USC Viterbi School of Engineering

SAIL

Robust speaker recognition using unsupervised adversarial invariance

Raghuveer Peri, Monisankha Pal, Arindam Jati, Krishna Somandepalli, Shrikanth Narayanan

ICASSP2020 Barcelona

Presented by

Raghuveer Peri
Signal Analysis and Interpretation Laboratory
University of Southern California

USC Viterbi School of Engineering

Introduction

SAIL

Goal

Extract robust, low-dimensional, speaker-discriminative representations ("*speaker embeddings*") from speech signal

Introduction

Goal

Extract robust, low-dimensional, speaker-discriminative representations ("*speaker embeddings*") from speech signal

Applications

- Automatic Speaker Verification (ASV): Verify identity of person from speech signal
- Speaker diarization: Determine who spoke when in multi-party conversations
- Automatic Speech Recognition: Speaker-adapted speech recognition models

Introduction

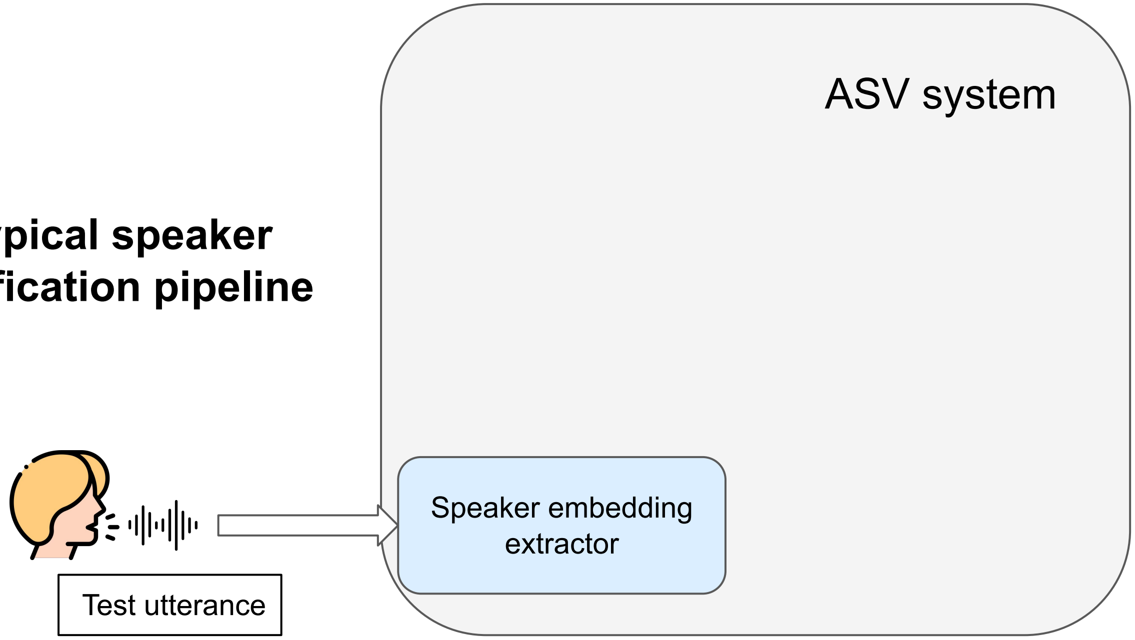
Goal

Extract robust, low-dimensional, speaker-discriminative representations (“*speaker embeddings*”) from speech signal

Applications

- Automatic Speaker Verification (ASV): Verify identity of person from speech signal
- Speaker diarization: Determine who spoke when in multi-party conversations
- Automatic Speech Recognition: Speaker-adapted speech recognition models

Typical speaker verification pipeline



Introduction

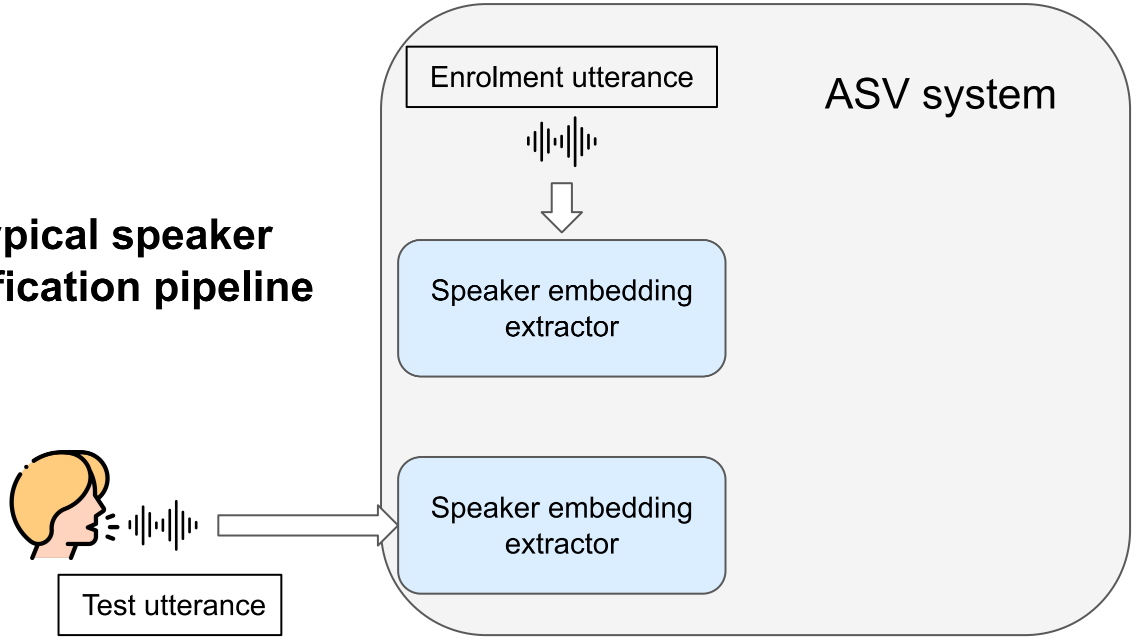
Goal

Extract robust, low-dimensional, speaker-discriminative representations ("*speaker embeddings*") from speech signal

Applications

- Automatic Speaker Verification (ASV): Verify identity of person from speech signal
- Speaker diarization: Determine who spoke when in multi-party conversations
- Automatic Speech Recognition: Speaker-adapted speech recognition models

Typical speaker verification pipeline



Introduction

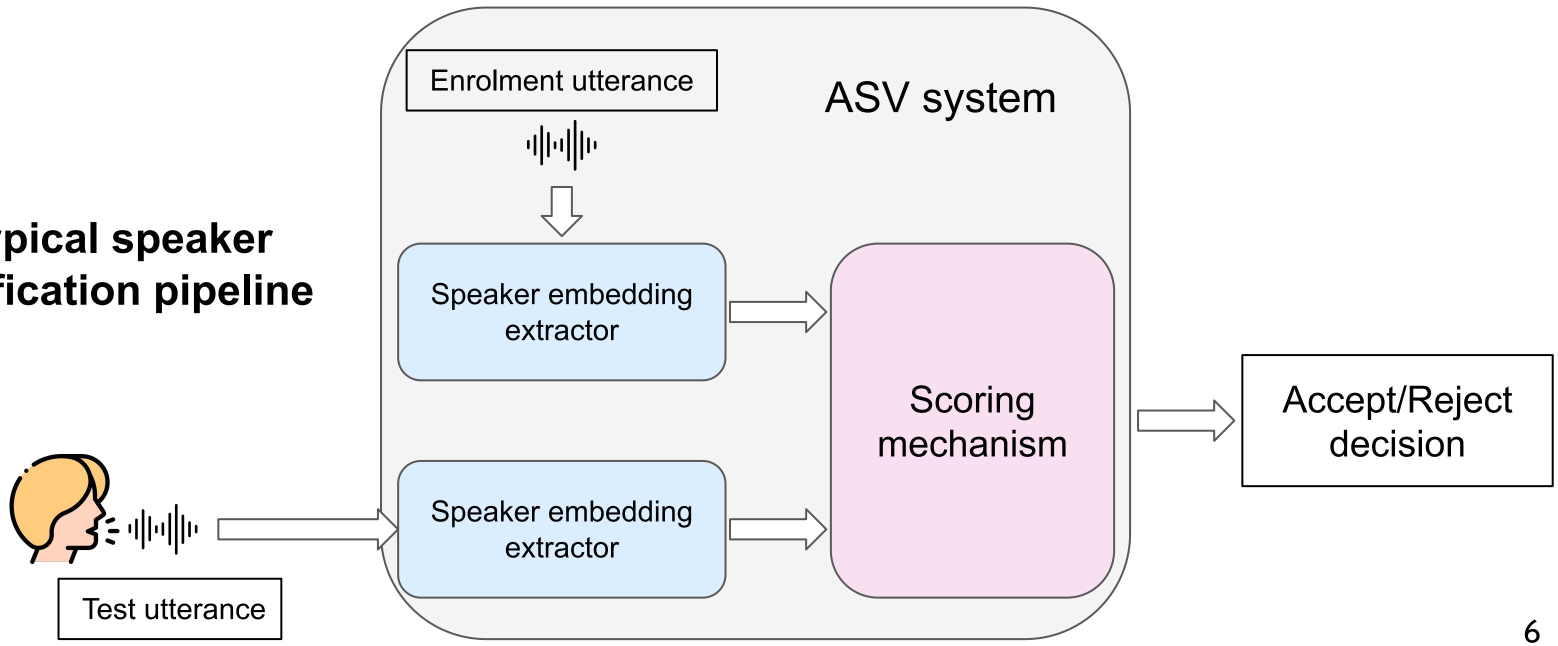
Goal

Extract robust, low-dimensional, speaker-discriminative representations ("*speaker embeddings*") from speech signal

Applications

- Automatic Speaker Verification (ASV): Verify identity of person from speech signal
- Speaker diarization: Determine who spoke when in multi-party conversations
- Automatic Speech Recognition: Speaker-adapted speech recognition models

