On the Correlation and Transferability of Features between Automatic Speech Recognition and Speech Emotion Recognition

Haytham M. Fayek[1] | Margaret Lech[1] | Lawrence Cavedon[2]

[1]School of Engineering | [2]School of Science
RMIT University | Melbourne | Australia

haytham.fayek@ieee.org

INTERSPEECH
12 September 2016

Outline

Outline

Background

- The relationship between Automatic Speech Recognition (ASR) and Speech Emotion Recognition (SER) is ill-defined.

- Acoustic models in ASR utilise a few frames to recognise phonemes that are later decoded into a transcription.

  Acoustic models in SER require a larger number of frames to recognise emotions.

- Most work in ASR considers the presence of paralinguistics (e.g. Emotions) in speech a form of distortion.

  Improvement was reported in SER in the presence of a linguistic input.

Hybrid ASR-SER

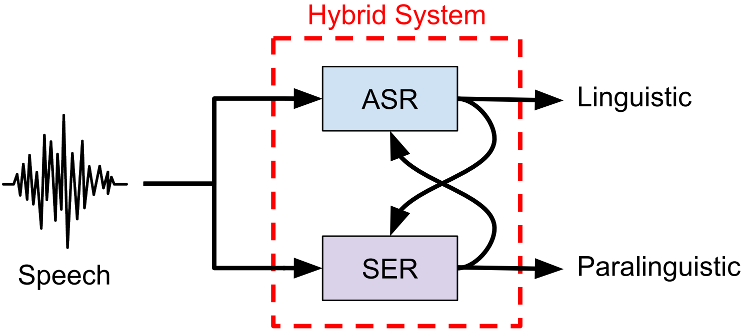

Figure: A Hybrid ASR-SER System.

Relation Between ASR and SER

- Deep learning is the state-of-the-art approach for ASR and SER.
- The relation between ASR and SER must be studied prior.

Relation Between ASR and SER

- Deep learning is the state-of-the-art approach for ASR and SER.
- The relation between ASR and SER must be studied prior.

We can study the relation between ASR and SER by studying the relation and relevance of the features learned in both tasks using transfer learning.

Related Work

- Deep neural networks tend to learn low-level features in initial layers and transition to high-level features in final layers.
- Yosinski *et al.* (2014) showed on a computer vision task that there is a correlation between the benefit of feature transfer and the distance between both tasks.
- Transfer Learning has been used in:
  - ASR: cross-language, speaker adaptation, etc.
  - SER: cross-corpus, music, etc.

Transfer Learning: ASR $\leftrightarrow$ SER

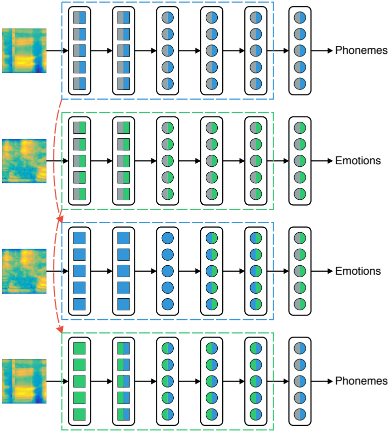

Figure: Transfer Learning between ASR and SER.

Outline

Data

- ASR Data:

  TIMIT: 630 speakers from 8 major american english dialects.

  - Training set: Complete set of 462 speakers without SA utterances.
  - Development Set: 50-speaker set.
  - Test Set: Core set of 24 speakers.

- SER Data:

  IEMOCAP: 10 speakers producing 12 hours of audiovisual recordings.

  - Classes: anger, happiness + excitement, neutral, sadness.
  - 8-Fold Leave-One-Speaker-Out (LOSO) cross-validation.
  - 2 Speakers left out as a validation set.

