Noise Robust Multi-Band Speech Recognition Combining Gabor Filters and Time Delay Neural Networks*

György Kovács[1] and Marcus Liwicki[1]

Embedded Internet Systems Lab, Luleå University of Technology, Luleå, Sweden
gyorgy.kovacs, marcus.liwicki@ltu.se

Abstract. Here, we investigate the combination of two complementary speech processing methods designed to increase robustness of speech recognition, namely multi-band processing and spectro-temporal feature extraction (using Gabor filters). We do so by first optimising the meta-parameters of multi-band processing using the TIMIT speech corpus. Then we apply the resulting meta-parameters on the Aurora-4 corpus to validate their cross-corpus viability. Lastly, we combine multi-band processing with the recently proposed band dropout method. The resulting error rates are significantly lower than those got with the individual methods. Furthermore, these error rates compare favourably to those recently reported in the speech recognition literature for the same task.

Keywords: Multi-band processing, band dropout, noise robust ASR

1 Introduction

A spectro-temporal processing method we found effective in our earlier studies [2] is that of processing by Gabor filters. For this here, we combine the technique of filtering with Gabor filters with that of multi-band processing. This means that the spectro-temporal features extracted from different frequency regions (or bands) are processed independently by dedicated Time Delay Neural Nets (TDNNs). Then, the outputs of these TDNNs (in our work this is the output of a bottleneck layer) are concatenated and processed by a combinational network.

2 Experimental setup

Two English language corpora were used in our experiments. First, the **TIMIT** [3] database, in the standard train/test partitioning (3696/192 utterances). The proposed methods were also evaluated using the **Aurora-4** [4] corpus. Here, only the train cleaning scenario (3692 utterances) was used so as to highlight the robustness of the proposed method to unseen noise types. The trained models were evaluated on 14 different versions of the test set (330 utterances).

* Published in: G. Kovács, L. Tóth, G. Gosztolya, "Multi-Band Processing With Gabor Filters and Time Delay Neural Networks for Noise Robust Speech Recognition," in Proc. IEEE SLT 2018. pp 242–249

A pivotal question for multi-band processing is the formation of bands. Here, we examined several aspects of this, including the number of bands to create, the possible overlap between bands. We also examined the case where instead of using information from one band in each TDNN, we use information from all but one bands. And as a baseline, we also experimented with the case where information from all bands are concatenated and processed together.

Two types of neural networks were applied in our experiments. One was the **band classifier**, created to process information from individual bands (or combination of bands). These are 4-layer TDNNs extended with sub-sampling, applying the ReLu activation function in all layers, with the exception of the output layer, and the linear bottleneck layer directly preceding the output layer. The **combinational** neural networks (processing the combined information of the band classifiers) were simple fully connected rectifier neural nets with three hidden layers (each containing a thousand neurons), and a softmax output layer. For each experiment we averaged the error rates got using ten such networks.

3 Experiments and Results

Our experiments on the **TIMIT** corpus had two important outcomes. First, the results showed which meta-parameters to use in future experiments. Then the error scores we got affirmed the viability of multi-band processing, by demonstrating a capacity for increasing the robustness of ASR (leading to a relative error rate reduction - RER - of close to 6% in real-life noise types, and 25% in band-limited noise), while still being competitive in clean speech. Our experiments on the **Aurora-4** corpus also demonstrated a significant improvement in error rates when using multi-band processing (a RER of 17% was attained with fewer trainable parameters compared to the baseline). And the improvement was particularly pronounced in noise contaminated speech (RER: 21%).

Lastly, we combined the technique of multi-band processing with that of band dropout [2]. This means that during the training of the combinational network, information from up to six (out of ten) bands was "dropped" with a probability of 60%, so as to prevent the network from over-relying on any given band. The resulting overall 25% error rate was a 8% relative improvement over our earlier results, and a 2% relative improvement over the results of Alam et al. [1].

References

1. Alam, M.J., Kenny, P., O'Shaughnessy, D.: Robust feature extraction based on an asymmetric level-dependent auditory filterbank and a subband spectrum enhancement technique. Digital Signal Processing 29(Complete), 147–157 (2014)
2. Kovács, G.: Noise Robust Automatic Speech Recognition Based on Spectro-Temporal Techniques. Ph.D. thesis, University of Szeged (2018)
3. Lamel, L.F., Kassel, R., Seneff, S.: Speech database development: design and analysis of the acoustic-phonetic corpus. In: Proc. DARPA SR Workshop (1986)
4. Parihar, N., Picone, J.: DSR front end LVCSR evaluation. Aurora Working Group AU/384/02, Institute for Signal and Information Processing (December 2002)