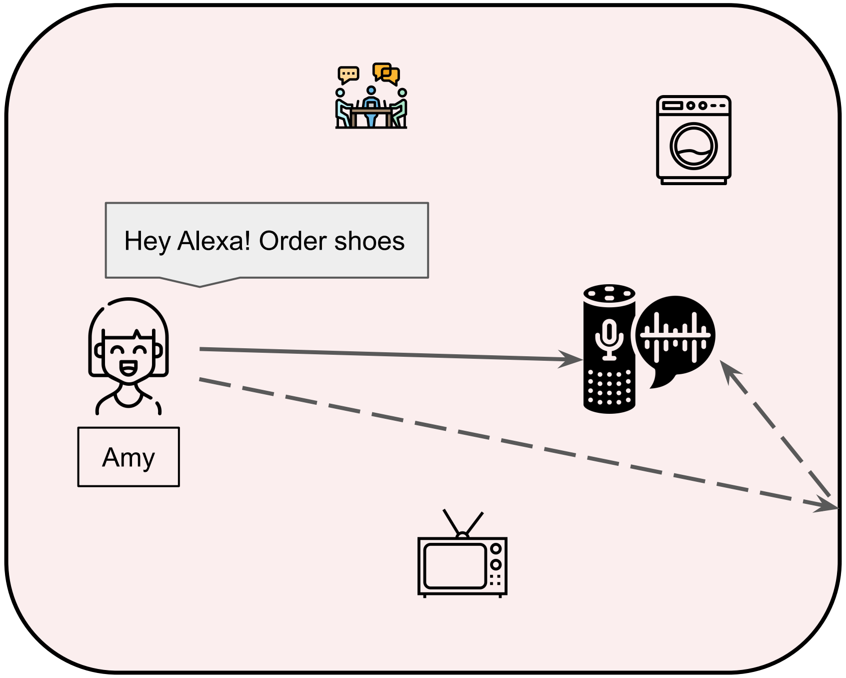
Typical speaker verification pipeline



Introduction

Challenges

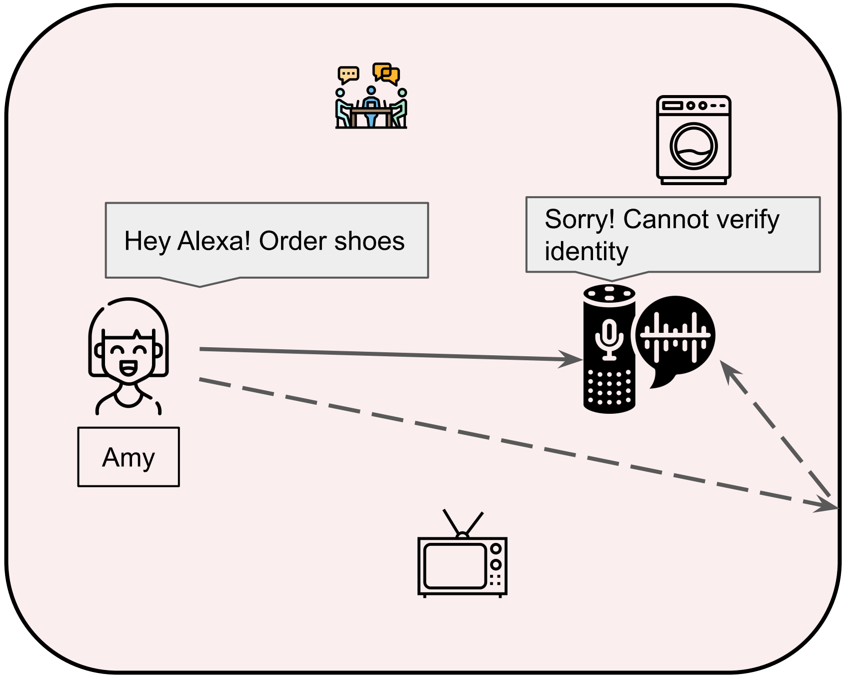
- Speech is an information-rich signal
- *Nuisance factors* unrelated to speaker identity entangled in signal
  - Channel factors
    - Acoustic noise (TV, babble etc.)
    - Reverberation
  - Content factors
    - Affective state (happy, angry etc.)
    - Linguistic content



Introduction

Challenges

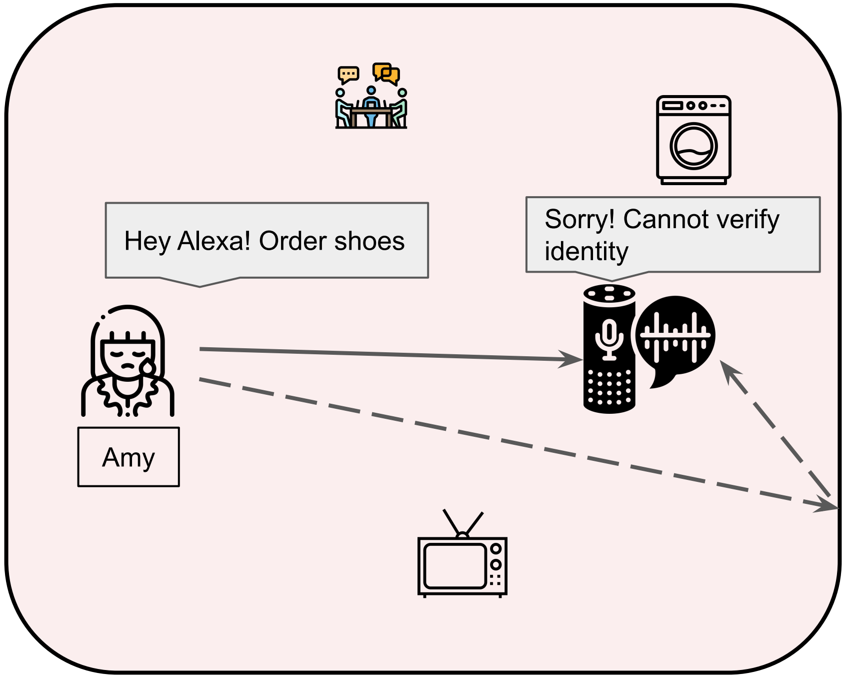
- Speech is an information-rich signal
- *Nuisance factors* unrelated to speaker identity entangled in signal
  - Channel factors
    - Acoustic noise (TV, babble etc.)
    - Reverberation
  - Content factors
    - Affective state (happy, angry etc.)
    - Linguistic content



Introduction

Challenges

- Speech is an information-rich signal
- *Nuisance factors* unrelated to speaker identity entangled in signal
  - Channel factors
    - Acoustic noise (TV, babble etc.)
    - Reverberation
  - Content factors
    - Affective state (happy, angry etc.)
    - Linguistic content



Prior work

- Total Variability Modeling (i-vectors - *Dehak et al., 2011*)
  - Capture all factors of variability in total variability space
  - Perform additional channel compensation steps, such as length normalization
- Deep learning methods (x-vectors - *Snyder et al., 2017*)
  - Train deep models on artificially augmented audio using various noise and reverberation.
  - Extract hidden layer representations as utterance-level features.
- More recent supervised domain adversarial training techniques (*Bhattacharya et al., 2019*)
  - Train models to discriminate speakers
  - Simultaneously made robust to “specific” factors of variability by training adversarially, such as known noise type or channel conditions.

Disentangling speaker representations

Prior work

- Total Variability Modeling (i-vectors - Dehak et al., 2011)
  - Capture all factors of variability in total variability space
  - Perform additional channel compensation steps, such as length normalization
- Deep learning methods (x-vectors - Snyder et al., 2017)
  - Train deep models on artificially augmented audio using various noise and reverberation.
  - Extract hidden layer representations as utterance-level features.
- More recent supervised domain adversarial training techniques (Bhattacharya et al., 2019)
  - Train models to discriminate speakers
  - Simultaneously made robust to "specific" factors of variability by training adversarially, such as known noise type or channel conditions.

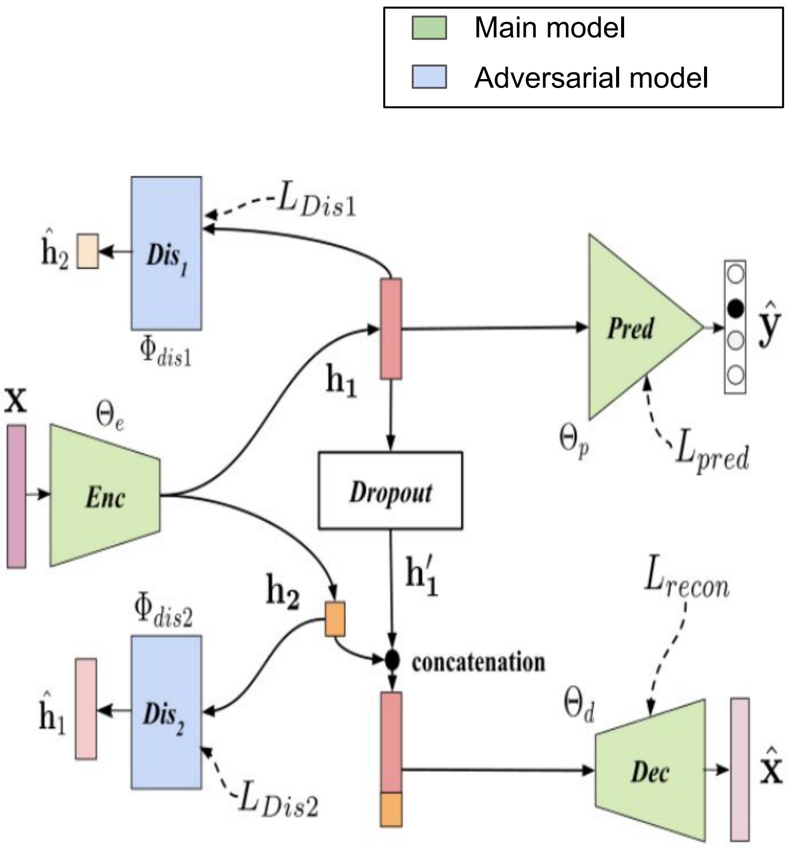
Proposed work

- Disentangle speech representations into two embeddings
  - Speaker factors
  - Nuisance factors
- **No assumptions on specific factors of variability**

