Multi-Task Joint-Learning for Robust Voice Activity Detection

Yimeng Zhuang, Sibo Tong, Maofan Yin, Yanmin Qian, Kai Yu

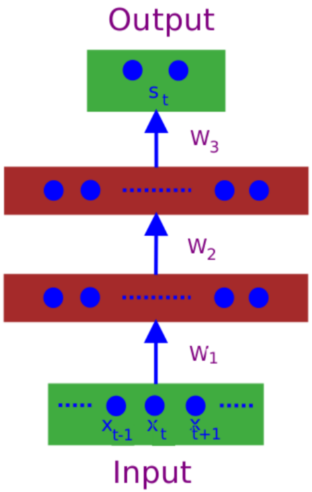
Speech Lab
Department of Computer Science & Engineering
Shanghai Jiao Tong University

October 2016

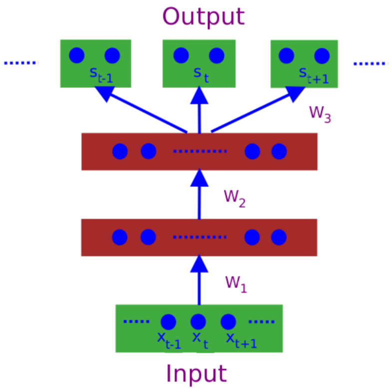
VAD Overview

- Voice activity detection
  - A technique used in speech processing in which the presence or absence of human speech is detected
- Model based VAD
  - Zero crossings rate
  - Energy
  - Long term spectral
  - Gaussian mixture model(GMM)
  - Deep neural network(DNN) based VAD

Basic DNN based VAD

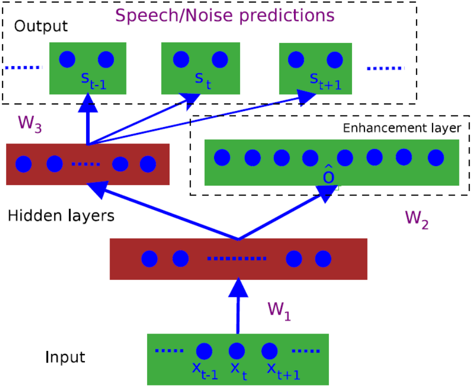


Multi-frame prediction



$$\mathcal{L}_{vad}(\mathbf{W}) = -\frac{1}{N}\sum_{n=1}^{N}\sum_{t=-M}^{M}\lambda_t\sum_{i=1}^{2} d_{s_{(n+t)i}} \log P(s_{(n+t)i}|\mathbf{o}_n, \mathbf{W}) \quad (1)$$

Train multi-frame DNN with multi-task joint-learning



$$\mathcal{L}(\mathbf{W}) = \mathcal{L}_{vad}(\mathbf{W}) + \frac{1}{N}\sum_{n=1}^{N} \| \hat{\mathbf{o}}_n - \mathbf{o}_n \|_2^2 + \kappa \| \mathbf{W} \|_2^2 \quad (2)$$

Prediction

- Enhancement layer is removed
- Functions to combine multiple prediction results

- Maximum:

$$P(s_t|\mathbf{o},\mathbf{W}) = \max_{-M \le i \le M} \{P(s_t|\mathbf{o}_{t+i},\mathbf{W})\} \tag{3}$$

- Arithmetic mean:

$$P(s_t|\mathbf{o},\mathbf{W}) = \frac{1}{2M+1}\sum_{i=-M}^{M} P(s_t|\mathbf{o}_{t+i},\mathbf{W}) \tag{4}$$

- Harmonic mean:

$$\frac{1}{P(s_t|\mathbf{o},\mathbf{W})} = \frac{1}{2M+1}\sum_{i=-M}^{M} \frac{1}{P(s_t|\mathbf{o}_{t+i},\mathbf{W})} \tag{5}$$

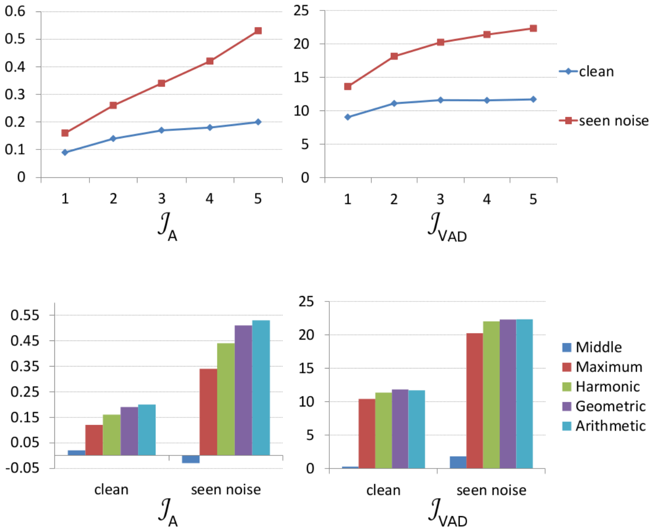
- Geometric mean:

$$\log P(s_t|\mathbf{o},\mathbf{W}) = \frac{1}{2M+1}\sum_{i=-M}^{M} \log P(s_t|\mathbf{o}_{t+i},\mathbf{W}) \tag{6}$$

Experiment Setup

- Aurora 4 dataset is used
- Six different types of noises, including airport, babble, car, restaurant, street and train
- 10-20 dB SNR
- 7 test sets, including the clean set and six noise sets (seen noise)
- To simulate a more realistic scenario, an unseen noise test set is designed with 100 noise types

Choosing context window size and score combination methods



Frame-level evaluation (AUC)

Hidden layers	Noise condition	Single frame	Multi-frame	Multi-frame + Multi-task
2 (1+1)	clean	99.75	99.78	**99.79**
	seen	98.85	98.95	**99.00**
	unseen	96.62	97.35	**97.72**
3 (2+1)	clean	99.76	**99.79**	**99.79**
	seen	98.90	99.03	**99.08**
	unseen	96.82	97.58	**97.95**

- The model of multi-frame prediction with multi-task joint-learning yields best results
- The multi-task approach is an effective method to further impove VAD performance at frame-level.

Segment-level evaluation ($\mathcal{J}_{VAD}$)

Hidden layers	Noise condition	Single frame	Multi-frame	Multi-frame +Multi-task
2 (1+1)	clean	81.6	90.28	**91.0**
	seen	55.4	71.81	**71.9**
	unseen	45.9	63.80	**65.7**
3 (2+1)	clean	82.2	90.23	**91.3**
	seen	56.5	71.89	**75.1**
	unseen	46.0	63.86	**66.6**

- $\mathcal{J}_{VAD}$ is sensitive to boundary accuracy and the total number of speech/non-speech segments. Improved $\mathcal{J}_{VAD}$ suggests that the proposed approaches produce more accurate boundaries and less fragiles.

Conclusion

- Multi-frame prediction with multi-task joint learning is utilized for VAD
- The proposed approach need to predict classification posteriors covering the neighboring multiple frames
- A speech enhancement task is jointly trained in order to generate better regression ability
- Future work
    - More experiments are needed to exam whether other score combination functions can get a better performance
    - Also it is worth exploiting a postprocessing method that suits this new proposed approach