Lasso-based Reverberation Suppression In Automatic Speech Recognition

Xuewei Zhang, Yiye Lin, Dong Wang
Center for Speech and Language Technologies (CSLT)
Tsinghua University

Overview

- **Far-field automatic speech recognition(ASR) is challenging**
- **Lasso is a novel linear sparse prediction model which estimates the late reflection**
- **We apply three Lasso-based de-reverberation approaches to far-field speech recognition based on deep neural networks**

Lasso-based de-reverberation

- **Far field signal**

$$x[t]=\mathrm{s}[t] * (r_e[t]+r_f[t])+\mathrm{n}[t]$$

- ➢ $x[t]$ **: the received reverberated signal**
- ➢ $s[t]$ **: the direct signal**
- ➢ $n[t]$ **: the background noise**
- ➢ $r_e[t]$ **: the early room impulse response**
- ➢ $r_f[t]$ **: the late room impulse response**

Lasso-based de-reverberation

- **Reverberated signal**

$$X_{k,n} = S_{k,n} + \sum_{i=0}^{I-1} \beta_{k,n,i} X_{k,n-i} + \sum_{l=0}^{L-1} \alpha_{k,n,l} X_{k,n-\delta-l}$$

- ➢ $S_{k,n}$ **follows a zero-mean Gaussian distribution**
- ➢ $\{\alpha_{k,n,i}\}$, $\{\beta_{k,n,i}\}$ **represent the model parameters**
- ➢ I **represents the maximum delay of the early reflection**
- ➢ L **represents the maximum delay of the late reflection**

Lasso-based de-reverberation

- **The sparse linear prediction model**

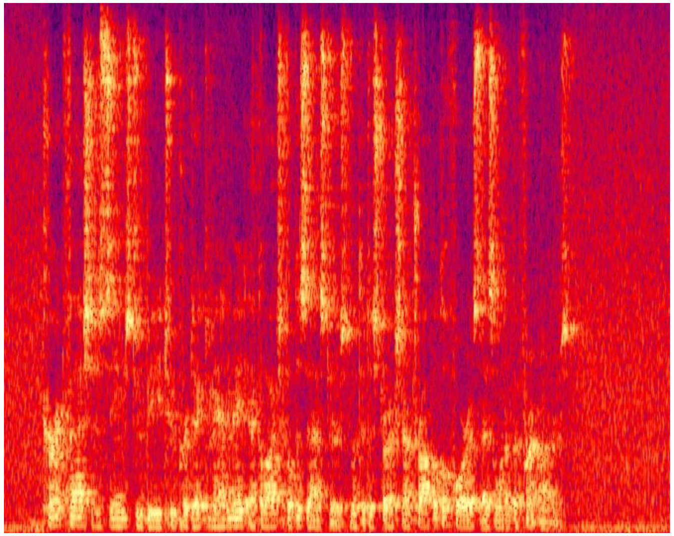
$$\min_{\{\alpha_{k,n,l}\}} \left| x_{k,n} - \sum_{l=0}^{L-1} \alpha_{k,n,l}\, x_{k,n-\delta-l} \right|^2$$

$$s.t \sum_{l=0}^{L-1} \left|\alpha_{k,n,l}\right| \leq \lambda$$

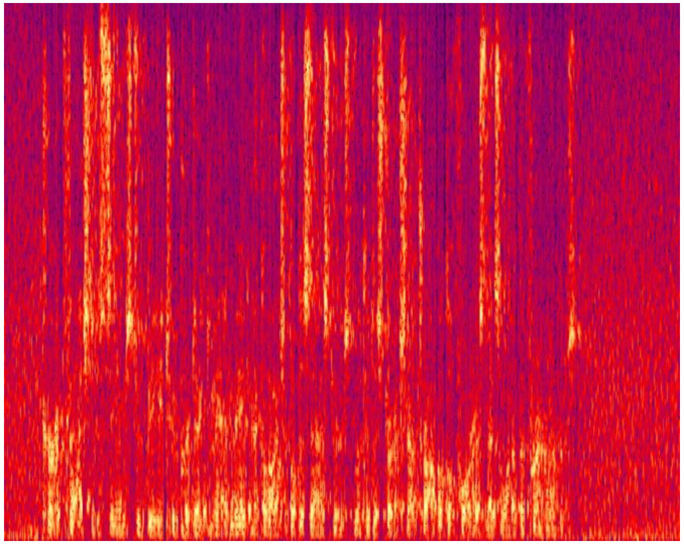
➢ **λ : a regularization parameter**

Lasso-based de-reverberation

- **Reverberated speech signal and the Lasso-based de-reverberation signal**



(a) reverberated signal



(b) dereverberated signal

Lasso-based de-reverberation for speech recognition

- **Although promising in perceptual experiments, it is unknown if the Lasso-based dereverberation can improve far-field ASR**
- **Inferring the regression coefficients $\alpha_{k,n,l}$ for each frame and each frequency channel involves very demanding computation**

Lasso-based de-reverberation for speech recognition

- **FBank element-based Lasso**
  - **the Mel channels are independent**
  - **FBank-based Lasso is easily integrated in the frontend pipeline of the ASR system**

Lasso-based de-reverberation for speech recognition

- **FBank frame-based**

$$\min_{\{\alpha_{n,l}\}} ||x_n - \sum_{l=0}^{L-1} \alpha_{n,i}\, x_{n-\delta-l}||^2$$

$$s.t \sum_{l=0}^{L-1} |\alpha_{n,l}| \leq \lambda$$

➢ **The late reflection contributes to all channels in the same way, so that the regression coefficients can be shared**

