \frac{1}{C}\sum_{j=1}^{C} log(softmax(f_{res})_j),$$

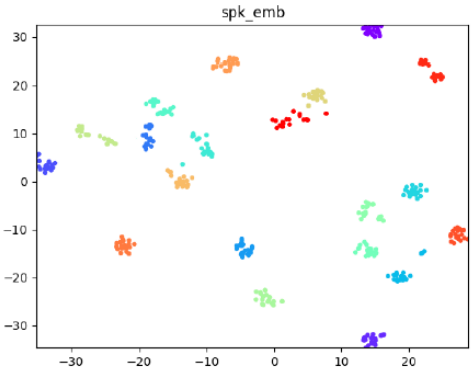
$$L_{MI} = \mathbb{E}[T_\theta(f^A_{spk}, f^{A'}_{spk})] - log\left(\mathbb{E}\left[e^{T_\theta(f^A_{spk}, f^A_{res})}\right]\right)$$
$$+\mathbb{E}[T_\theta(f^{A'}_{spk}, f^A_{spk})] - log\left(\mathbb{E}\left[e^{T_\theta(f^{A'}_{spk}, f^{A'}_{res})}\right]\right),$$



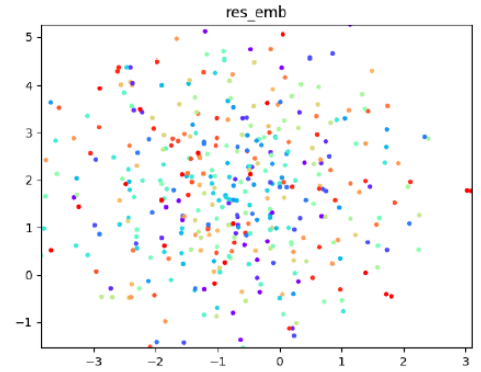
(a)



(b)



(c)



(d)

Thinking

- Flow-based disentanglement learning ?
  - Not only preserve overall information, but also constrain distribution property.

- Can we include both speaker and phonetic labels, and build a factorization model.
  - The MI and IC criteria may improve the performance of voice conversion.