Generating and Protecting Against Adversarial Attacks for Deep Speech-based Emotion Recognition Models

Zhao Ren[1], Alice Baird[1], Jing Han[1], Zixing Zhang[2], Björn Schuller[1, 2]

[1]Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany
[2] GLAM – Group on Language, Audio & Music, Imperial College London, UK

Zhao Ren
07. 05. 2020

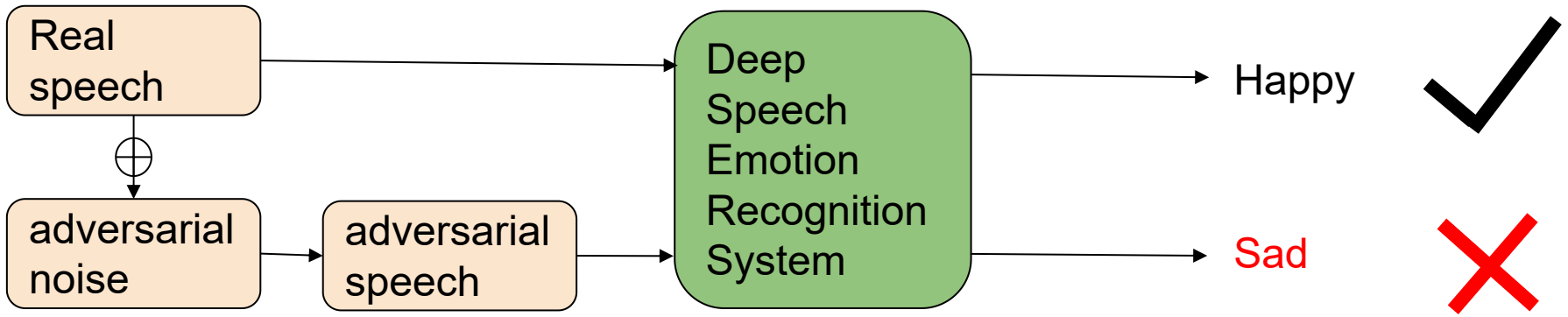
ICASSP2020 Barcelona

Outline

- Motivation
- Methodology
  - Adversarial-based Augmentation and Attacks
  - Adversarial-based Training for Defense
  - Deep Convolutional Neural Networks
- Experimental Results
- Conclusions and Future Work