➢ **$||\cdot||$:the Frobenius norm**

Lasso-based de-reverberation for speech recognition

- **FBank utterance-based Lasso**

$$\min_{\{\alpha_l\}} ||x_n - \sum_{l=0}^{L-1} \alpha_l \, x_{n-\delta-l}||^2$$

$$s.t \sum_{l=0}^{L-1} |\alpha_l| \leq \lambda$$

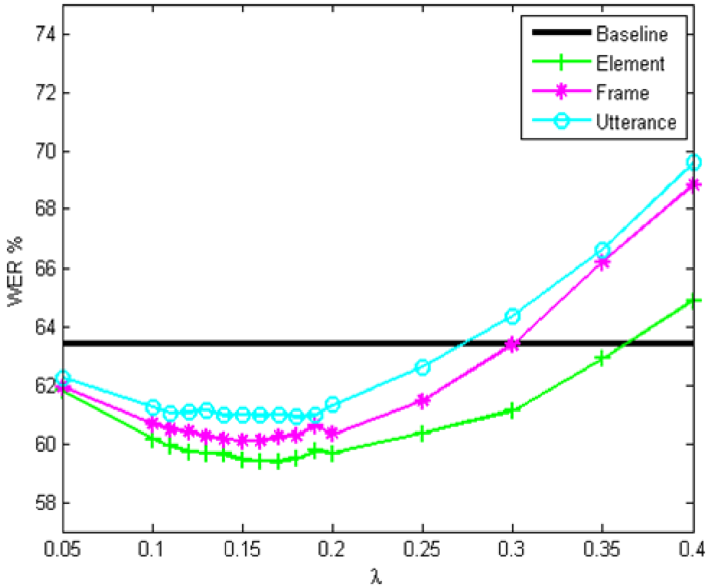
- ➢ **Reducing computation cost in the frontend of ASR systems**
- ➢ **Considering that in a stationary environment where the locations of the speaker and the microphone are both unchanged, the regression coefficients should be shared among all the frames**

Experiments

- **Experimental settings**
  - **The wsj dev93 dataset (503 utterances) and eval92 dataset (333 utterances) were used to conduct the development set and evaluation set**
  - **Two approaches were used to generate the reverberated version**
  - **using the Kaldi**
  - **40-dimensional Fbanks feature**
  - **The DNN architecture involves 4 hidden layers and each layer consists of 1200 units. The output layer is composed of 3447 units**
  - **Mini batch size is set to 256 frames**
  - **The learning rate started from a relatively large value 0.008**

Experiments

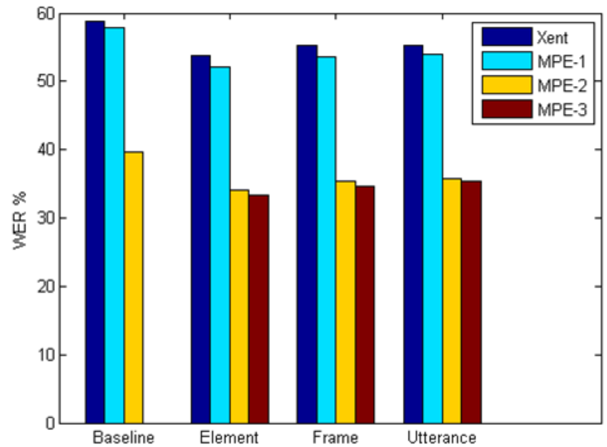
- **Estimate λ**



- **Using element-based, frame-based and utterance-based methods, the corresponding optimal λ is 0.17, 0.15 and 0.14.**
- **The computation speed of Lasso based on utterance is twice faster than that of the other two methods**
- **The utterance-based method is particularly suitable for real- time ASR.**

Experiments

- **Results on simulated data**

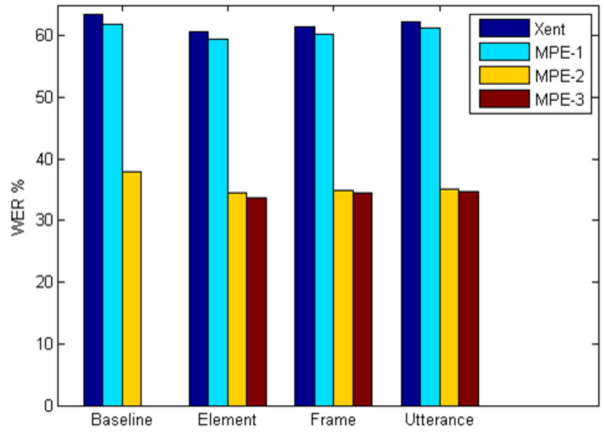


In any case (Xent and MPEs), the Lasso-based de-reverberation delivers clear performance improvement compared to the baseline results

- **The element-based method is slightly better**

Experiments

- **Results on real reverberated data**



- **We can draw similar conclusions as with the simulated data**

Conclusions

- **This paper experimented with a Lasso-based de-reverberation approach in DNN-based speech recognition**
- **The new de-reverberation approach can deliver significant performance improvement on both simulated and real reverberated speech data**
- **The utterance-based method is much faster than the element and frame-based methods, so it is suitable to be applied to real-time ASR**