MULTI-HEAD ATTENTION FOR SPEECH EMOTION RECOGNITION WITH AUXILIARY LEARNING OF GENDER RECOGNITION

Anish Nediyanchath, Periyasamy Paramasivam, Promod Yenigalla
{anish.n, periyasamy.p, promod.y}@samsung.com
Samsung R&D Institute India - Bangalore, India

ICASSP 2020

Agenda

- Motivation & Proposed Method
- Raw Audio Processing
- Multi Head Attention Network For SER
- Multi Head Attention Network For SER + MTL (Gender)
- Network Parameters
- Dataset
- Results Comparison
- Conclusion

Motivation & Proposed Method

Motivation

- Voice assistants are becoming ubiquitous
- Emotional understanding of users makes for better companions
- Humans always carries emotion
- We express our emotions through
  - What we speak (extrinsic)
  - Voice
  - Expressions (Face, Gesture, Posture etc.)
- Targeted for products like chat-bots, voice assistants and social robot

The effect of emotions on the human voice

	fear	anger	sorrow	joy	disgust	surprise
speech rate	much faster	slightly faster	slightly slower	faster or slower	very much slower	much faster
pitch average	very much higher	very much higher	slightly lower	much higher	very much lower	much higher
pitch range	much wider	much wider	slightly narrower	much wider	slightly wider	
intensity	normal	higher	lower	higher	lower	higher
voice quality	irregular voicing	breathy chest tone	resonant	breathy blaring	grumbled chest tone	
pitch changes	normal	abrupt on stressed syllable	downward inflections	smooth upward inflections	wide downward terminal inflections	rising contour
articulation	precise	tense	slurring	normal	normal	

Proposed Method

- Tranformer Encoder based MHA network for SER with LFBE fearure input & Position Embedding
- Addition of MTL on MHA network SER with gender prediction as auxiliary task

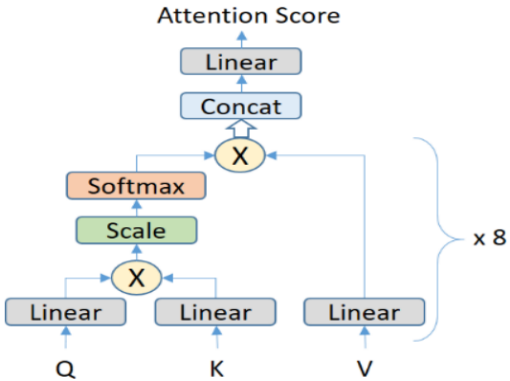
Raw Audio Preprocessing

- 16KHz, 16-bit PCM format audio of 6 seconds input length
- Divide in 46 ms frames with 23 ms stride
- Calculate Log Mel-Filter Bank Energies for each frame, 64 filter banks
- Final feature matrix for training model will be of size 260 * 64
- Scipy wavfile, python speech features library

Multi Head Attention Network For SER

Attention

- An **attention** mechanism allows neural network to focus parts on input relevant to given context
- Attention parameters are a projection of Query (Q) on Key(K)-Value(V) pair vectors.
- Self attention is a variant of attention mechanism where all Q, K, V are from the same input vector
- Self attention networks replaced sequence based methods like RNN, LSTM and GRU's
- We use Multi-Head Attention (MHA), which divides each vector into *n*
- For audio sequences, neighboring frames will carry similar acoustic characteristics.
- MHA allows model to relate to other parts of the sequence as well if similar characteristics appear



$$\left.\begin{aligned} Q_i &= X * W_i^Q \\ K_i &= X * W_i^K \\ V_i &= X * W_i^V \\ H_i &= Softmax\left(\frac{Q_i K_i^T}{\sqrt{h}}\right) * V_i \end{aligned}\right\} \forall\, i$$

