An Analysis Of Speech Enhancement And Recognition Losses In Limited Resource Multi-Talker Single Channel Audio-Visual ASR

Luca Pasa, Giovanni Morrone, Leonardo Badino

ICASSP 2020

ISTITUTO ITALIANO DI TECNOLOGIA

UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

ICASSP2020 Barcelona

ASR and the Cocktail-Party Problem

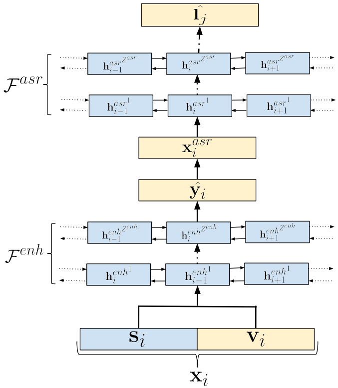
- State-of-the-art ASR can be very accurate but **performance drops significantly in a cocktail party scenario**
- Recognizing the speech of a target speaker mixed with other people speech's in a single-channel audio is an **ill-posed problem**
  - Many different hypotheses about what the target speaker says are consistent with the mixture signal, **we do not know which utterance corresponds to the target speaker**
  - We addressed this problem by exploiting an additional information: **the video of talking face of the target speaker**

- Some robust ASR systems process the audio signal through a speech enhancement or separation stage
- Jointly training the ASR and enhancement modules can be more beneficial than training them separately
- **Goal: analyze the interaction between the ASR and enhancement tasks**
  - Understand whether (and how) it is advantageous to train them jointly
- **How?**
  - Train and analyze a simple AV-ASR model
  - Analyze whether adding a preliminary speech enhancement stage helps in performing the ASR task

Joint Model

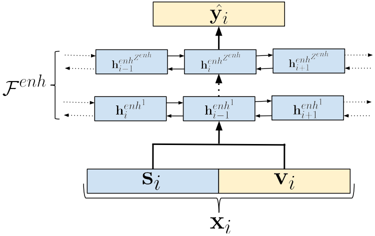
We analyze a simple and common architecture:

- Based on deep-BLSTM
- Composed of 2 sub-models:
  - Enhancement Model
  - ASR Model
- With the following model inputs:
  - Noisy Audio information: $\mathbf{s} = [\mathbf{s}_1, \ldots, \mathbf{s}_T]$
  - Face Motion vector: $\mathbf{v} = [\mathbf{v}_1, \ldots, \mathbf{v}_T]$
- Where only the enhancement part exploits the visual information, while the ASR part receives in input only the output of the speech enhancement module



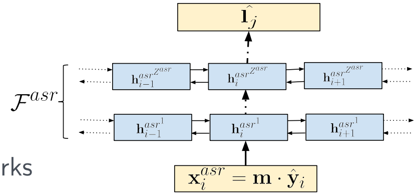
Enhancement Model

- **Goal:** de-noising the speech of the speaker of interest
- **Input** at time step i: $\mathbf{x}_i = \begin{bmatrix} \mathbf{s}_i \\ \mathbf{v}_i \end{bmatrix}$,
- **Target:** a slice of the spectrogram of the clean utterance spoken by the target speaker.
- **Loss function:** Mean Squared Error (*MSE*)
  - $\mathcal{L}^{enh}(\mathbf{y}_i, \hat{\mathbf{y}}_i) = MSE(\mathbf{y}_i, \hat{\mathbf{y}}_i)$.



ASR Model

- Input: computes the mel-scale filter bank representation derived from the spectrogram s_i
- Maps $\mathbf{x}_i^{asr}$ to the phone label $\hat{\mathbf{l}}_i$ by using Z^{asr} BLSTM layers
- Uses the CTC loss
  - $\mathcal{L}^{asr}(\mathbf{l}_j, \hat{\mathbf{l}}_j) = CTC_{loss}(\mathbf{l}_j, \hat{\mathbf{l}}_j)$
- 3 different versions:
  1. Fed with acoustic features
     $$\mathbf{x}_i^{asr} = \mathbf{s}_i^m$$
  2. Fed with motion vector computed from face landmarks
     $$\mathbf{x}_i^{asr} = \mathbf{v}_i$$
  3. Uses both audio and visual features
     $$\mathbf{x}_i^{asr} = \begin{bmatrix} \mathbf{s}_i^m \\ \mathbf{v}_i \end{bmatrix}$$



Two Training Strategies

Goal: Analyse the behaviors of the ASR and enhancement loss

- **Joint training**
  - $\mathcal{L}_{join} = \lambda \cdot \mathcal{L}^{enh} + \mathcal{L}^{asr}$
  - We explored 2 different types of λ:
    - Constant
    - Adaptive: $\lambda_{adapt} = 10^{\lfloor \log_{10}(\mathcal{L}^{asr}) \rfloor} / 10^{\lfloor \log_{10}(\mathcal{L}^{enh}) \rfloor}$
- **Alternated training**
  - Alternation of speech enhancement and ASR training phases
  - Performs a few steps of each phase several times
  - **Alternated two full phases training**
    - the two phases are performed only one time each
  - The $\mathcal{L}^{asr}$ optimization phase updates both θ^{enh} and θ^{asr} parameters
  - **Weight freezing**: optimize $\mathcal{L}^{asr}$ by only updating θ^{asr}

