Audio Augmentation for Speech Recognition

Tom Ko[1], Vijayaditya Peddinti[2], Daniel Povey[2,3], Sanjeev Khudanpur[2,3]

[1]*Huawei Noah's Ark Research Lab, Hong Kong, China*
[2]*Center for Language and Speech Processing &*
[3]*Human Language Technology Center of Excellence,*
Johns Hopkins University, Baltimore, MD, 21218, USA

Background

- Data augmentation is a common strategy adopted to
  - Increase the quantity of training data.
  - Avoid overfitting.
  - Improve robustness of the models.
- Advantages of audio augmentation in speech recognition:
  - Low implementation cost.
  - Easy to adopt by different system architectures.

Objectives

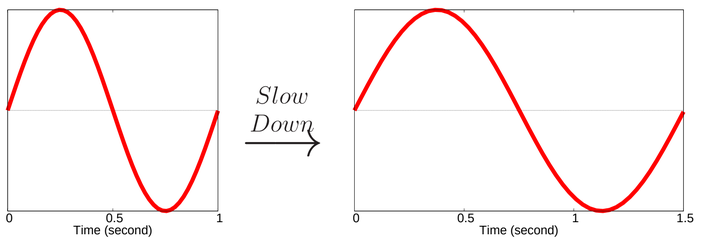
- To investigate the effectiveness of several audio augmentation techniques using state-of-the-art DNN systems.
- Propose a new audio augmentation approach which perturb the speed of the audio signal.

Previous Work

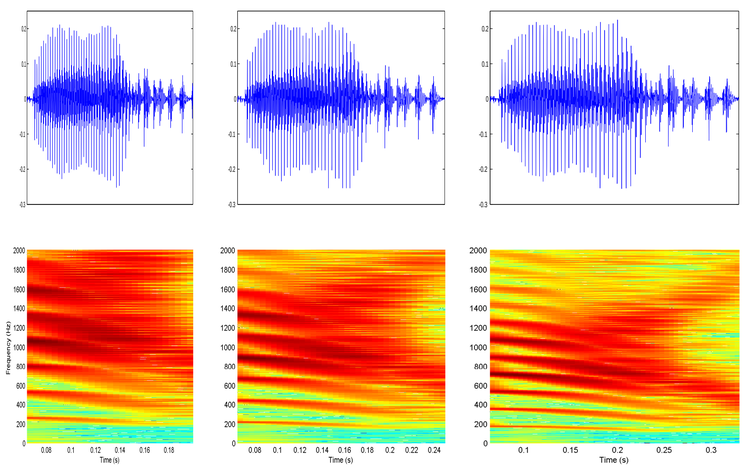
- VLTP [2], Noise [1], Tempo [3] perturbation

Speed Perturbation

- Given an audio signal $x(t)$, time warping by a factor α gives the signal $x(\alpha t)$. It can be seen from the Fourier transform of $x(\alpha t)$, $\alpha^{-1}\hat{x}(\alpha^{-1}\omega)$, that the warping factor produces shifts in the frequency components of the $\hat{x}(\omega)$ by an amount proportional to frequency ω.
- On mel averaging the variations in speed are converted to perturbation in mel spectral envelope. Further it also produces tempo perturbation.



- Signal with 80%, 100% and 120% speed and corresponding spectrograms



Acoustic Modeling

- Time-delay neural network (TDNN) with 4 hidden layers.
- *p*-norm non-linearity.
- Each *p*-norm layer was followed by a normalization layer. This layer scales the input vector by its root mean square value.

Experimental Setup

- The following augmentation methods are used to create multiple copies of the original training set.
- Parallel training of the DNNs using up to 18 GPUs was done using the model averaging technique to keep similar training time of the systems.