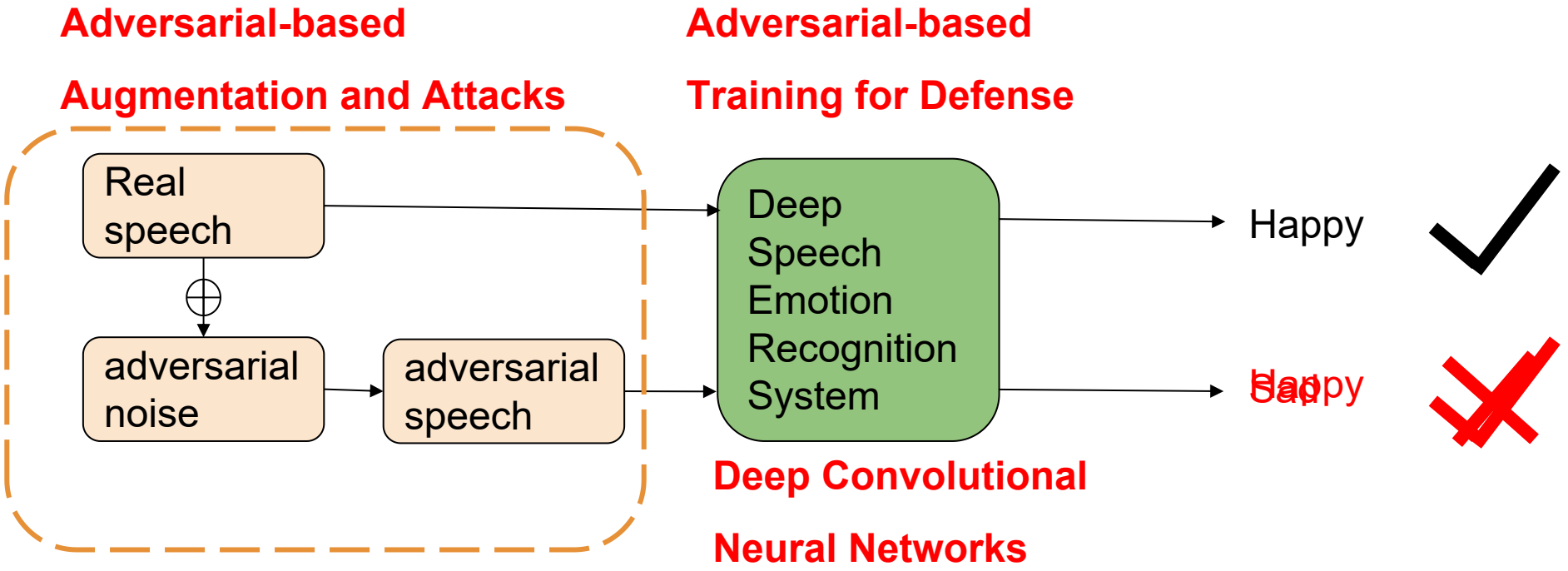
Motivation

- **Deep learning** based **Speech Emotion Recognition** has become a popular research topic.
- Real-world models are **vulnerable** to external attacks, particularly **adversarial attacks**.
  - destructive and misinterpreted interactions.



- **Improving the robustness of deep learning models** is now an important factor in AI research.

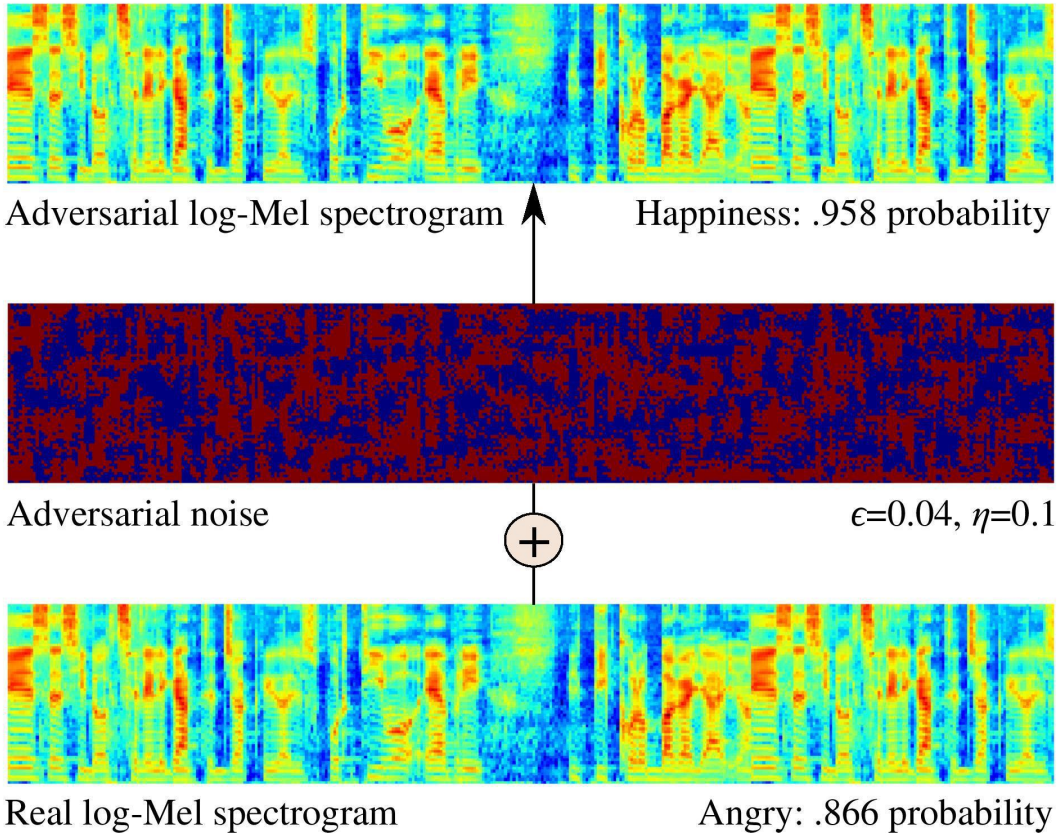
Methodology



Adversarial-based Augmentation and Attacks

An adversarial attack fools a neural network through adding **perturbations** to the data.