Preprocessing

- Preprocessing:
  - Speech analysed 25ms Hamming window with a stride of 10ms.
  - 40-coefficient Log Mel-scale Fourier-transform based filter banks.
  - Speaker-independent mean and variance normalisation with training subset.
- ASR Labels:
  - Force-aligned labels were obtained with a GMM-HMM system with MFCCs using the standard Kaldi recipe.
- SER Labels:
  - Frame labels were inherited from the parent utterance labels.
  - A VAD was then used to label silent and unvoiced frames and a *Silence* label was added as an extra class.

ConvNet Acoustic Model

Table: *Convolutional Neural Network Architecture.*

No.	Type	Size	Other
1	Convolution	$64, 5 \times 4$	$l_2 = 1 \times 10^{-3}$
	BatchNorm	-	-
	ReLU	-	-
	Max Pooling	2×2	Stride = 2
2	Convolution	$128, 3 \times 3$	$l_2 = 1 \times 10^{-3}$
	BatchNorm	-	-
	ReLU	-	-
	Max Pooling	2×2	Stride = 2
3	Fully-Connected	1024	-
	BatchNorm	-	-
	ReLU	-	-
	Dropout	-	$Dropout = 0.6$
4	Fully-Connected	1024	-
	Batch Norm	-	-
	ReLU	-	-
	Dropout	-	$Dropout = 0.6$
5	Fully-Connected	1024	-
	BatchNorm	-	-
	ReLU	-	-
	Dropout	-	$Dropout = 0.6$
6	Fully-Connected	$144/5$	-
	Softmax	-	-

System Architecture and Training

- ASR System:

  31 Frames + ConvNet Acoustic Model + 3-State HMM Bi-Gram LM.

- SER System:

  31 Frames + ConvNet Acoustic Model.

- Training:

  - Parameters were initialised from a Gaussian distribution with zero mean and $\sqrt{2/n}$ standard deviation.
  - Mini-batch SGD and RMSProp with respect to a CE cost function.
  - Validation set was used for early stopping.
  - Trained on a cluster of Tesla K40 GPUs.

Transfer Learning: ASR $\leftrightarrow$ SER

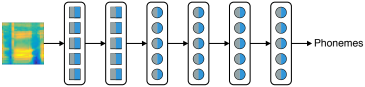

Figure: Transfer Learning between ASR and SER.

Transfer Learning: ASR $\leftrightarrow$ SER

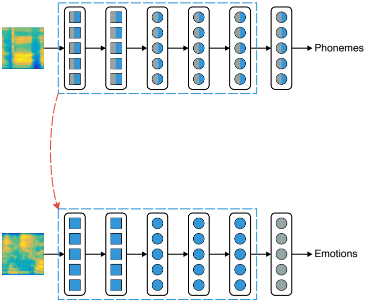

Figure: Transfer Learning between ASR and SER.

Transfer Learning: ASR $\leftrightarrow$ SER

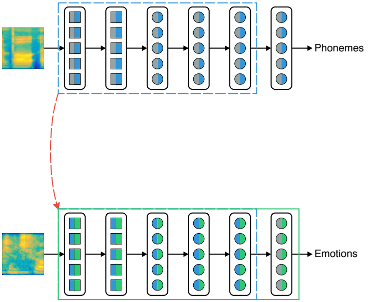

Figure: Transfer Learning between ASR and SER.

Transfer Learning: ASR $\leftrightarrow$ SER

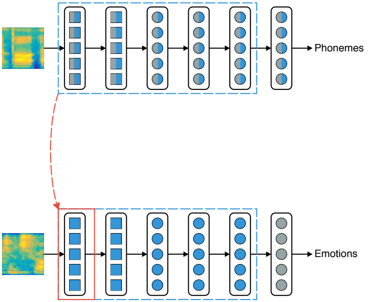

Figure: Transfer Learning between ASR and SER.

Transfer Learning: ASR $\leftrightarrow$ SER

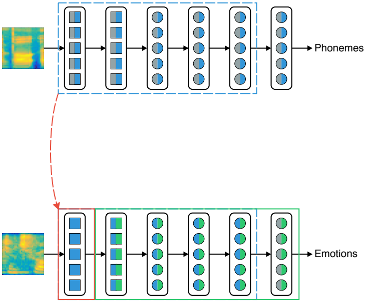

Figure: Transfer Learning between ASR and SER.