Speed Perturbation

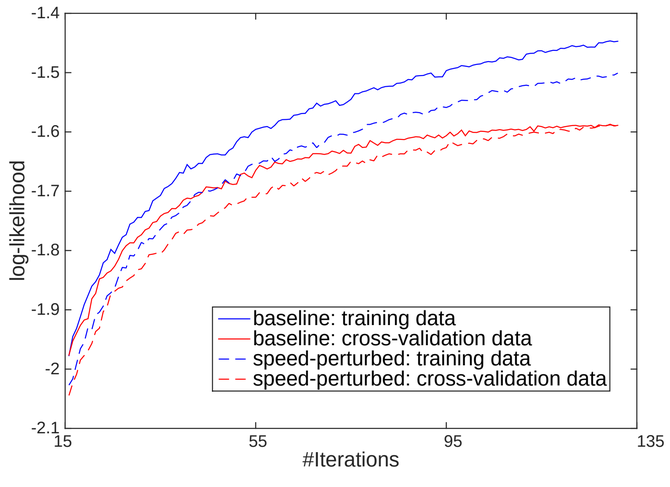
- To modify the speed of a signal we resample the signal with the *speed* command of the *SoX* tool.
- The pitch and spectral evelope of the signal are both changed.
- Two additional copies of the original training data were created by modifying the speed to 90% and 110% of the original rate.

Tempo Perturbation

- The pitch and spectral evelope of the signal does not change.
- The WSOLA based implementation in the *tempo* command of the *SoX* tool was used in our work.

Vocal Tract Length Perturbation (VLTP)

- Two sets of warping factors, $\{0.9, 1.0, 1.1\}$ and $\{0.9, 0.95, 1.0, 1.05, 1.1\}$, are used to create 3 and 5 copies of the original feature vectors.



Results

- Results for various systems on the Hub5 00 evaluation set using a 4-gram language model.

System	Fold	Epochs	SWB	CHE	Total
Baseline	1	6	13.7	27.7	20.7
VTLP	3	2	13.1	26.5	19.9
VTLP	3	6	12.9	26.5	19.7
VTLP	5	2	13.2	26.7	20.0
VTLP + time-warp	3	2	13.3	26.8	20.1
Tempo-perturbed	3	2	13.5	27.0	20.3
Speed-perturbed	3	2	13.1	26.1	19.7
Speed-perturbed	3	6	12.9	25.7	**19.3**

- Results on the GALE Mandarin test set using a trigram language model.

System	Fold	Epochs	Pitch	Total
Baseline	1	6	N	18.46
Baseline	1	12	N	18.63
Speed-perturbed	3	2	N	18.34
Speed-perturbed	3	6	N	**18.09**
Baseline	1	6	Y	17.51
Baseline	1	12	Y	17.63
Speed-perturbed	3	2	Y	17.56
Speed-perturbed	3	6	Y	**17.16**

- Comparison of baseline and speed-perturbation on various tasks.

LVCSR task	Hrs	WER		Relative
		Baseline	Sp	imp.
GALE Mandarin	100	17.51	17.16	2.0
Tedlium	118	17.9	17.2	3.9
Switchboard	300	20.7	19.3	6.7
Librispeech	960	12.93	12.51	3.2
ASpIRE	5500	30.8	30.7	0.32

Conclusions

- We presented an audio augmentation technique with low implementation cost.
- Speed perturbation, which emulates both VTLP and tempo perturbation, is shown to give more WER improvement than either of those methods.

Future Work

- Investigation of perturbing the audio in other ways (e.g. simulated channel distortion).

References

[1] Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, and Andrew Y. Ng. Deep speech: Scaling up end-to-end speech recognition. *CoRR, abs/1412.5567*, 2014.

[2] N. Jaitly and G. E. Hinton. Vocal tract length perturbation (VTLP) improves speech recognition. In *International Conference on Machine Learning (ICML) Workshop on Deep Learning for Audio, Speech, and Language Processing*, 2013.

[3] N. Kanda, R. Takeda, and Y. Obuchi. Elastic spectral distortion for low resource speech recognition with deep neural networks. In *ASRU*, 2013.