NTT

ICASSP 2021
 # 4128

Speech Emotion Recognition based on Listener Adaptive Models

Atsushi Ando[1], Ryo Masumura[1], Hiroshi Sato[1], Takafumi Moriya[1], Takanori Ashihara[1], Yusuke Ijima[1], Tomoki Toda[2]

[1] NTT Corporation, Japan [2] Nagoya University, Japan

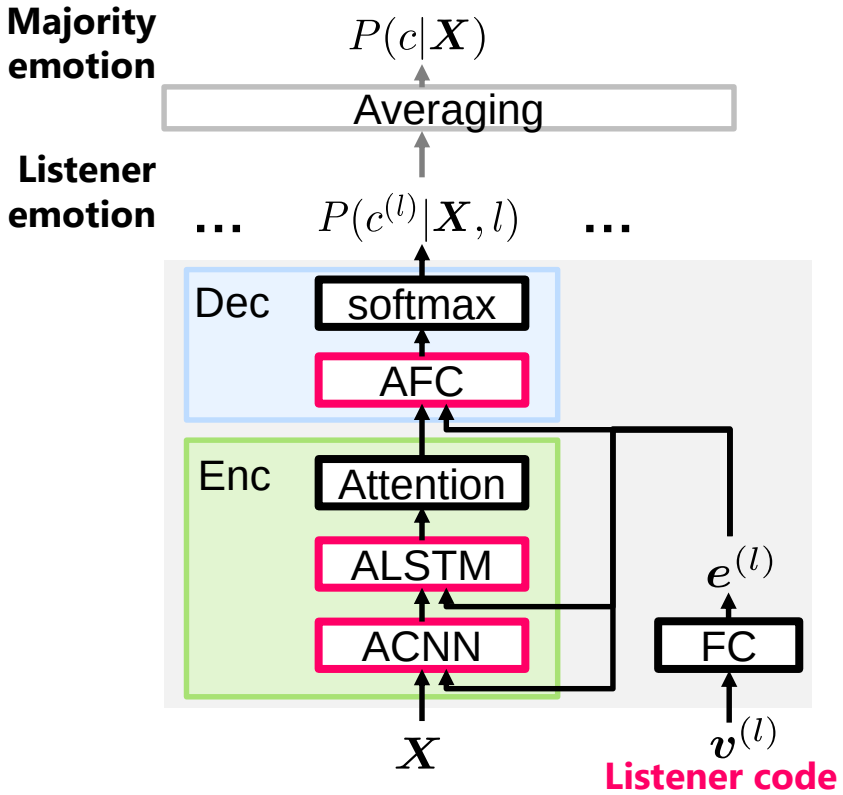
Quick Overview

Summary

- Conventional emotion recognitions estimate **the majority-voted emotion** of listeners
- Hypothesize **emotion perception is biased by each listener,** and propose a **Listener-Adaptive (LA) model** which **estimates listener-dependent perceived emotions**
  - Majority-voted emotions can also be estimated

Contributions

- Reveal that **listener-dependent perceptual biases exist** in natural speech
- The proposed LA model **significantly improves WAs with the same levels of UAs in both the majority-voted / the listener-dependent emotion recognition**



		Listener emo.		Majority emo.	
		WA	UA	WA	UA
Majority-voted model	single	41.0	40.8	42.9	46.4
	ens.	-	-	45.3	48.6
Soft-label model	single	45.5	**45.7**	49.3	49.4
	ens.	-	-	49.2	**50.2**
LA model w/ AFC layers		**63.2**	42.7	**59.7**	48.6

Background

- Speech Emotion Recognition (SER) is important to understand human communication

- Various SER methods have been developed
  - Heuristic feature-based
    - Utterance-level features + SVM/GMM [Luengo+,05][Rao+,13]
  - DNN-based
    - Low-level descriptors + RNN-Attention [Mirsamadi+,17]
    - Raw waveform + TDNN-RNN-Attention [Sarma+,18]
    - **Spectrogram + CNN-RNN-Attention [Tzirakis+,18][Li+,19]**