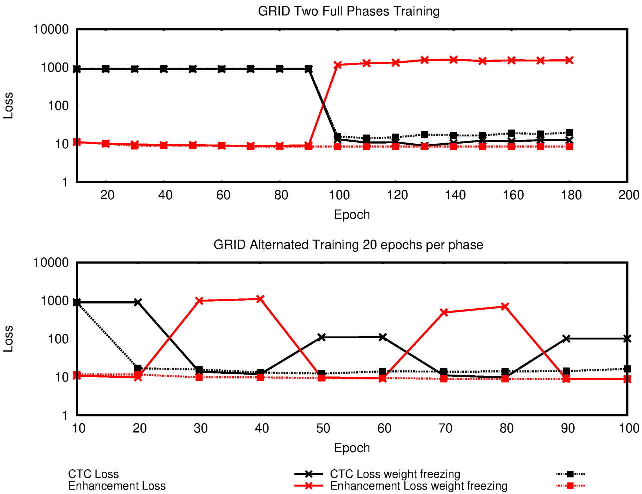
Experimental Setup

- Two Audio-visual limited-size datasets
  - **GRID**, **TCD-TIMIT**
  - **Speaker-independent**
  - Respectively split into disjoint sets of 25/4/4 and 51/4/4 speakers for training/validation/testing
  - Used standard TIMIT phone dictionary
    - GRID: 33 phones, TCD-TIMIT: 61 phones
- Baseline
  - ASR-only models
    - 2 layers of 250 hidden units and were trained by using back-propagation through time (BPTT) with Adam optimizer
- Joint Model
  - Same number of layers for both ASR and enhancement components

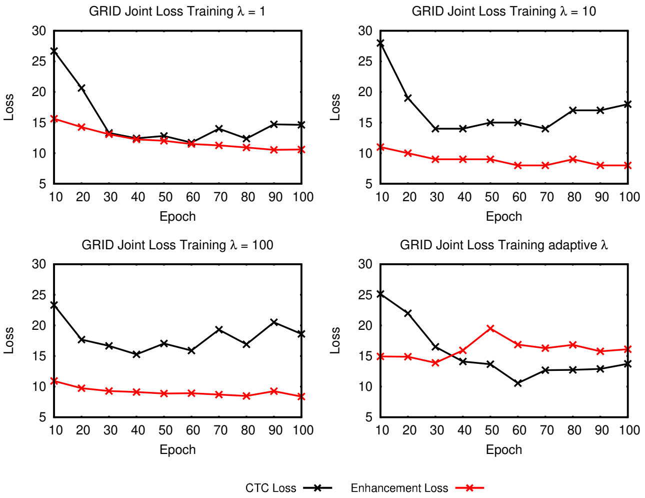
Experimental Results

Training Method	GRID PER	TCD-TIMIT PER-61	TCD-TIMIT PER-39
Baseline-ASR-Mod. Clean-Audio	5.8	46.7	40.6
Baseline-ASR-Mod. Mixed-Audio	49.4	78.4	71.3
Baseline-ASR-Mod. Mixed-A/V	49.9	77.2	70.9
Baseline-ASR-Mod. Visual	29.4	78.6	74.7
Joint-Mod. Joint Training	15.4	53.1	47.7
	$\lambda = 1$	$\lambda = \lambda_{adapt}$	
Joint-Mod. Alt. Training 2 full	16.0	45.6	41.2
Joint-Mod. Alt. Training 2 full freeze	18.7	**44.3**	**40.0**
Joint-Mod. Alt.Training	**13.9**	44.9	40.6
Joint-Mod. Alt. Training freeze	18.1	61.3	55.5
Joint-Mod. PIT Alt. Training	43.3	67.1	62.4

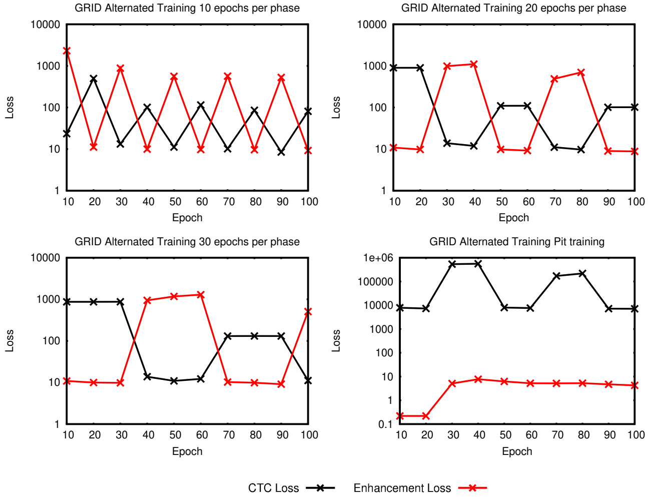
Alternate Training Analysis



Joint Loss Training Analysis



Alternated Training Analysis



Conclusion

- Jointly minimizing the speech enhancement loss and the CTC loss may not the best strategy to improve ASR
- Alternation of the speech enhancement and ASR training phases
  - The loss function that was not considered for the training phase tends to diverge
- The interaction between the two loss functions can be exploited in order to obtain better results
  - The alternated training method shows that the recognition error can be gradually reduced by wisely alternating the two training phases

Thanks for the attention!

Contacts:

Luca Pasa: lpasa@math.unipd.it
Giovanni Morrone: giovanni.morrone@unimore.it
Leonardo Badino: leobad08@gmail.com