The Joint Optimization of Spectro-Temporal Features and Deep Neural Nets for Robust ASR

György Kovács and László Tóth

One of the biggest challenges facing automatic speech recognition is to get an acceptable performance even in an adverse environment, e.g speech with background noise. One way of increasing the robustness of ASR systems is to apply spectro-temporal processing on the speech signal [1]. In this approach, the features for the subsequent classification are got by processing small, spectrally and temporally localized patches of the spectrogram. Hence, unlike in traditional processing methods, some of the features may be unnaffected by the noise component, making the whole feature set more robust. The spectro-temporal processing of the patches can be performed by using the two-dimensional discrete cosine transform (2D DCT) or Gabor filters. Good recognition results were reported with both of these approaches earlier [1, 2, 3].

After the initial step of feature extraction, the features are passed on to a machine learning algorithm – in most cases a Hidden Markov Model (HMM) or an Artificial Neural Net (ANN). Normally, the feature extraction and recognition steps are performed separately. In our previous paper, we showed that the spectro-temporal feature extraction step and the ANN-based recognition step can be integrated, and in this way the parameters of the two phases can be trained together [4]. Our solution was based on the observation that the spectro-temporal filters can be treated as special types of neurons, and so the standard backpropagation training algorithm of ANNs can be extended to the feature extraction step as well. We experimented with neurons that simulated three types of spectro-temporal feature extraction methods. In the first case, a set of 2D DCT filters was applied; in the second we applied Gabor filters; while for the third configuration we used randomly generated filters. What we found was that in each case, our integrated method enhanced the performance of the filter sets by extending the scope of the backpropagation algorithm to the neurons simulating them.

In this study, we improve our system further by incorporating recent advances in neural networks into it. In the standard ANN implementations there are three layers, namely an input layer, an output layer (applying the softmax nonlinearity), and in between a hidden layer that uses a sigmoid activation function. Recently, it has been shown that significant improvements in performance can be achieved by increasing the number of hidden layers [5]. Unfortunately, training these 'deep' networks with many (three or more) hidden layers using the classic backpropagation algorithm has certain problems associated with it. A solution to these problems was given by Hinton et al., leading to a renaissance of ANN-based technologies in speech [5]. An even simpler solution was later given with the introduction of rectifier neural networks [6]. Here, we apply the latter in combination with deep neural nets to the model introduced in our earlier paper [4]. By evaluating our system in phone recognition tasks on the widely used TIMIT speech database, we will show that these techniques allow us to further improve the model performance in the case of clean speech and also noise contaminated speech. It is important to note that the training and testing was performed on different portions of the database, and that to obtain phone recognition results in noisy environment, we did not do any additional training of the models, but used the models that were trained on clean speech data.

Acknowledgements

This publication is supported by the European Union and co-funded by the European Social Fund. Project title: Telemedicine-focused research activities in the fields of mathematics, informatics and medical sciences. Project number: TÁMOP-4.2.2.A-11/1/KONV-2012-0073.

References

[1] Kleinschmidt M. Robust Speech Recognition Based on Spectro-Temporal Processing. PhD thesis at Carl von Ossietzky University, Oldenburg, Germany, 2002.

[2] Bouvrie, J., Ezzat, T., and Poggio, T. Localized spectro-temporal cepstral analysis of speech. *Proceedings of ICASSP* pp. 4733–4736, 2008.

[3] Kovács, G., Tóth, L. Phone Recognition Experiments with 2D DCT Spectro-Temporal Features. *Proceedings of SACI*, pp. 143–146, 2011.

[4] Kovács, G., Tóth, L. The Joint Optimization of Spectro-Temporal Features and Neural Net Classifiers. *Proceedings of TSD*, pp. 552–559, 2013.

[5] Hinton, G. et al. Deep Neural Networks for Acoustic Modeling in Speech Recognition. *in IEEE Signal Processing Magazine*, pp. 82–97, 2012.

[6] Tóth, L. Phone Recognition with Deep Sparse Rectifier Neural Networks. *Proceedings of ICASSP*, pp. 6985–6969, 2013.