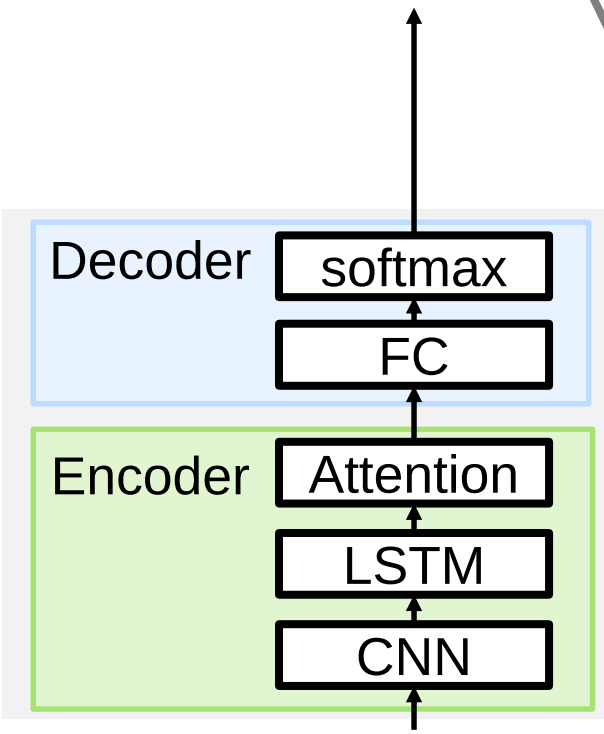
Conventional SER

- Estimate **the majority-voted emotion** perceived by multiple listeners

Posteriors of majority-voted emo.

$P(c|\boldsymbol{X})$

SER model



Acoustic features
(e.g. spectrogram)

$\boldsymbol{X}$

	Majority	**Listener perceptions**				
	c	$c^{(1)}$	$c^{(2)}$	$c^{(3)}$	$c^{(4)}$	…
	*Neu*	*Neu*	*Hap*	*Neu*		
	*Ang*		*Ang*	*Ang*	*Ang*	
	*Hap*	*Neu*	*Hap*	*Hap*		
…	…					…

