MULTI-CONDITION DEEP NEURAL NETWORK TRAINING

Matthew Gibson, Christian Plahl, Puming Zhan and Gary Cook

Nuance Communications Inc.

ABSTRACT

Multi-condition training (MCT) is the incorporation of data sourced from a diverse range of conditions into the training dataset. In the case of acoustic modelling, examples of such conditions may be broadly categorised as background noise conditions, reverberation conditions and acoustic channel (e.g. telephone or microphone). The aim of MCT is to deliver models which are more robust to conditions which are not represented in the original training dataset.

If transcribed speech data is readily available for a wide variety of acoustic conditions, then this data can be directly used for MCT. More typically, such data is not available and must be synthesised, a process referred to as speech corruption. While such corruption is reasonably well-documented, e.g. in the case of the Google Home product ([1, 2]), there is little published work which examines the relationships between the details of the corruption technique and the effectiveness of the resulting MCT process. The aim of this work is to investigate these relationships i.e. to measure the impact of a variety of speech corruption methods upon the accuracy of the resulting multi-condition trained models. The techniques are evaluated by measuring the error rate of the resulting models on a home automation speech dataset. In this work, speech signals will be corrupted by introducing reverberation and background noise at a specified signal-to-noise ratio (SNR). The corruption of a signal is highly configurable, with control over the following:

- the background noise source used.
- the SNR of the resulting corrupted signal (the target SNR).
- the reverberation source (in the form of a room impulse response).
- whether to treat the background noise as a point noise source (in which case it will also be reverberated) or additive noise.

Further, the corruption process may be configured as symmetric, asymmetric or noise-symmetric. Symmetric corruption outputs a corrupted version of each training set utterance for every incorporated condition combination (background noise source, target SNR and reverberation source). Symmetric corruption outputs a copy of the training dataset for every condition combination. To constrain the amount of data generated by symmetric corruption the number of considered conditions must often be limited. In the case of asymmetric corruption, for each training set utterance, a condition combination is randomly sampled from all possible such combinations and used to produce a corrupted utterance. Asymmetric corruption outputs only one corrupted version of the training dataset. Noise-symmetric corruption is the application of asymmetric corruption over target SNRs and reverberation sources for each noise source. So noise symmetric corruption outputs a copy of the training dataset for each noise source in the corruption configuration.

Another refinement of the corruption process is to (optionally) corrupt only those training set utterances which contain relatively little background noise or reverberation. To achieve this, thresholds are applied to the SNR and reverberation level (measured using the C50 metric) of the input utterance. This procedure is referred to as input filtering.

Specifically, in this work the following aspects of MCT will be examined and their impact up the resulting models measured:

- the list of noise sources incorporated.
- the list of target SNRs.
- the list of reverberation sources incorporated.
- the reverberation of point noise sources.
- asymmetric versus noise-symmetric corruption.
- input filtering.

1. REFERENCES

[1] B. Li, T. Sainath, A. Narayanan, J. Caroselli, M. Bacchiani, A. Misra, I. Shafran, H. Sak, G. Pundak, K. Chin, K. C. Sim, R. J. Weiss, K. Wilson, E. Variani, C. Kim, O. Siohan, M. Weintraub, E. McDermott, R. Rose, and M. Shannon, "Acoustic modeling for Google Home," in *Proceedings Interspeech*, 2017.

[2] C. Kim, A. Misra, K. Chin, T. Hughes, A. Narayanan, T. Sainath, and M. Bacchiani, "Generation of large-scale simulated utterances in virtual rooms to train deep-neural networks for far-field speech recognition in Google Home," in *Proceedings Interspeech*, 2017.