Disentangling speaker representations

Input

- Speech representations (MFCC, x-vectors etc)
- Speaker labels



Disentangling speaker representations

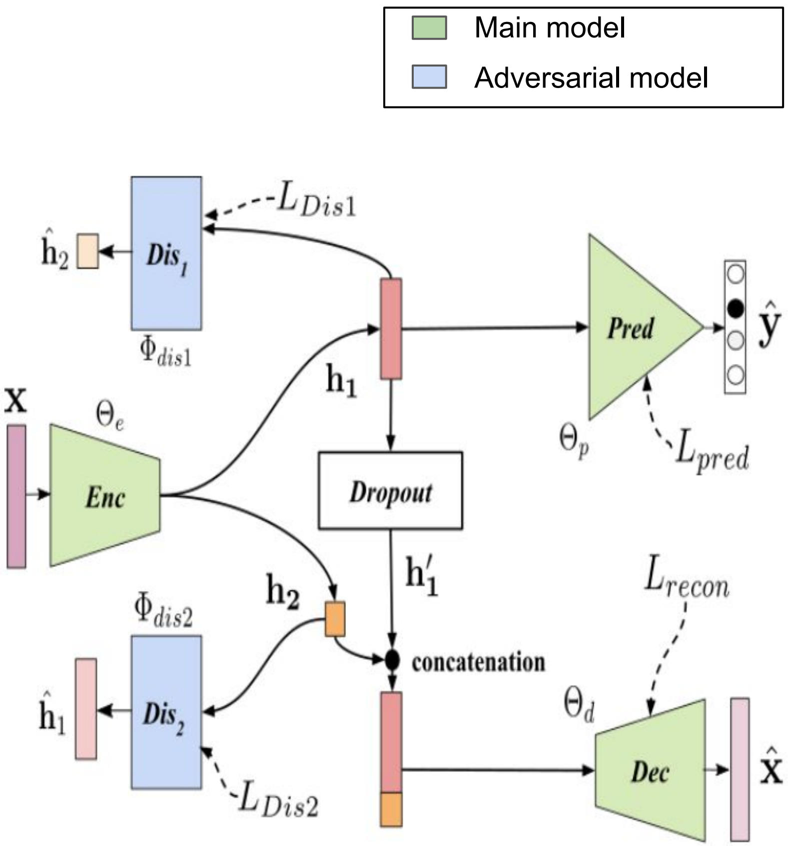
Input

- Speech representations (MFCC, x-vectors etc)
- Speaker labels

Main model

- Predictor (*Pred*): Predicts speakers

- Decoder (*Dec*): Reconstruct input



Disentangling speaker representations

Input

- Speech representations (MFCC, x-vectors etc)
- Speaker labels

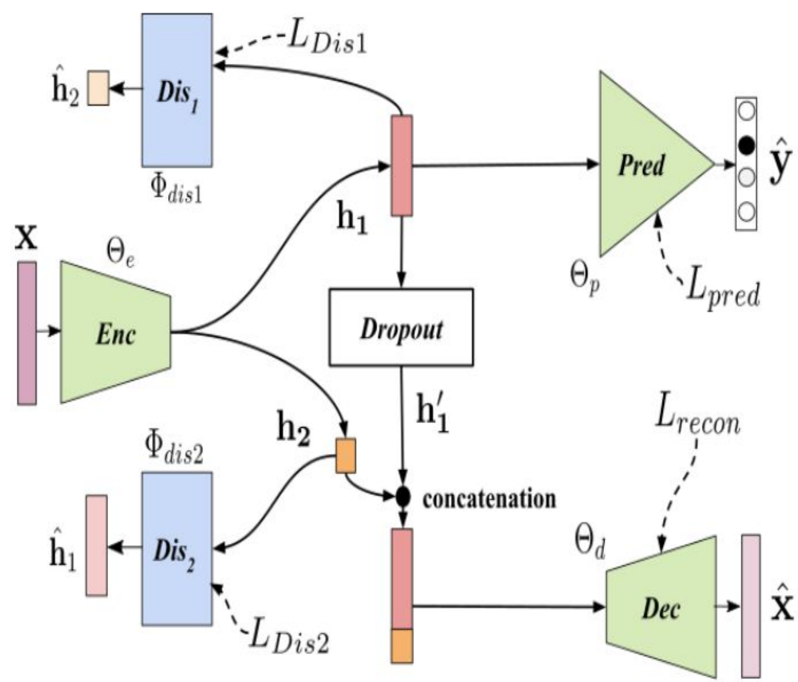
Main model

- Predictor (*Pred*): Predicts speakers
- Decoder (*Dec*): Reconstruct input

Adversarial model

- Disentanglers (*Dis*$_1$ and *Dis*$_2$): Make h$_1$ and h$_2$ poor predictors of each other

Main model
Adversarial model



Disentangling speaker representations

Input

- Speech representations (MFCC, x-vectors etc)
- Speaker labels

Main model

- Predictor (*Pred*): Predicts speakers
- Decoder (*Dec*): Reconstruct input

Adversarial model

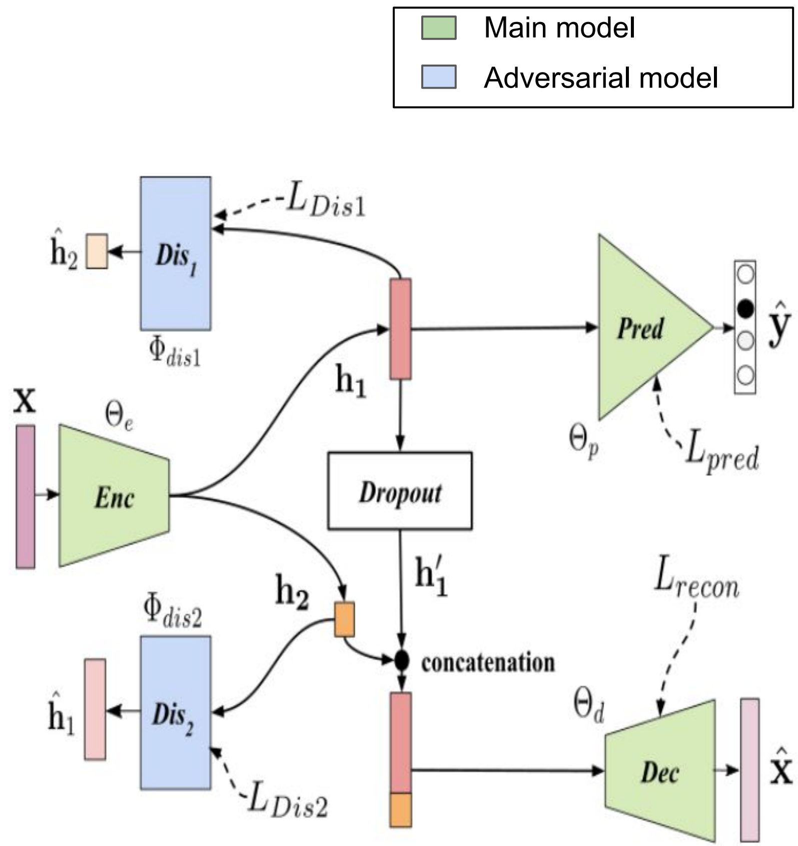
- Disentanglers (*Dis*$_1$ and *Dis*$_2$): Make h_1 and h_2 poor predictors of each other

Adversarial Training*

$$L_{main} = \alpha L_{pred}(\mathbf{y}, \hat{\mathbf{y}}) + \beta L_{recon}(\mathbf{x}, \hat{\mathbf{x}})$$

$$L_{adv} = L_{Dis1}(\mathbf{h}_2, \hat{\mathbf{h}}_2) + L_{Dis2}(\mathbf{h}_1, \hat{\mathbf{h}}_1)$$

$$\min_{\Theta_e, \Theta_d, \Theta_p} \max_{\Phi_{dis1}, \Phi_{dis2}} L_{main} + \gamma L_{adv}$$



*Jaiswal, A., Wu, R.Y., Abd-Almageed, W. and Natarajan, P., 2018. Unsupervised adversarial invariance. In Advances in Neural Information Processing Systems (pp. 5092-5102).

Disentangling speaker representations

Input

- Speech representations (MFCC, x-vectors etc)
- Speaker labels

Main model

- Predictor (*Pred*): Predicts speakers
- Decoder (*Dec*): Reconstruct input

Adversarial model

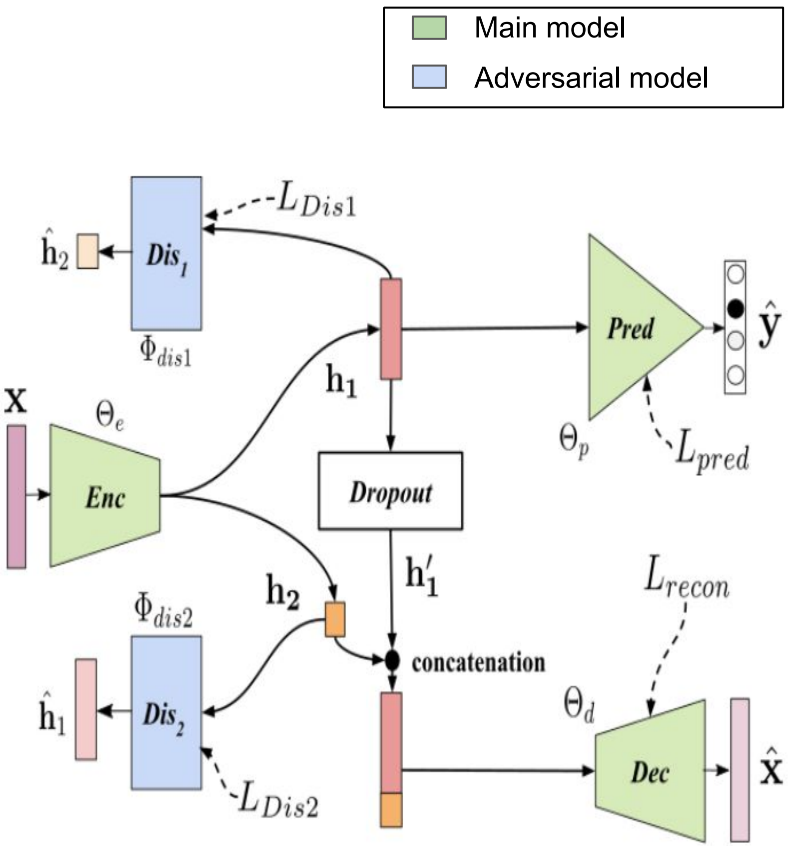
- Disentanglers (*Dis*$_1$ and *Dis*$_2$): Make h_1 and h_2 poor predictors of each other