Problem

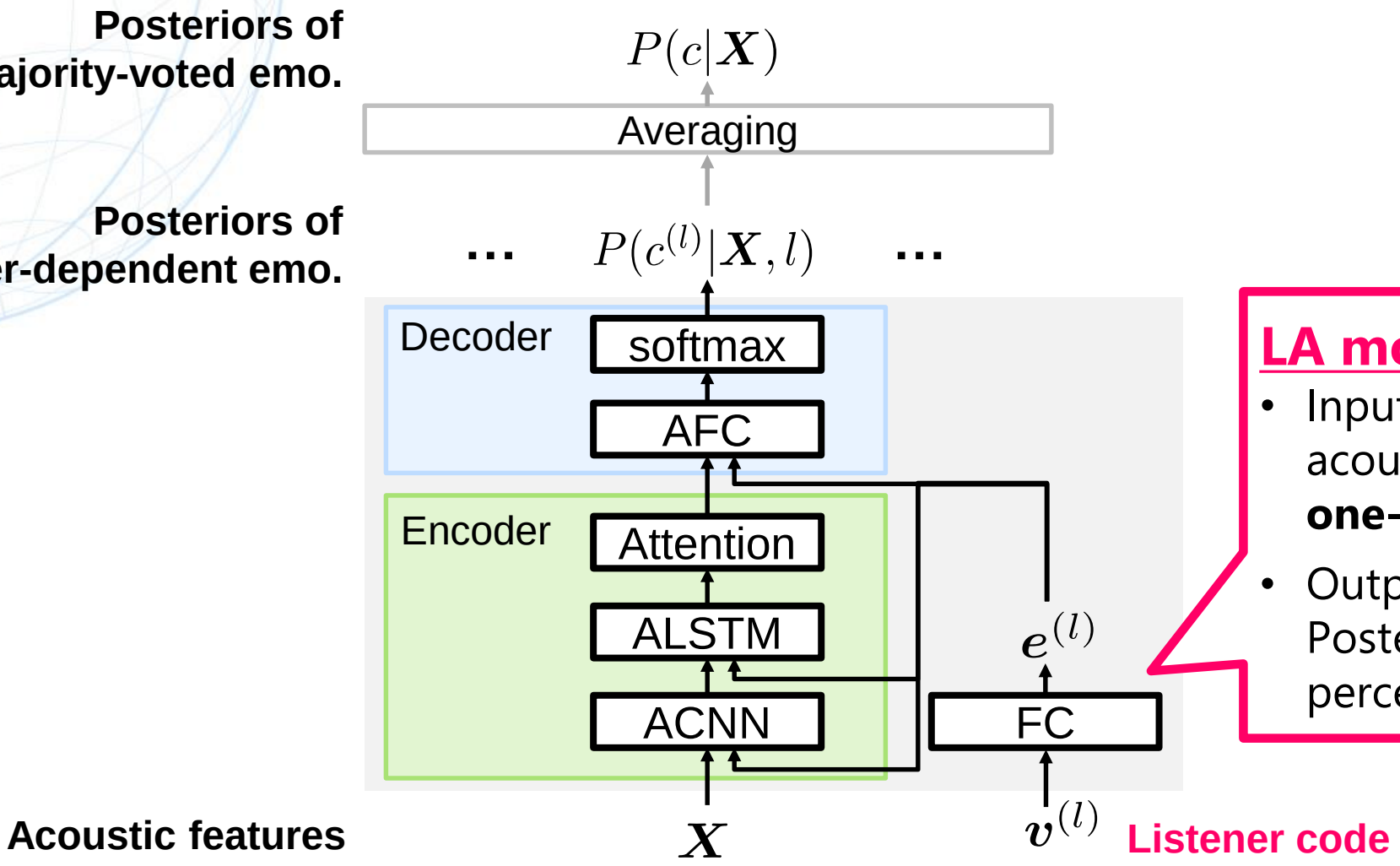
- **Emotion perceptions may be biased by individual listeners**
  - Emotion perception depends on listener's age, gender, and cultures [Dang+,10][Zhao+,19]
  - Majority-voted emotion is usually determined by different sets of listeners for each utterance

	Majority	Listener perceptions				
	c	$c^{(1)}$	$c^{(2)}$	$c^{(3)}$	$c^{(4)}$	…
	*Neu*	*Neu*	*Hap*	*Neu*		
	*Ang*		*Ang*	*Ang*	*Ang*	
	*Hap*	*Neu*	*Hap*	*Hap*		
…	…					…
		mostly *Neu*		no bias		

- Research questions:
  1. **Are there any perceptual biases between listeners?**
  2. **Is it better to make listener-dependent models than the majority-voted model?**