Transfer Learning: ASR $\leftrightarrow$ SER

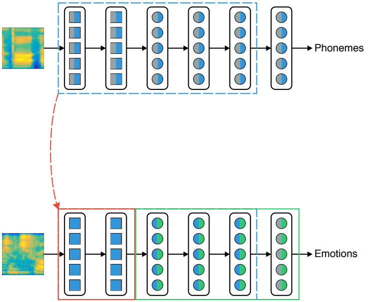

Figure: Transfer Learning between ASR and SER.

Transfer Learning: ASR $\leftrightarrow$ SER

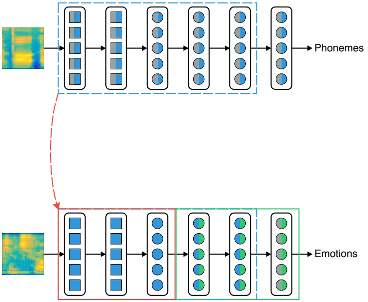

Figure: Transfer Learning between ASR and SER.

Transfer Learning: ASR $\leftrightarrow$ SER

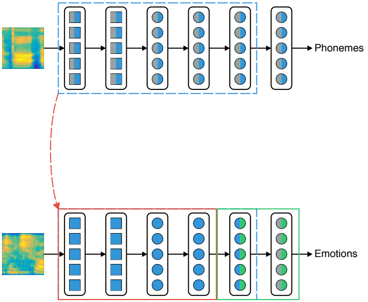

Figure: Transfer Learning between ASR and SER.

Transfer Learning: ASR $\leftrightarrow$ SER

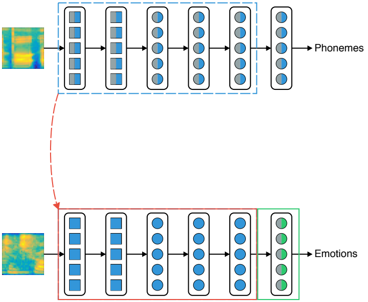

Figure: Transfer Learning between ASR and SER.

Transfer Learning: ASR $\leftrightarrow$ SER

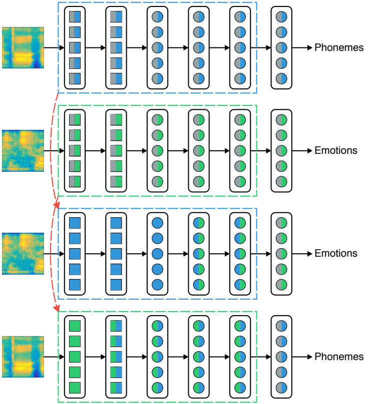

Figure: Transfer Learning between ASR and SER.

Outline

Learned Features

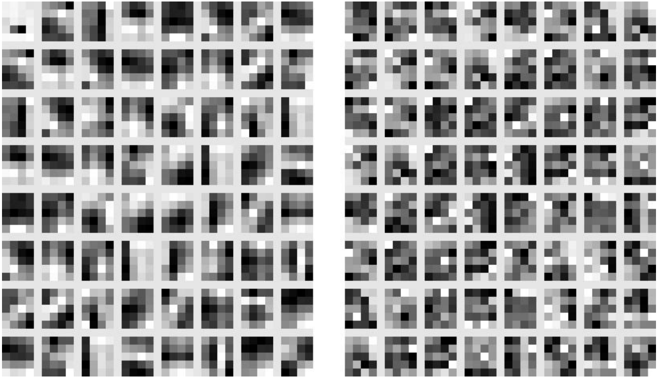

Figure: Learned Features from ASR (left) and SER (right).