Adversarial Training*

$$L_{main} = \alpha L_{pred}(\mathbf{y}, \hat{\mathbf{y}}) + \beta L_{recon}(\mathbf{x}, \hat{\mathbf{x}})$$

$$L_{adv} = L_{Dis1}(\mathbf{h}_2, \hat{\mathbf{h}}_2) + L_{Dis2}(\mathbf{h}_1, \hat{\mathbf{h}}_1)$$

$$\boxed{\min_{\Theta_e, \Theta_d, \Theta_p} \max_{\Phi_{dis1}, \Phi_{dis2}} L_{main} + \gamma L_{adv}}$$



*Jaiswal, A., Wu, R.Y., Abd-Almageed, W. and Natarajan, P., 2018. Unsupervised adversarial invariance. In Advances in Neural Information Processing Systems (pp. 5092-5102).

Disentangling speaker representations

Input

- Speech representations (MFCC, x-vectors etc)
- Speaker labels

Main model

- Predictor (*Pred*): Predicts speakers
- Decoder (*Dec*): Reconstruct input

Adversarial model

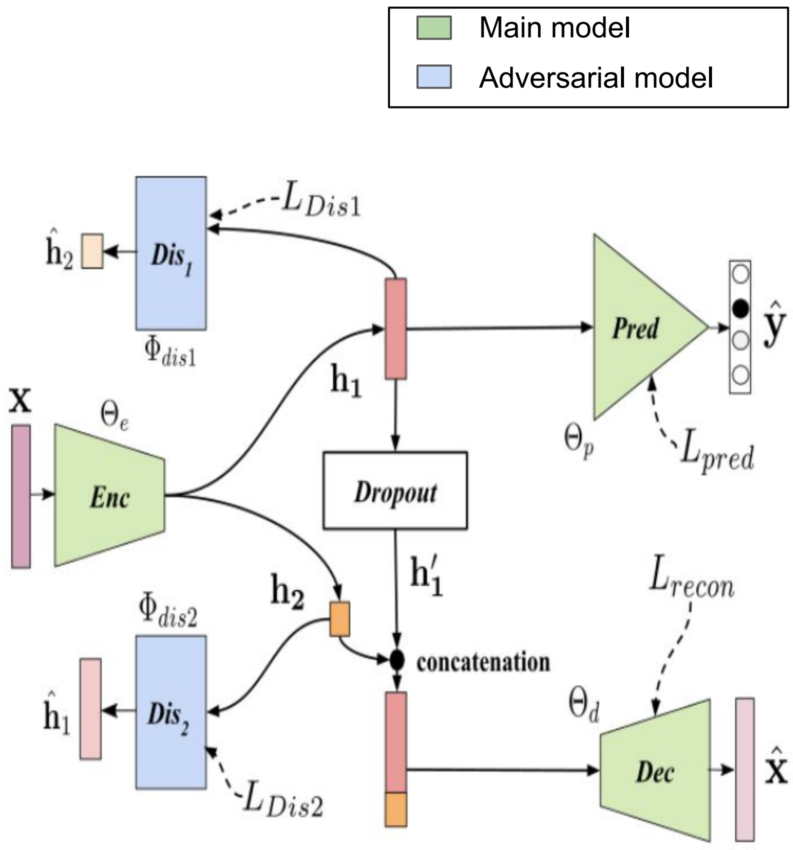
- Disentanglers (*Dis*$_1$ and *Dis*$_2$): Make h_1 and h_2 poor predictors of each other

Adversarial Training*

$$L_{main} = \alpha L_{pred}(\mathbf{y}, \hat{\mathbf{y}}) + \beta L_{recon}(\mathbf{x}, \hat{\mathbf{x}})$$

$$L_{adv} = L_{Dis1}(\mathbf{h}_2, \hat{\mathbf{h}}_2) + L_{Dis2}(\mathbf{h}_1, \hat{\mathbf{h}}_1)$$

$$\boxed{\min_{\Theta_e, \Theta_d, \Theta_p} \max_{\Phi_{dis1}, \Phi_{dis2}} L_{main} + \gamma L_{adv}}$$



- h_1 : speaker discriminative information
- h_2 : nuisance information

*Jaiswal, A., Wu, R.Y., Abd-Almageed, W. and Natarajan, P., 2018. Unsupervised adversarial invariance. In Advances in Neural Information Processing Systems (pp. 5092-5102).

Dataset and Features

Training data (VoxCeleb[1])

- Training set of VoxCeleb
  - Vox 1 (Dev)
  - Vox2 (Dev and test)
- No artificial augmentation
- 1.2M data samples
- 7323 unique speakers

Input features

- x-vectors using pre-trained model[2]

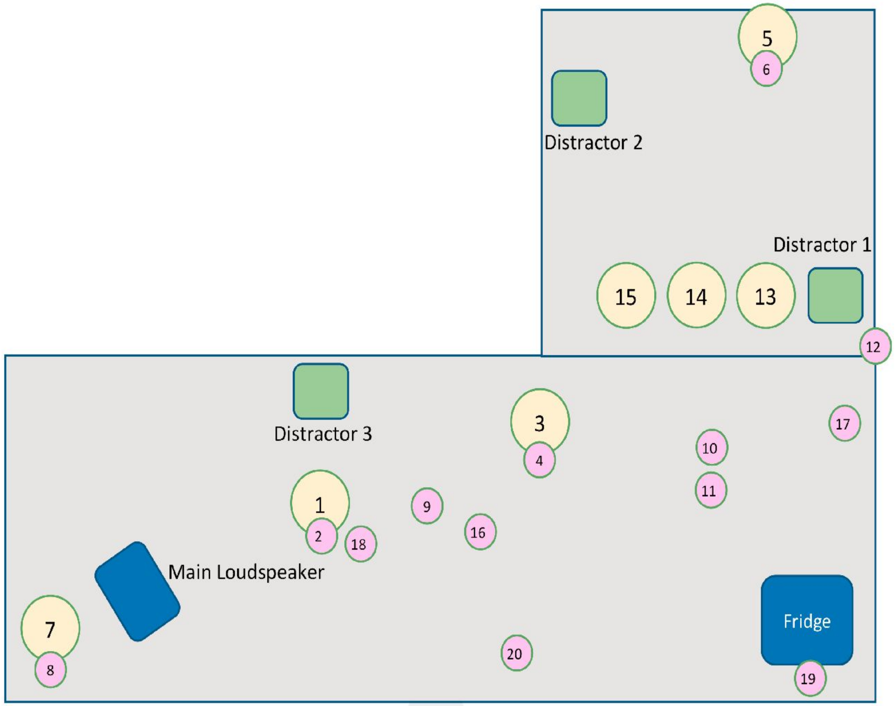
Dataset and Features

Training data (VoxCeleb[1])

- Training set of VoxCeleb
  - Vox 1 (Dev)
  - Vox2 (Dev and test)
- No artificial augmentation
- 1.2M data samples
- 7323 unique speakers

Input features

- x-vectors using pre-trained model[2]