Proposed: Listener-Adaptive (LA) model

- Adapts to each listener **by listener code + adaptation layers**

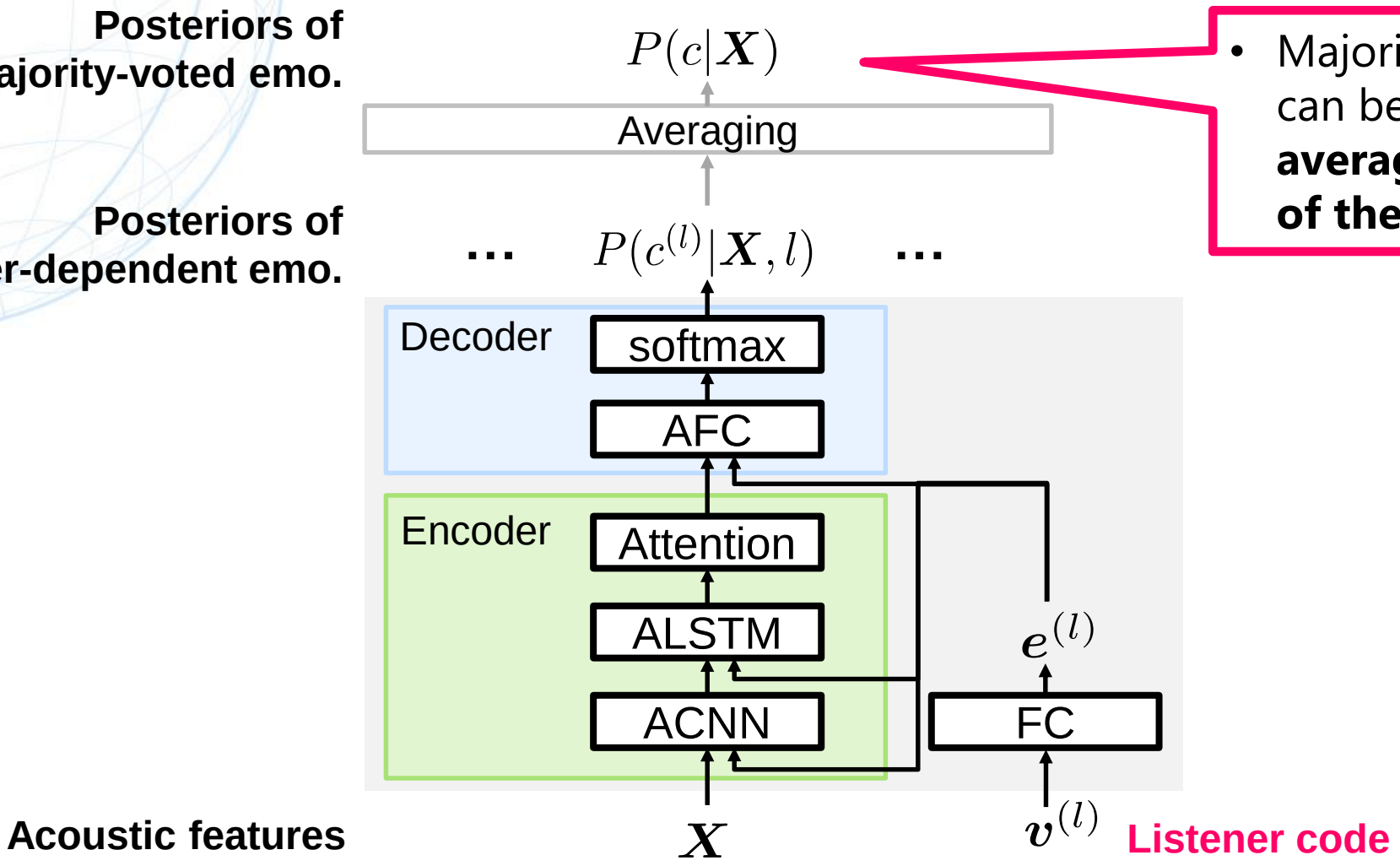


LA model

- Input:
  acoustic features + **one-hot listener code**
- Output:
  Posteriors of listener l perceived emotions