$$MHA = Concat(H_1, H_2 \dots H_n) * W$$

Multi Head Attention Network For SER

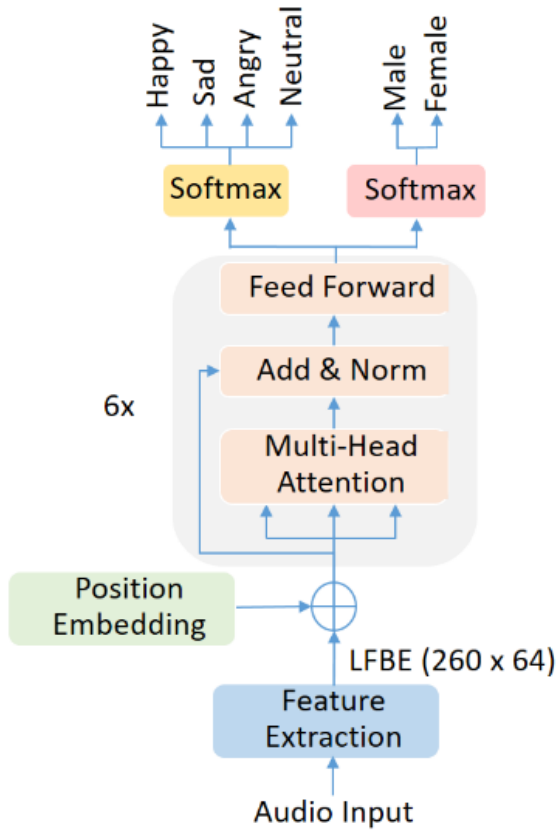
Position Embedding

- MHA identifies acoustic events relevant for emotion, but not the sequence of events itself
- Position encoding is used in tranformer networks to capture sequence information
- Position encoding is fixed positional representation of input features
- BERT extends the idea of position encoding making position information learnable
- We follow the BERT approach by adding a learned position information to input LFBE features
- Position embedding vectors are initialized with random weights and learned as part of training

Multi Head Attention Network For SER + MTL (Gender)

Multi Task Learning

- MTL allows better genarlization by learning to ignore task specific noise patterns
- Contrast between relevant v/s irrelevant features
- Sharing features relevant for different tasks
- Other researches have demonstrated that knowing gender can improve SER
- We introduced gender recognition as auxiliary task to improve SER
- Adding auxiliary MTL task make MHA learn multiple representations relevant in granular space common to both tasks
- Position Embedding and MHA layers are shared by both tasks
- Two independent softmax activations are used for emotion and gender classification



Network Parameters

Attribute	Value
LFBE Frames	260
LFBE Features	64
Input Sequence Length	260
Batch Size	16
Number of MHA Layers	6
MHA Heads	8
MHA Dropout	0.1
Feed Forward Layers	6
Feed Forward Size	256
MHA Activation	gelu
Output Activation	softmax
Learning Rate	0.001
Optimizer	Adam

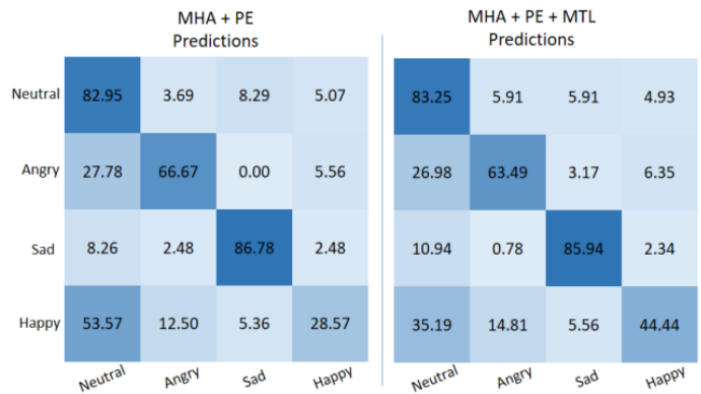
Dataset

- IEMOCAP (Interactive Emotional Motion Capture) from University Of Southern California
- It is a multimodal dataset having around 12 hours of audiovisual data
- Consists of five diadic sessions where actors perform improvisations or scriped scenarios to represent emotional experssions
- We use only improvised raw audio samples as it has strong correlation to labelled emotions and are close to natural speech
- Four emotions of Neutral, Happy, Sad and Angry are explored
- IEMOCAP class distributions are skewed – Neutral 49%, Happy 12%, Sad 27%, Angry 12%
- Because of data imbalance, we report both Weighted Average (WA) and Unweighted Average (UA) on test data.
- Dataset is split into 80:20 train and test split for all experiments
- We use five fold cross validation on this dataset for all the reported results

Results Comparison

- Results from experiments are compared using IEMOCAP dataset are presented
- Comparison is performed against state-of-the-art models using four emotion classes – neutral, angry, happy and sad
- MHA attention model increases the overall accuracy by 3% compared to state-of-the-art
- Adding poisition emebedding improved results over MHA model by 0.6%, average class accuracy by 2%
- MTL network with MHA and PE gives overall accuracy of 76.4, 5.3% higher than state-of-the-art by 5.3%
- The average class accuracy is 70.1, and improvement of 6.2%

Methods	Overall Accuracy (WA)	Class Accuracy (UA)
Lee [18] (Bi-LSTM)	62.8	63.9
Satt [19] (CNN + LSTM)	68.8	59.4
Ramet [20] (Attn. Bi-LSTM)	68.8	63.7
Zhang [21] (Attn. CNN)	70.4	63.9
Yenigalla [22] (CNN)	71.3	61.9
MHA	74.1	64.2
MHA + PE	74.7	66.2
MHA + PE + MTL	**76.4**	**70.1**



Conclusion

- This work demonstrated MHA network for SER
- We also employed Multi Task Learning with auxiliary task of gender recognition
- Using self attention to attend to different sections of speech features can improve SER accuracy
- Demonstrated an overall improvement of accuracy to 76.4%
- We believe our work as a step forward for SER towards building conversation systems with better emotion recognition capabilities
- In future, we plan to extend this work by bringing in speaker identification to MTL, add more emotions and explore noisy speech

Thank You