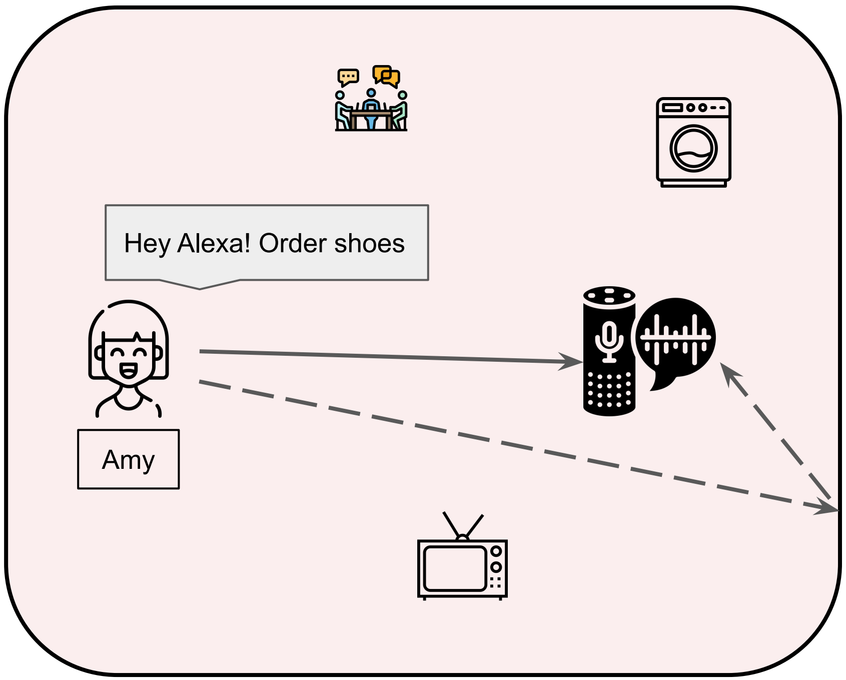


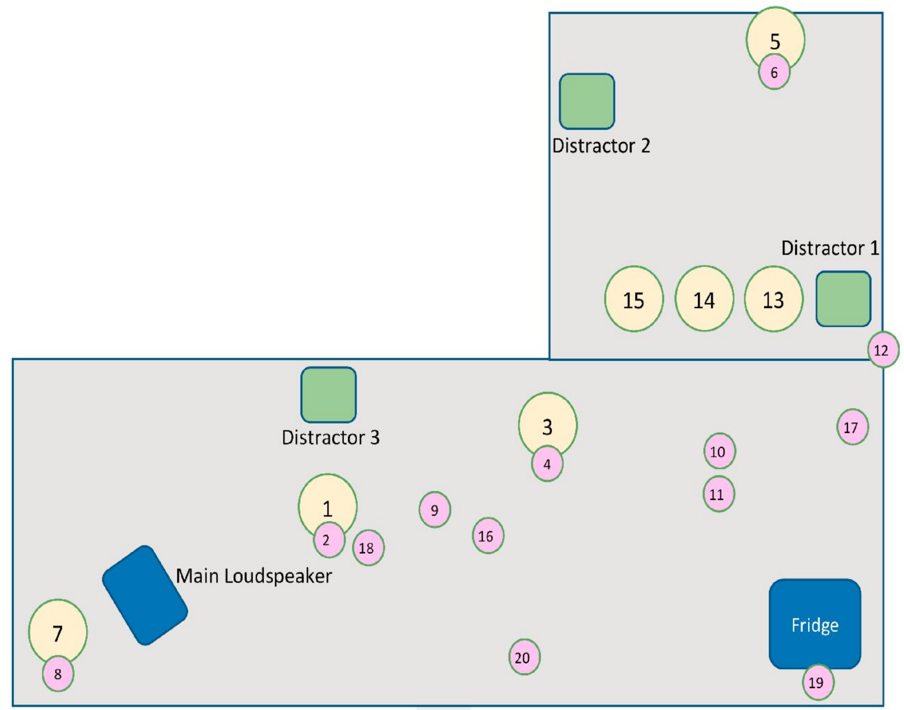
Evaluation data (VOiCES[3])

- 4 rooms
- 12-18 microphones
- 200 unique speakers
- 2 subsets: Voices-dev, Voices-eval

1. Chung, J.S., Nagrani, A. and Zisserman, A., 2018. Voxceleb2: Deep speaker recognition. *arXiv preprint arXiv:1806.05622*.
2. https://kaldi-asr.org/models/m7
3. Richey, Colleen, Maria A. Barrios, Zeb Armstrong, Chris Bartels, Horacio Franco, Martin Graciarena, Aaron Lawson et al. "Voices obscured in complex environmental settings (voices) corpus." *arXiv preprint arXiv:1804.05053* (2018).

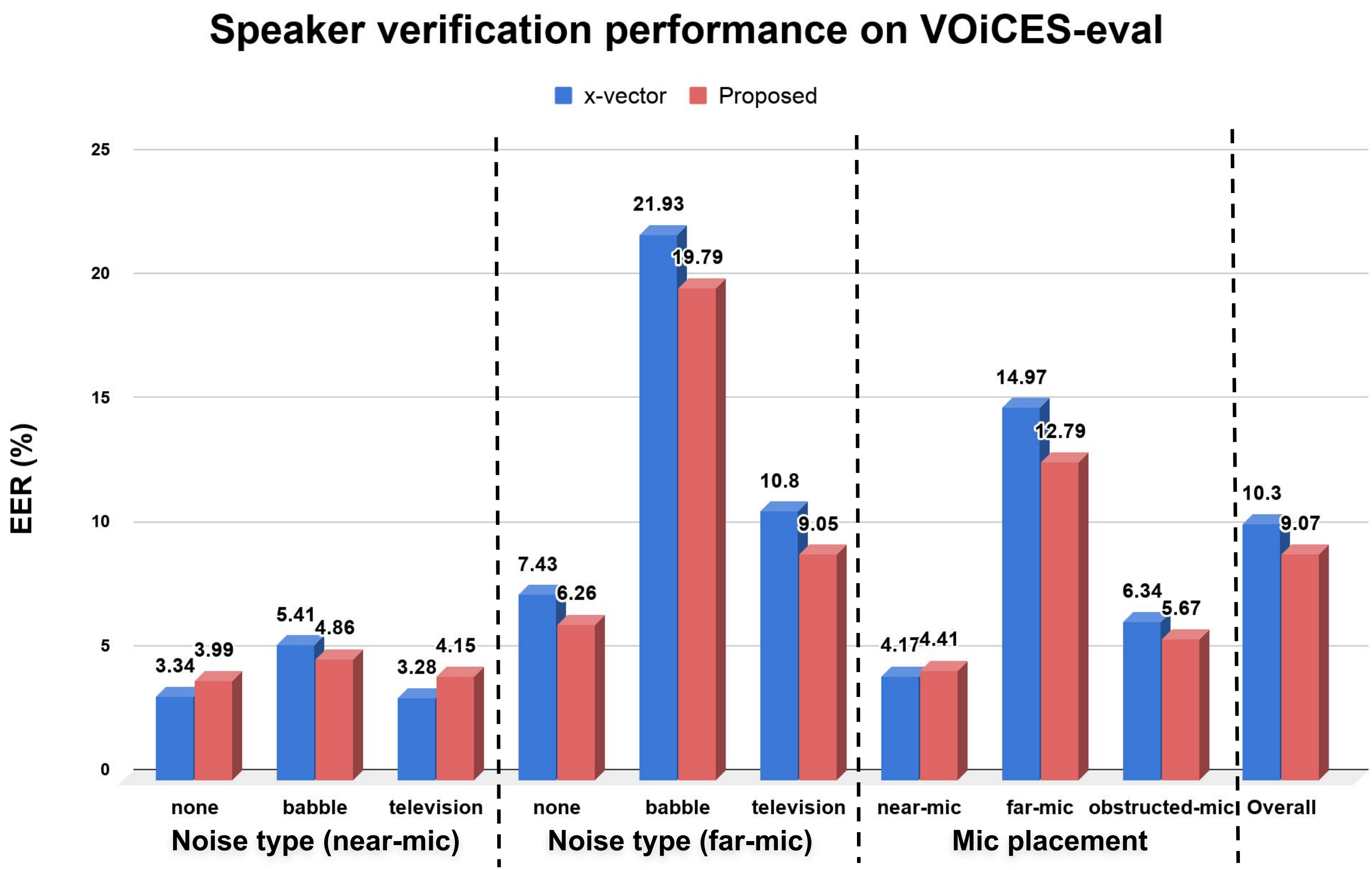
Dataset and Features





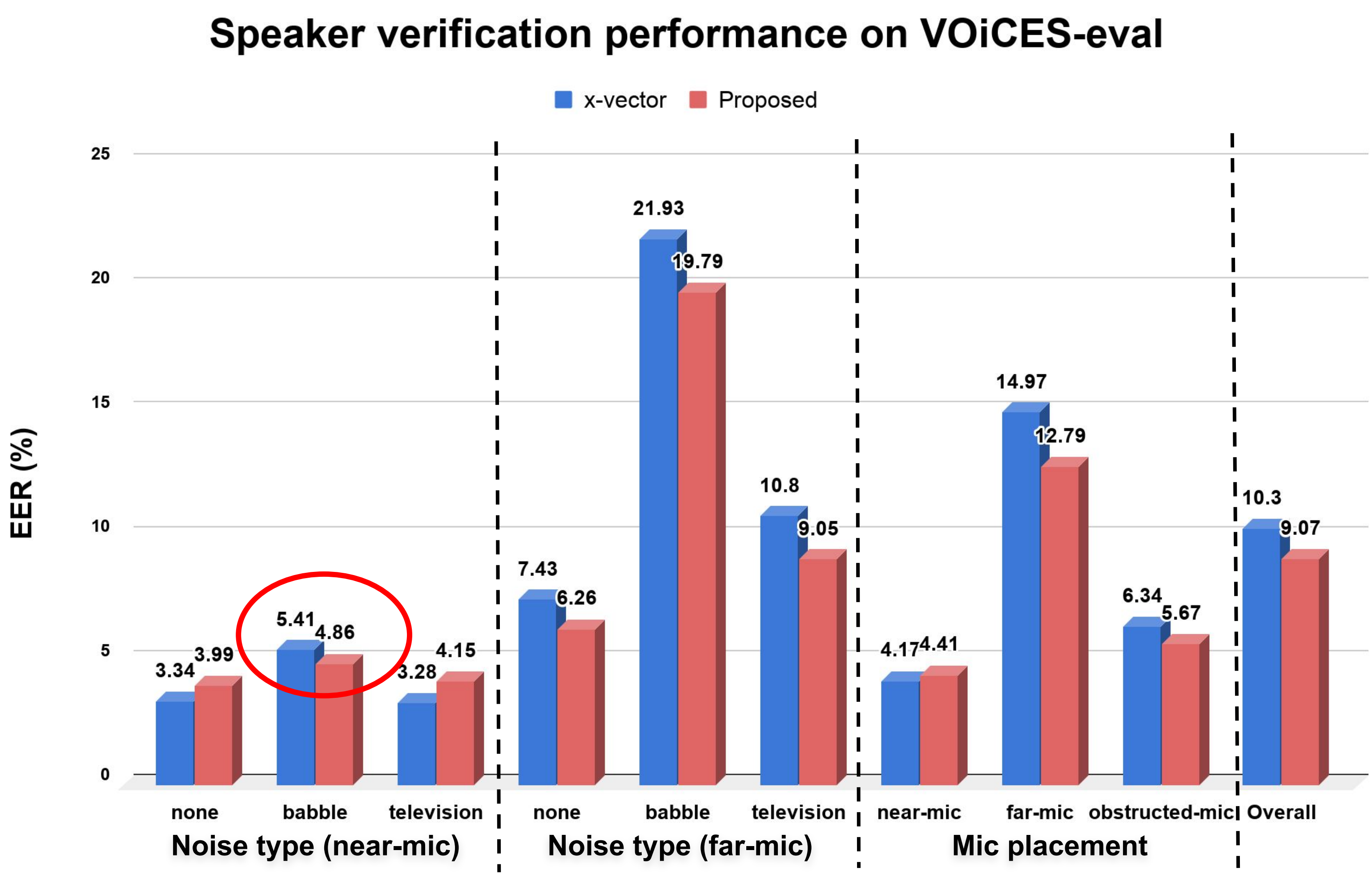
Results (Speaker Verification)