Proposed: Listener-Adaptive (LA) model

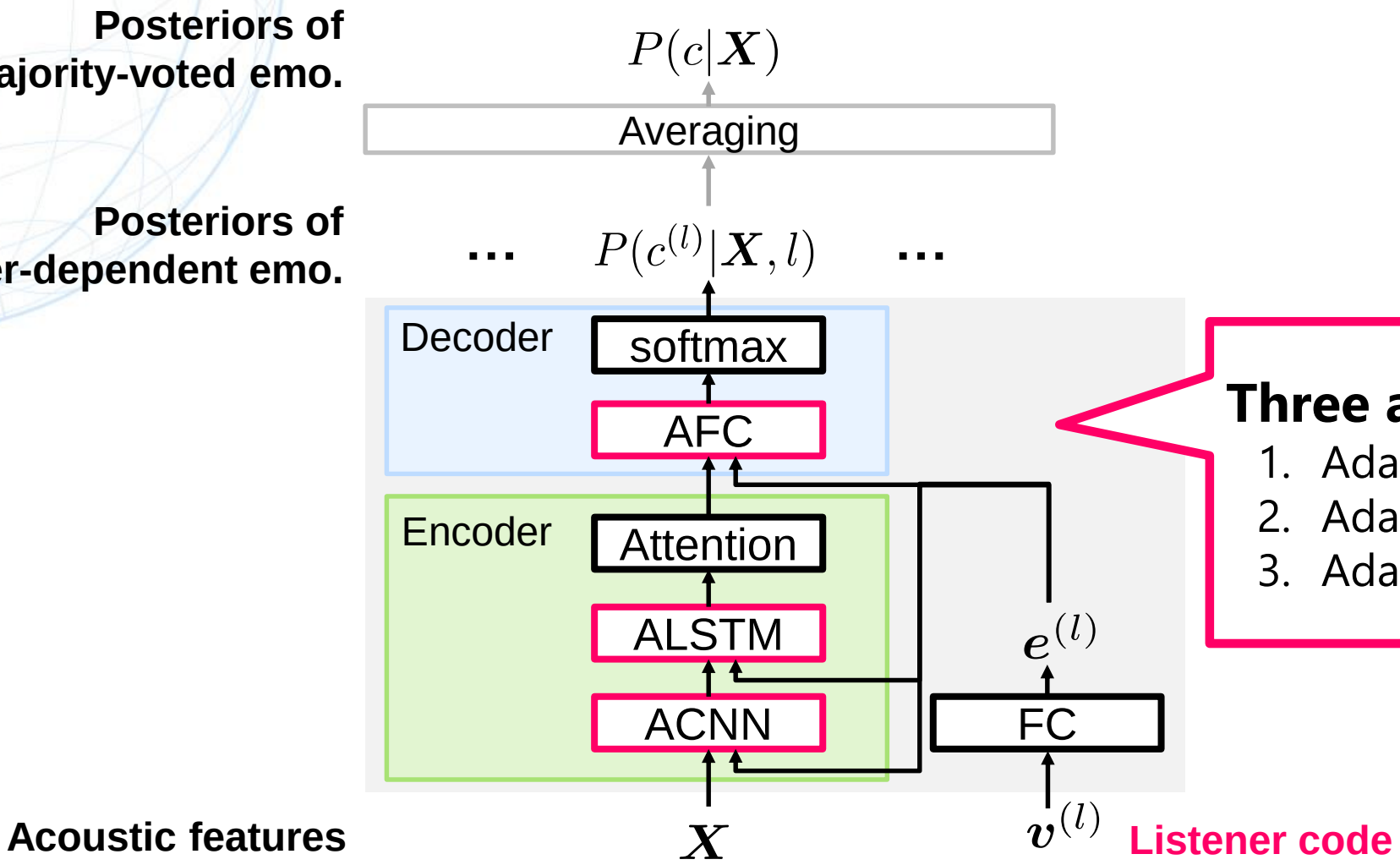
- Adapts to each listener **by listener code + adaptation layers**



- Majority-voted emotion can be estimated **by averaging the outputs of the LA model**