- help for **data augmentation** while training the models
- **attack a pre-trained model** during the validation



Fast Gradient Sign Method (FGSM)

Fake data: x'

Prediction: y'

$$\mathbf{x}' = \mathbf{x} + \epsilon \times \mathbf{sign}(\nabla_{\mathbf{x}} \mathbf{L}(\mathbf{w}, \mathbf{x}, \mathbf{y}))$$

$$\mathbf{x}' = \mathbf{clip}(\mathbf{x}', \mathbf{x} - \eta, \mathbf{x} + \eta)$$

Input: x

Target: y

Adversarial-based Training for Defense

Adversarial training aims to train on both real and fake data → more robust classification results.

- the training set is larger (a necessity for deep networks)
- the parameters are regularised against more fine-grained differences

Vanilla Adversarial Training

$$\hat{L}(\mathbf{w}, \mathbf{x}, \mathbf{y}) = \alpha * \mathbf{L}(\mathbf{w}, \mathbf{x}, \mathbf{y}) + (1 - \alpha) * \mathbf{L}(\mathbf{w}, \mathbf{x}', \mathbf{y})$$

Similarity-based Adversarial Training

$$\hat{L}(\mathbf{w}, \mathbf{x}, \mathbf{y}) = \beta \times \mathbf{L}(\mathbf{w}, \mathbf{x}, \mathbf{y}) + \gamma \times \mathbf{L}(\mathbf{w}, \mathbf{x}', \mathbf{y}) + (1 - \beta - \gamma) \times \|\mathbf{v} - \mathbf{v}'\|_{\mathbf{n}}$$

To defend the attacks, the feature vectors from the final fully connected layer should be close in each pair of real and fake data.

Deep Convolutional Neural Networks

CNN-4	VGG-16	ResNet-50
Input: log Mel spectrogram		
1×conv5-64	2×conv3-64	1×conv7-64, stride 2
maxpooling 2		maxpooling 3, stride 2
1×conv5-128	2×conv3-128	3× { conv1-64, conv3-64, conv1-256 }
maxpooling 2		
1×conv5-256	3×conv3-256	4× { conv1-128, conv3-128, conv1-512 }
maxpooling 2		
1×conv5-512	3×conv3-512	6× { conv1-256, conv3-256, conv1-1024 }
maxpooling 2		
-	3×conv3-512	3× { conv1-512, conv3-512, conv1-2048 }
-	maxpooling	average pooling
fully connected layer, softmax		

Database of Elicited Mood in Speech (DEMoS)

Partitioning of the data (train, development, and test) was made speaker-independently with consideration to gender and emotional class balancing.

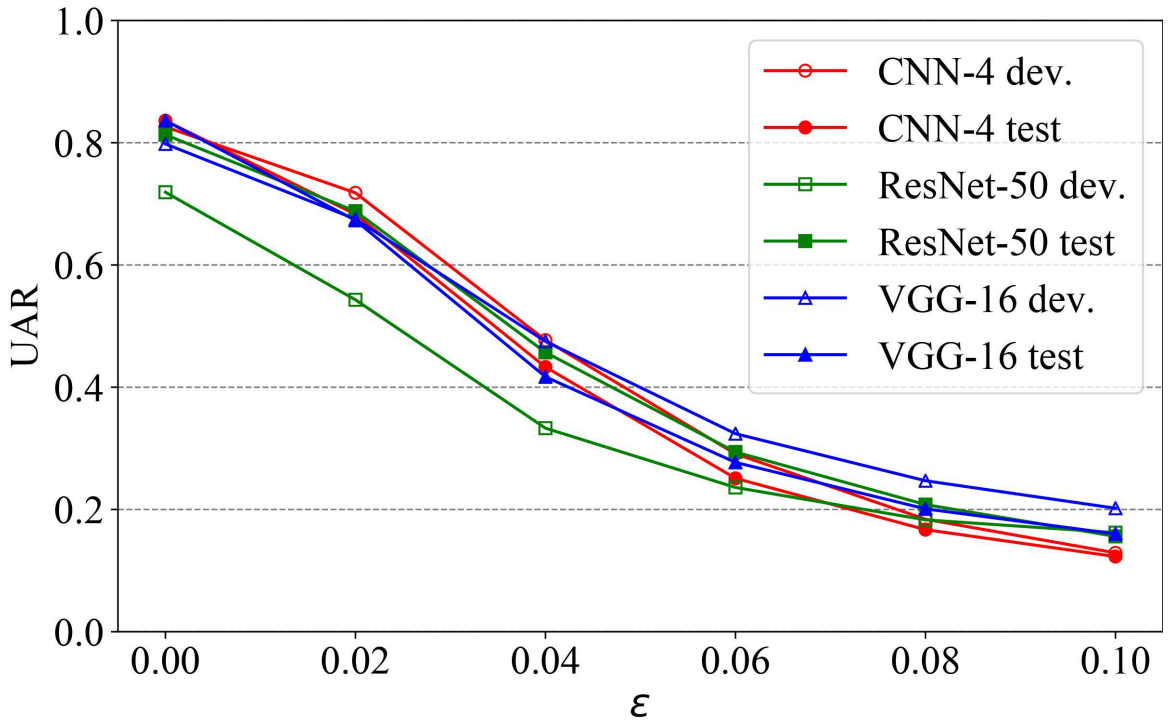
#	Train	Dev.	Test	**Sum**	Gender(F:M)
Speakers	24	22	22	68	23: 45
Anger	492	472	513	1477	516: 961
Disgust	525	556	597	1678	596:1082
Fear	380	383	393	1156	415: 741
Guilt	351	366	412	1129	400: 729
Happiness	447	434	514	1395	524: 871
Sadness	493	486	551	1530	532: 998
Surprise	336	327	337	1000	349: 651
Sum	3024	3024	3317	9365	3332:6033