- Dimensionality reduction: Linear Discriminant Analysis (LDA)
  - x-vector - dimension 150, Proposed - dimension 96
- Verification scoring: Probabilistic LDA (PLDA)



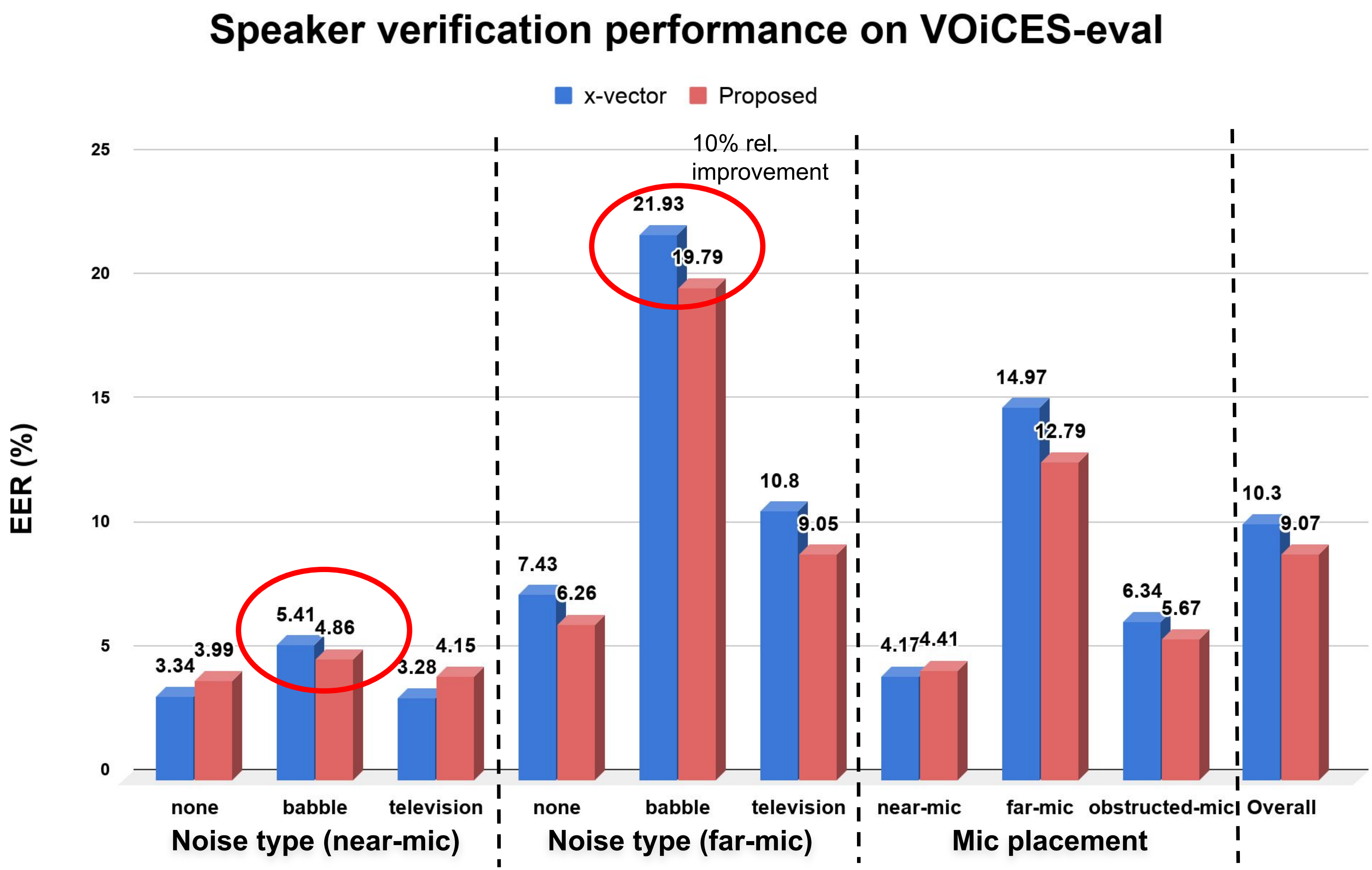
Results (Speaker Verification)

- Dimensionality reduction: Linear Discriminant Analysis (LDA)
  - x-vector - dimension 150, Proposed - dimension 96
- Verification scoring: Probabilistic LDA (PLDA)



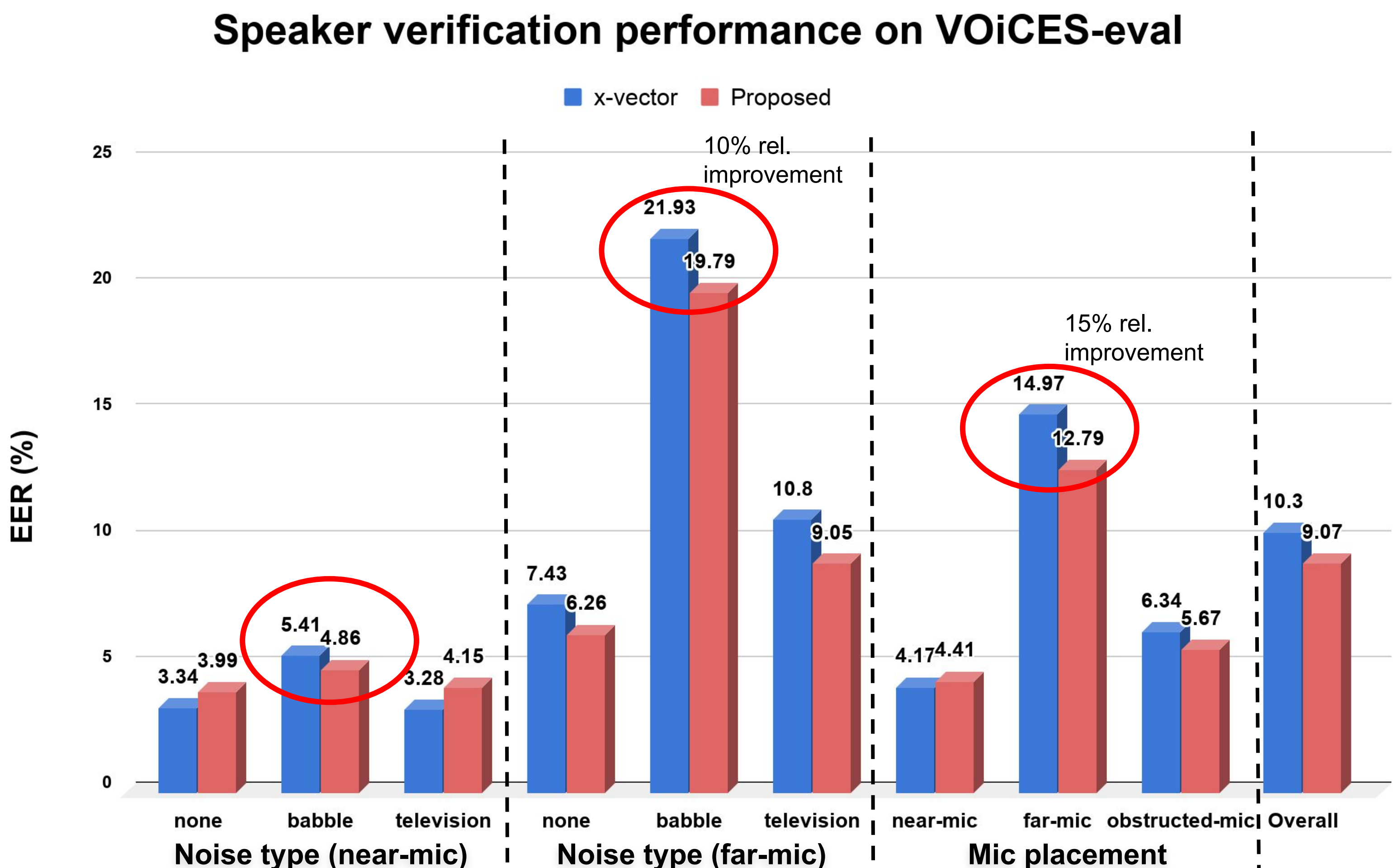
Results (Speaker Verification)

- Dimensionality reduction: Linear Discriminant Analysis (LDA)
  - x-vector - dimension 150, Proposed - dimension 96
- Verification scoring: Probabilistic LDA (PLDA)