Proposed: Listener-Adaptive (LA) model

- Adapts to each listener **by listener code + adaptation layers**



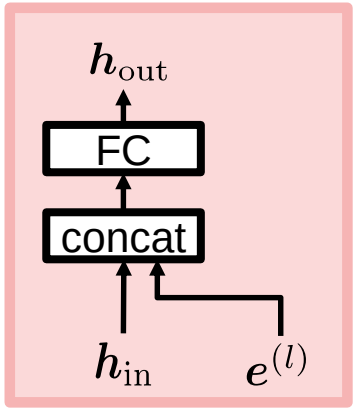
Three adaptation layers

1. Adaptive FC (AFC)
2. Adaptive LSTM (ALSTM)
3. Adaptive CNN (ACNN)

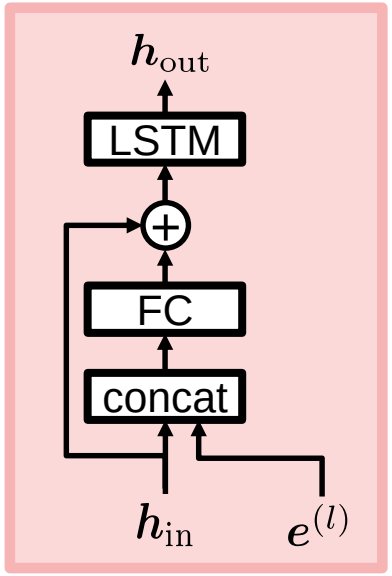
Proposed: Listener-Adaptive (LA) model

- Three types of adaptation layers
  1. **Adaptive FC (AFC)**
     - Concatenate the input and the auxiliary vector
     - Used in speech recognition / synthesis
  2. **Adaptive LSTM (ALSTM)**
     - Transform the LSTM input [Miao+,15]
     - Used in speech recognition
  3. **Adaptive CNN (ACNN)**
     - CNN filter parameters are determined by the auxiliary vector [Kang+,17]
     - Used in object detection

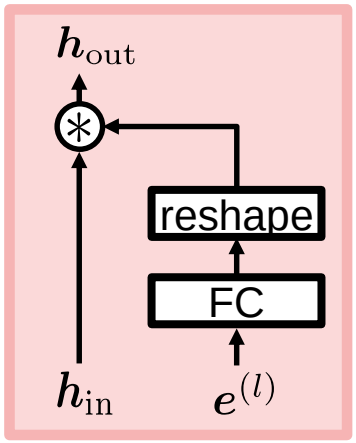
AFC



ALSTM



ACNN



Experiments

- Task: 4-class emotion classification (*Neu, Hap, Sad, Ang*)
  - Estimation target: (1)Majority-voted emo., (2)Listener-dependent emo.
- Dataset: MSP-Podcast, IEMOCAP

	MSP-Podcast [Reza+,17]	IEMOCAP [Busso+,08]
Emo type	Natural	Acted
Task	Podcast clip	two-actor dialog
# utts / spks	40227 utts / 1000~ spks	2943 utts / 10 spks
# listeners	154 + *rest* (orig: 11010)	3 + *rest* (orig: 6)

MSP-Podcast

		Neu	*Hap*	*Sad*	*Ang*
Majority		22681	12302	2351	2893
Listener	1	5475	380	27	59
	2	1130	1026	120	69
	3	421	1072	191	128

IEMOCAP

		Neu	*Hap*	*Sad*	*Ang*
Majority		1099	947	608	289
Listener	1	412	1166	589	284
	2	951	876	586	269
	3	1225	717	324	155

Experiments

- Analysis: Kappa coefficients of listener-wise perceived emotions

MSP-Podcast

	lis1	lis2	lis3	lis4	lis5	lis6	lis7	lis8	lis9	lis10
lis1	1.00	0.14	0.11	0.52	0.15	0.12	0.23	0.15	0.60	0.50
lis2	0.14	1.00	0.38	-	-	0.37	0.22	0.32	0.21	0.15
lis3	0.11	0.38	1.00	-	-	0.41	0.07	0.58	0.21	0.15
lis4	0.52	-	-	1.00	0.27	0.60	-	-	-	-
lis5	0.15	-	-	0.27	1.00	0.64	-	-	-	-
lis6	0.12	0.37	0.41	0.60	0.64	1.00	0.12	0.45	0.11	0.14
lis7	0.23	0.22	0.07	-	-	0.12	1.00	0.16	0.24	-
lis8	0.15	0.32	0.58	-	-	0.45	0.16	1.00	0.28	0.08
lis9	0.60	0.21	0.21	-	-	0.11	0.24	0.28	1.00	-
lis10	0.50	0.15	0.15	-	-	0.14	-	0.08	-	1.00

Only several pairs are > 0.4

IEMOCAP

	lis1	lis2	lis3
lis1	1.00	0.58	0.47
lis2	0.58	1.00	0.69
lis3	0.47	0.69	1.00

All > 0.4

There are listener-dependent perceptual biases in MSP-Podcast!