Experimental Results

Test on the real data: $\epsilon = 0.00$

Test on the adversarial data: $\epsilon > 0.00$

- All of the six models perform well on the real data at around the UAR of 0.8.
- The UAR values are decreasing when the value of epsilon increases - attack successfully.



Experimental Results

On the real data, the performance are mostly improved because of data augmentation.
On the fake data, the two adversarial training approaches perform well.

		CNN-4				ResNet-50				VGG-16			
	UAR	Real		Fake		Real		Fake		Real		Fake	
NN	ϵ	*Dev.*	*Test*	*Dev.*	*Test*	*Dev.*	*Test*	*Dev.*	*Test*	*Dev.*	*Test*	*Dev.*	*Test*
Single Training	.00	.826	.836	–	–	.719	.813	–	–	.798	.836	–	–
Van. Adv. Training	.02	.825	.856	.744	.800	.699	.817	.620	.774	.850	.847	.794	.806
Van. Adv. Training	.04	.817	.871	.657	.749	.813	.850	.685	.755	.849	.855	.743	.783
Van. Adv. Training	.06	.854	.869	.576	.671	.774	.839	.595	.672	.871	.878	.741	.770
Van. Adv. Training	.08	.853	.858	.520	.540	.813	.855	.607	.710	**.875**	**.867**	.709	.756
Van. Adv. Training	.10	.866	.880	.457	.570	.823	.845	.602	.678	.842	.870	.638	.716
Sim. Adv. Training	.02	.844	.797	.827	.784	.743	.798	.732	.771	.847	.823	**.839**	**.821**
Sim. Adv. Training	.04	.822	.825	.772	.769	.708	.798	.652	.759	.814	.842	.806	.820
Sim. Adv. Training	.06	.775	.824	.675	.731	.723	.788	.630	.728	.786	.839	.753	.815
Sim. Adv. Training	.08	.739	.806	.610	.674	.631	.803	.450	.714	.792	.653	.730	.550
Sim. Adv. Training	.10	.727	.752	.526	.585	.407	.734	.316	.498	.804	.833	.716	.769

The performances are becoming worse on the adversarial data.

The vanilla adversarial training performs better on the real data than the similarity-based one mostly.
The similarity-based adversarial training can defend against attacks more effectively than the other.

Experimental Results

UAR	Dev.	Test
WaveNet (two classes)	.857	.741
Raw audio augmentation by random noise	.795	.833
Spectrogram augmentation by random noise	.808	.833
Our proposed approach	**.875**	**.867**

- WaveNet was applied only for two classes (happiness and sadness).
- Our training result performs significantly better than data (raw audio and log-Mel spectrogram image) augmentation methods using random noise.

Conclusions and Future work

Conclusions:

- We proposed a system for training a deep speech emotion recognition Convolutional Neural Network (CNN) model to be robust against adversarial attacks.
- We applied the vanilla and similarity-based adversarial training for defense to three deep CNN models, namely CNN-4, VGG-16, and ResNet-50.

Future Work:

- We will investigate generating black-box fake data for attacking deep learning models.
- Transferring the fake data across deep models will help to validate model robustness.
- To improve the performance when using the fake data, we look to train a detector for recognising this.