Results: SER to ASR

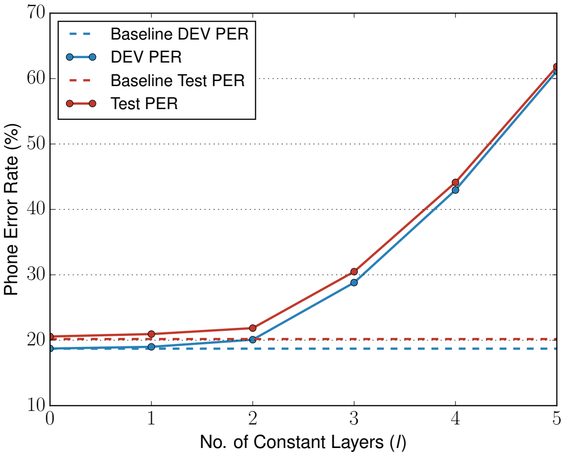

Figure: Transfer Learning Performance from SER to ASR.

Results: SER to ASR

Table: *Transfer Learning Performance SER to ASR.*

No. Constant Layers (l)	FER		PER	
	Dev	Test	Dev	Test
Baseline	30.53%	31.61%	18.71%	20.18%
5	71.09%	71.64%	61.15%	61.82%
4	53.26%	53.92%	42.96%	44.13%
3	40.29%	40.97%	28.81%	30.48%
2	31.75%	32.87%	20.08%	21.85%
1	30.83%	32.01%	18.99%	20.94%
0	30.62%	31.65%	18.73%	20.57%

Results: ASR to SER

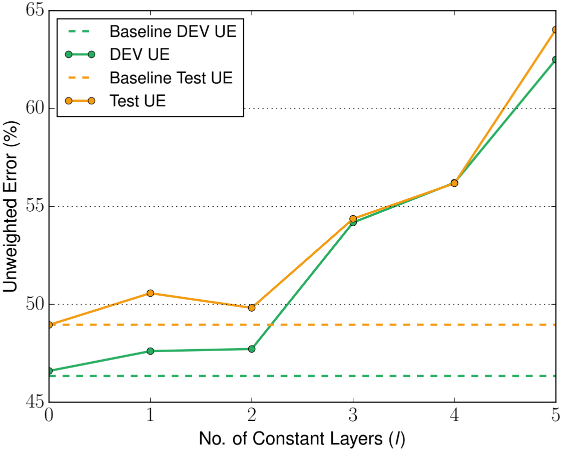

Figure: Transfer Learning Performance from ASR to SER.

Results: ASR to SER

Table: *Transfer Learning Performance ASR to SER.*

No. Constant Layers (*l*)	E		UE	
	Dev	Test	Dev	Test
Baseline	44.63%	46.44%	46.34%	48.96%
5	52.55%	59.20%	62.50%	64.03%
4	51.94%	53.34%	56.21%	56.18%
3	50.22%	52.01%	54.18%	54.37%
2	47.39%	48.50%	47.72%	49.82%
1	46.37%	48.36%	47.61%	50.57%
0	45.26%	46.97%	46.60%	48.95%

Results

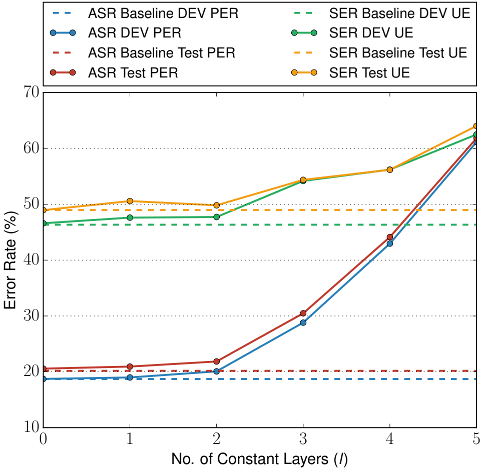

Figure: Transfer Learning Performance between ASR and SER.

Outline

Conclusion

- The relevance of features learned and information propagation in ConvNets between ASR and SER was studied using transfer learning.
- Results attested to the feasibility of transfer learning between both tasks.
- Initial layers in the network were more transferable between both tasks and the relevance of features decays gradually through deep layers.

Thank you

Questions & Discussion