Experiments

- Setups
  - Comparisons
    - Baseline: Majority-voted emotion model, Soft-target model (single, ensembled)
    - Proposed: Listener-Adaptive model with AFC, ALSTM, ACNN layers
  - Input: 400-dim log power spectrogram
  - Structure: CNN 3layer - BLSTM 128*1layer - SelfAtt 4head - FC 64*2layer
  - Training: Adam w/ Earlystop, SpeedPerturb, SpecAugment [Park+,19]

- Metrics
  - **Majority-voted emotion recognition**: WA, UA
  - **Listener-dependent emotion recognition**: Macro-avg. of listener-wise WA, UA

Results

- LA model **significantly improved majority-/listener-emotion WAs with equivalent UAs** in MSP-Podcast

		MSP-Podcast				IEMOCAP			
		Majority emo.		Listener emo.		Majority emo.		Listener emo.	
		WA	UA	WA	UA	WA	UA	WA	UA
Majority-voted	single	42.9	46.4	41.0	40.8	60.2	63.2	58.3	62.7
	ens.	45.3	48.6	-	-	61.5	**64.9**	-	-
Soft label	single	49.3	49.4	45.5	**45.7**	60.3	62.8	58.2	61.7
	ens.	49.2	**50.2**	-	-	**61.8**	64.6	-	-
Listener Adaptive (LA) model	AFC	**59.7**	48.6	**63.2**	42.7	61.6	**64.9**	**60.1**	**65.9**
	ALSTM	55.5	45.3	59.5	39.3	58.3	61.5	56.3	61.9
	ACNN	54.2	38.0	58.9	37.0	58.6	62.5	57.0	62.8
	AFC +ALSTM +ACNN	57.9	34.3	60.5	36.2	60.4	62.6	58.7	63.3

Results

- LA model **significantly improved majority-/listener-emotion WAs with equivalent UAs** in MSP-Podcast

		MSP-Podcast				IEMOCAP			
		Majority emo.		Listener emo.		Majority emo.		Listener emo.	
		WA	UA	WA	UA	WA	UA	WA	UA
Majority-voted	single	42.9	46.4	41.0	40.8	60.2	63.2	58.3	62.7
	ens.	45.3	48.6	-	-	61.5	**64.9**	-	-
Soft label	single	49.3	49.4	45.5	**45.7**	60			
	ens.	49.2	**50.2**	-	-	**61.**			
Listener Adaptive (LA) model	AFC	**59.7**	48.6	**63.2**	42.7	61.			
	ALSTM	55.5	45.3	59.5	39.3	58.3	61.5	56.3	61.9
	ACNN	54.2	38.0	58.9	37.0	58.6	62.5	57.0	62.8
	AFC +ALSTM +ACNN	57.9	34.3	60.5	36.2	60.4	62.6	58.7	63.3

Significantly improve WA with equivalent UA

($p < .05$ in paired t-test)

Results

- LA model **significantly improved majority-/listener-emotion WAs with equivalent UAs** in MSP-Podcast

		MSP-Podcast				IEMOCAP			
		Majority emo.		Listener emo.		Majority emo.		Listener emo.	
		WA	UA	WA	UA	WA	UA	WA	UA
Majority-voted	single	42.9	46.4	41.0	40.8	60.2	63.2	58.3	62.7
	ens.				-	61.5	**64.9**	-	-
Soft label	single				**5.7**	60.3	62.8	58.2	61.7
	ens.	49.2	**50.2**	-	-	**61.8**	64.6	-	-
Listener Adaptive (LA) model	AFC	**59.7**	48.6	**63.2**	42.7	61.6	**64.9**	**60.1**	**65.9**
	ALSTM	55.5	45.3	59.5	39.3	58.3	61.5	56.3	61.9
	ACNN	54.2	38.0	58.9	37.0	58.6	62.5	57.0	62.8
	AFC +ALSTM +ACNN	57.9	34.3	60.5	36.2	60.4	62.6	58.7	63.3