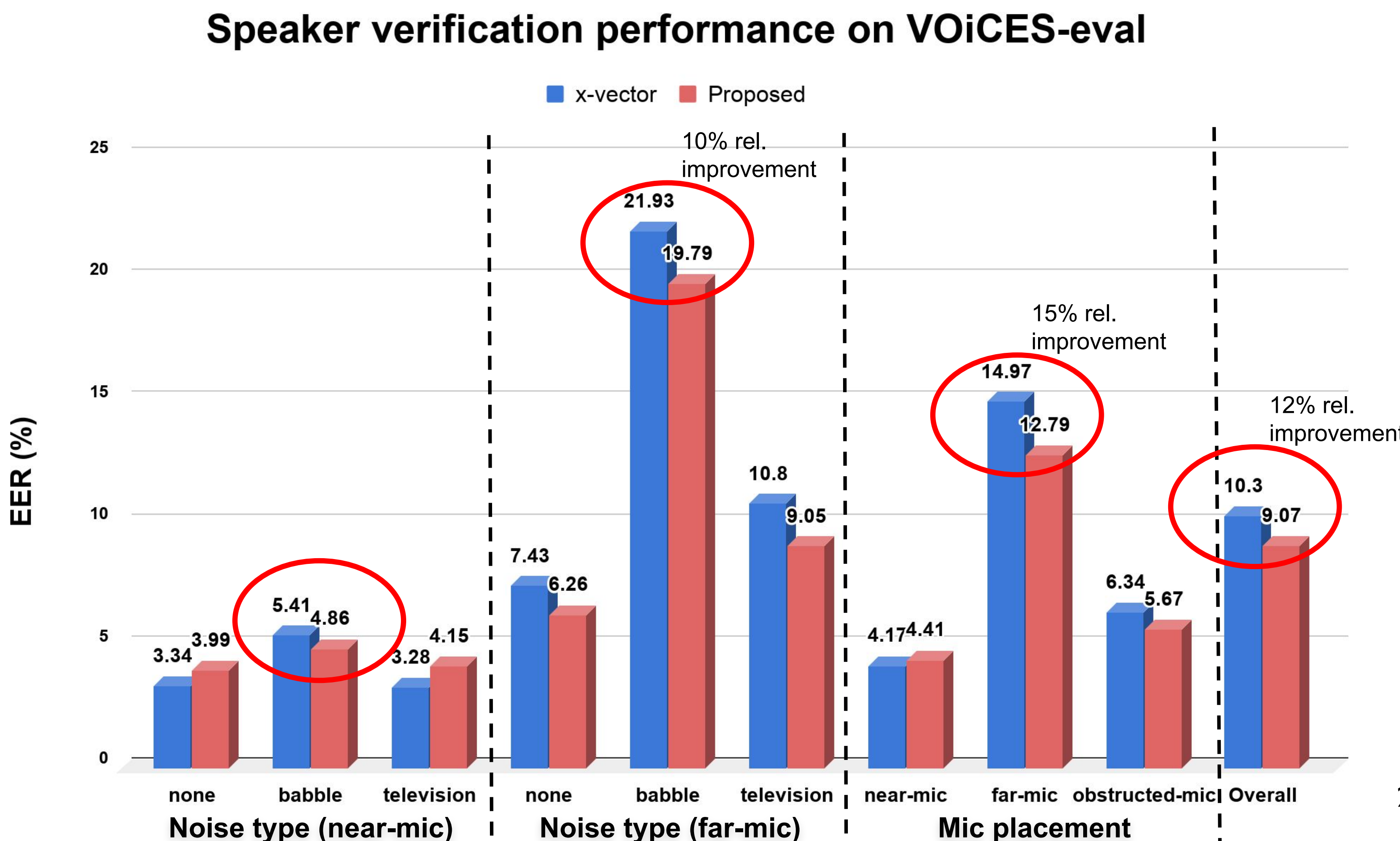
Results (Speaker Verification)

- Dimensionality reduction: Linear Discriminant Analysis (LDA)
  - x-vector - dimension 150, Proposed - dimension 96
- Verification scoring: Probabilistic LDA (PLDA)

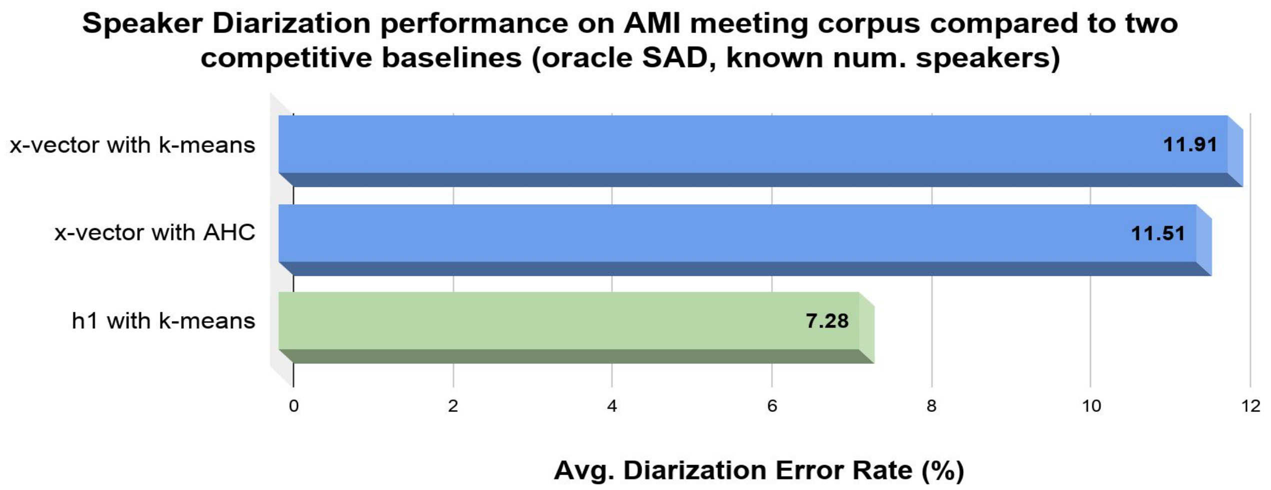


Results (Speaker Verification)

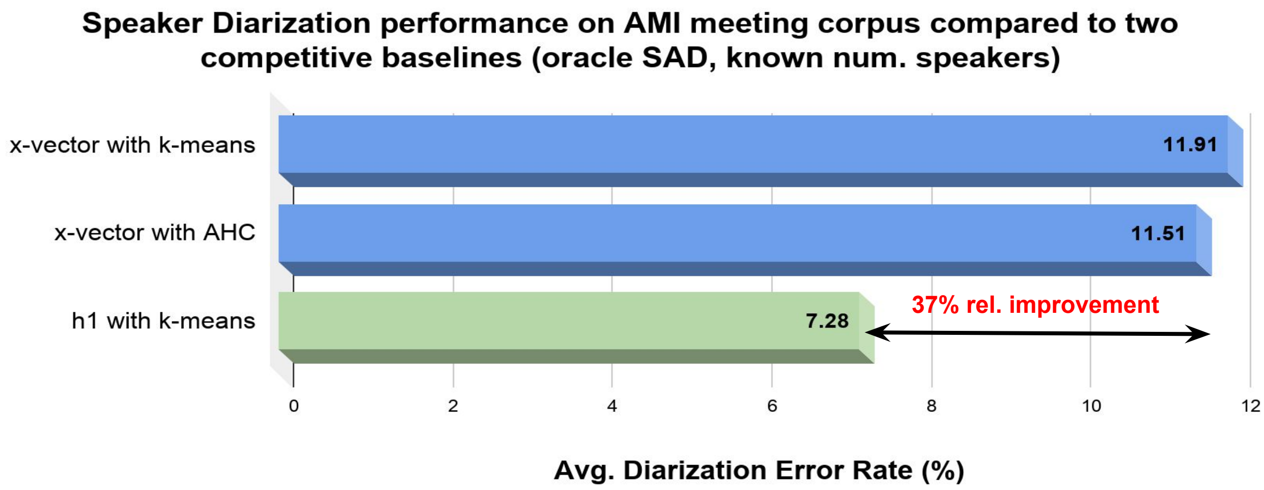
- Dimensionality reduction: Linear Discriminant Analysis (LDA)
  - x-vector - dimension 150, Proposed - dimension 96
- Verification scoring: Probabilistic LDA (PLDA)



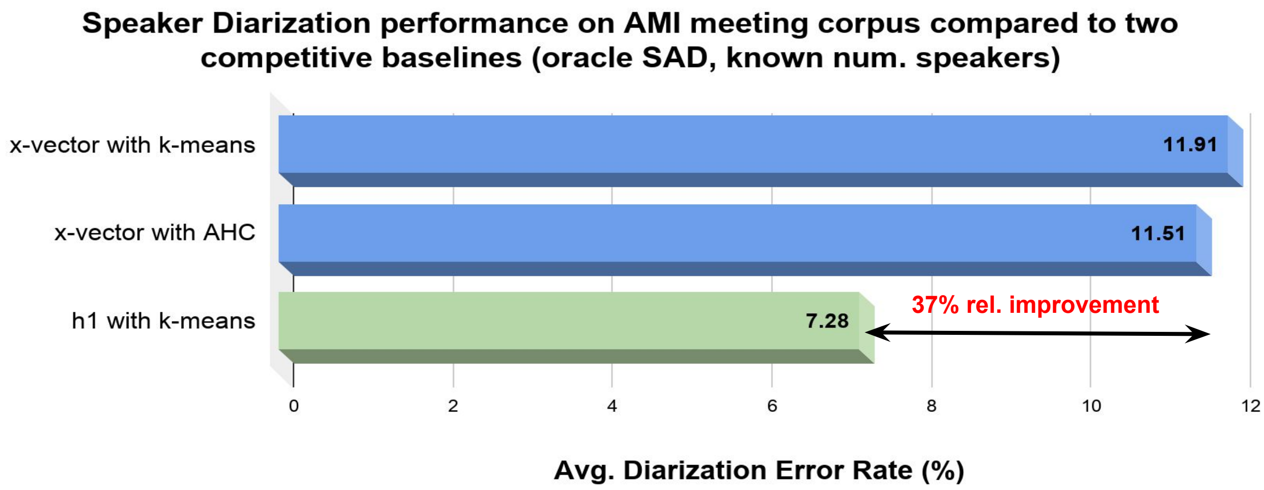
Conclusions



Conclusions



Conclusions

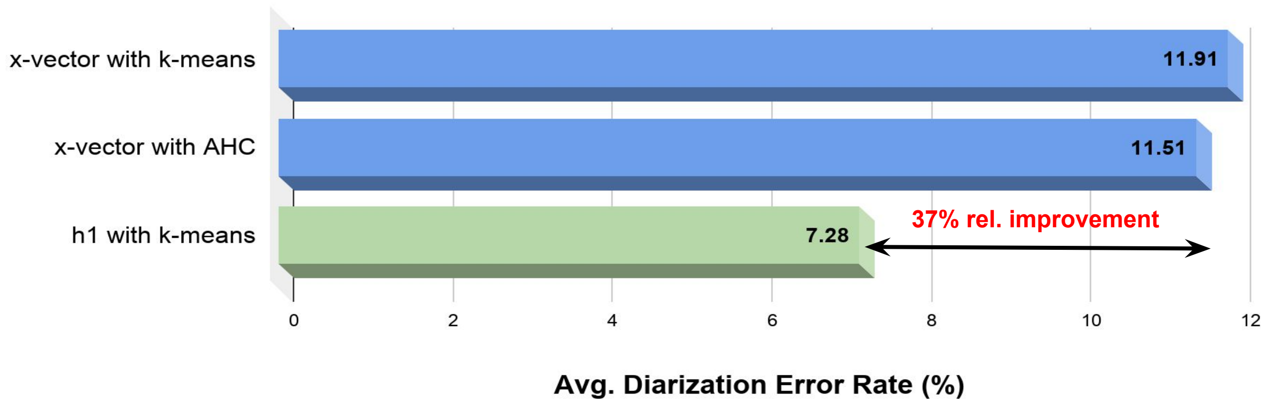


Conclusions

- Proposed novel speaker embeddings
  - Disentangled speaker and nuisance factors from speaker embeddings
  - No prior knowledge of specific nuisance factors during training