Almost the same WA/UA
(p > .05 in paired t-test)

Results

- LA model **significantly improved majority-/listener-emotion WAs with equivalent UAs** in MSP-Podcast

		MSP-Podcast				IEMOCAP			
		Majority emo.		Listener emo.		Majority emo.		Listener emo.	
		WA	UA	WA	UA	WA	UA	WA	UA
Majority-voted	single	42							62.7
	ens.	45							-
Soft label	single	49							61.7
	ens.	49.2	50.2		-	61.8	64.0	-	-
Listener Adaptive (LA) model	AFC	**59.7**	48.6	**63.2**	42.7	61.6	**64.9**	**60.1**	**65.9**
	ALSTM	55.5	45.3	59.5	39.3	58.3	61.5	56.3	61.9
	ACNN	54.2	38.0	58.9	37.0	58.6	62.5	57.0	62.8
	AFC +ALSTM +ACNN	57.9	34.3	60.5	36.2	60.4	62.6	58.7	63.3

AFC (decoder) is effective, ALSTM/ACNN (encoder) is not

↓

Listener-dependency may appear in the decision-making, not in the feature extraction?

Discussions: Confusion Matrix

- Recalls of ***Neu* improved**, while ***Sad, Ang* not improved**
  - Majority-voted emotion recognition (MSP-Podcast)

Majority (ens)		Pred.			
		Neu	*Hap*	*Sad*	*Ang*
Actu.	*Neu*	**2635**	882	1402	603
	Hap	869	**1741**	476	672
	Sad	211	50	**229**	46
	Ang	98	142	52	**369**

Soft-label (ens)		Pred.			
		Neu	*Hap*	*Sad*	*Ang*
Actu.	*Neu*	**2464**	1181	1191	686
	Hap	761	**2073**	300	624
	Sad	181	76	**222**	57
	Ang	81	148	37	**395**

LA model		Pred.			
		Neu	*Hap*	*Sad*	*Ang*
Actu.	*Neu*	**3765**	879	568	310
	Hap	1211	**2056**	181	310
	Sad	274	69	**175**	18
	Ang	162	207	35	**257**

  - Listener-dependent emotion recognition (MSP-Podcast)

Listener 1

Actu.	*Neu*	**1245**
	Hap	123
	Sad	4
	Ang	16

Soft-label		Pred.			
		Neu	*Hap*	*Sad*	*Ang*
Actu.	*Neu*	**365**	477	167	236
	Hap	22	**66**	10	25
	Sad	1	0	**1**	2
	Ang	3	4	1	**8**

LA model		Pred.			
		Neu	*Hap*	*Sad*	*Ang*
Actu.	*Neu*	**707**	399	1	138
	Hap	48	**64**	0	11
	Sad	3	0	**0**	1
	Ang	6	3	0	**7**

LA model may be affected by the data imbalance of each listener

Conclusion

- Summary
  - Hypothesized **emotion perception may be biased by individual listeners**
  - Proposed a **Listener-Adaptive (LA) model** that can estimate listener-dependent emotion perception results
    - **Adaptation by auxiliary input** of 1-hot listener-code
    - Three adaptation layers: AFC, ALSTM, ACNN
    - Majority-voted emotion can also be estimated by averaging of LA model outputs
  - Experimental results showed:
    - **Emotion perceptions are biased by listeners** in natural speech
    - **The proposed LA model significantly improved WAs** in both the majority-voted / the listener-dependent emotion recognition
- Future work
  - Adapt the LA model to **unseen listeners** (listeners not in the training set)
  - Improve robustness in data imbalance of each listener