Conclusions

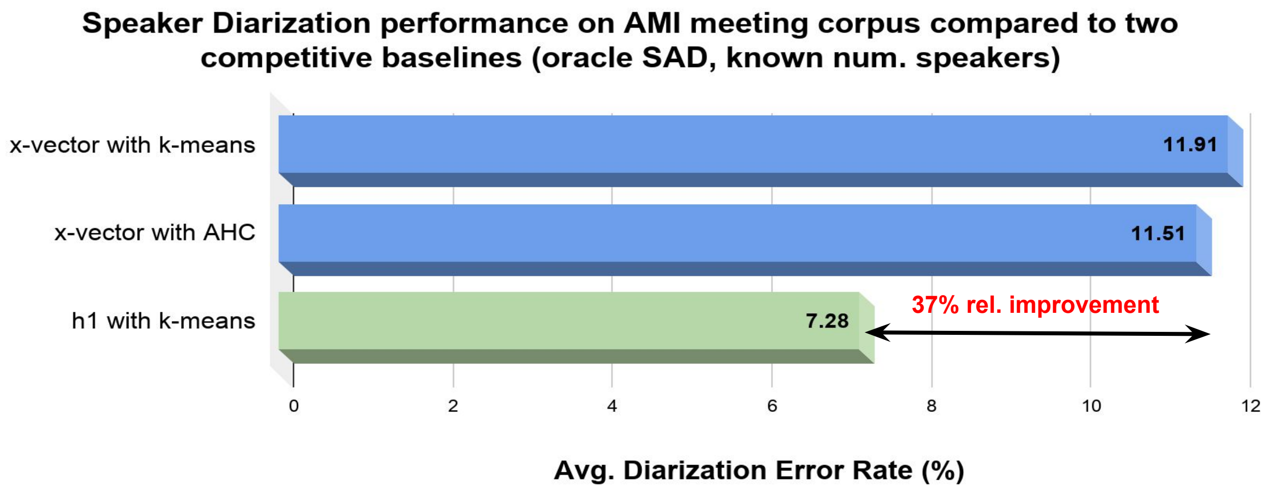
Speaker Diarization performance on AMI meeting corpus compared to two competitive baselines (oracle SAD, known num. speakers)



Conclusions

- Proposed novel speaker embeddings
  - Disentangled speaker and nuisance factors from speaker embeddings
  - No prior knowledge of specific nuisance factors during training
- Improves speaker verification performance in challenging conditions
  - Particularly babble noise (10% EER) and far-field recording conditions (15% EER)

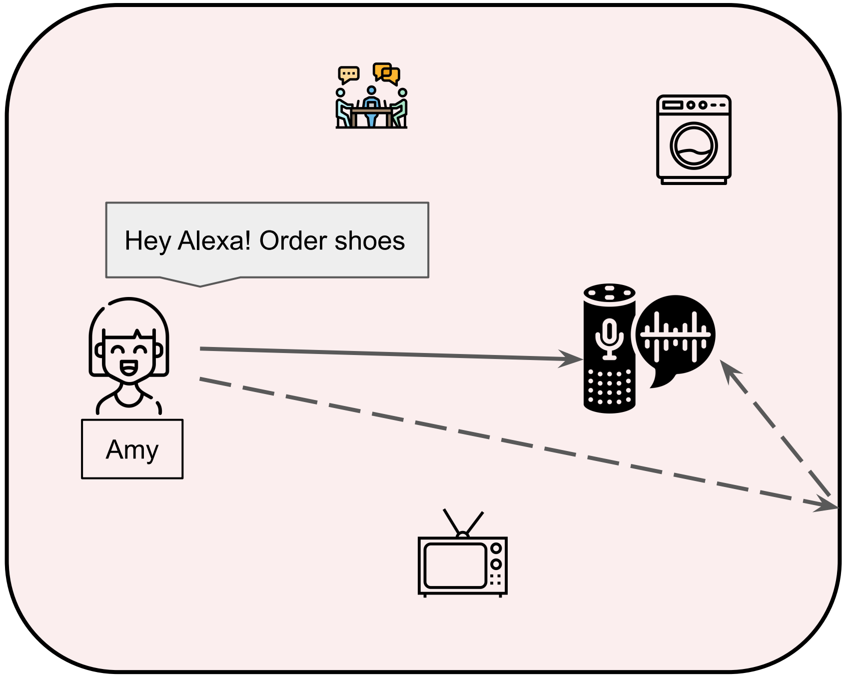
Conclusions



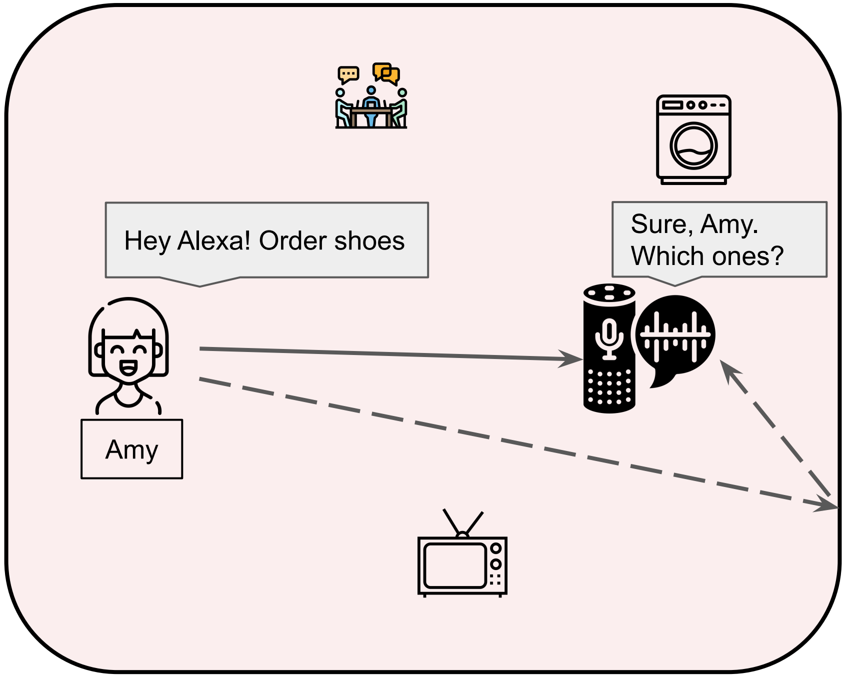
Conclusions

- Proposed novel speaker embeddings
  - Disentangled speaker and nuisance factors from speaker embeddings
  - No prior knowledge of specific nuisance factors during training
- Improves speaker verification performance in challenging conditions
  - Particularly babble noise (10% EER) and far-field recording conditions (15% EER)
- Improves speaker diarization performance on AMI meeting corpus (37% DER)

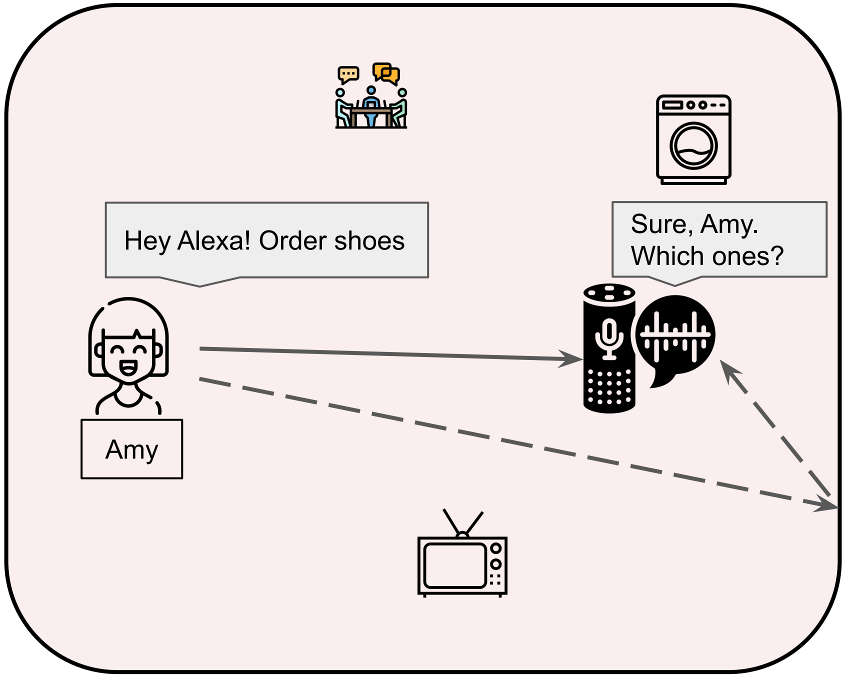
Conclusions



Conclusions



Conclusions



Future work

- Improve performance in babble noise scenario
- Evaluate disentangled speaker embeddings in presence of other nuisance factors, such as affective state, lexical content
- Train with more basic speech representations, which contain more variability useful for disentangling

We gratefully acknowledge the support of USG for this work

Contact Info
Name: Raghuveer Peri
Email: rperi@usc.edu