MULTI-BAND PROCESSING WITH GABOR FILTERS AND TIME DELAY NEURAL NETS FOR NOISE ROBUST SPEECH RECOGNITION

György Kovács[1]*, László Tóth*[2]*, Gábor Gosztolya*[1]

[1]MTA-SZTE Research Group on Artificial Intelligence, Szeged, Hungary
[2]Institute of Informatics, University of Szeged, Szeged, Hungary
{gykovacs, tothl, ggabor}@inf.u-szeged.hu

ABSTRACT

Spectro-temporal feature extraction and multi-band processing were both invented with the goal of increasing the robustness of speech recognisers. However, although these methods have been in use for a long time now, and they are evidently compatible, few attempts have been made to combine them. This is why here we investigate the combination of multi-band processing with the use of spectro-temporal Gabor filters. First, based on the TIMIT corpus, we optimise their meta-parameters like the overlap, and the number of bands. Then we verify the cross-corpus viability of our multi-band processing approach on the Aurora-4 corpus. Lastly, we combine our method with the recently proposed channel dropout method. Our results show that this combination not only leads to lower error rates than those got using either multi-band processing or channel dropout, but these results compare favourably to those recently reported for the clean training scenario on the Aurora-4 corpus.

Index Terms— Multi-band processing, Gabor filters, Time delay neural nets, TIMIT, Aurora-4

1. INTRODUCTION

Despite the steady progress made in Automatic Speech Recognition (ASR), computer systems have only recently approached Human Speech Recognition (HSR) performance [1, 2], and clearly fall behind when noise is introduced [3, 4, 5, 6]. This motivated researchers to develop a closer collaboration between the two areas [7, 8, 9, 10], leading to the advent of methods that seek to improve ASR performance based on an analysis of auditory processing.

In this study we combine two of these methods in an attempt to create a system that is more robust to additive noise. We do so without presuming any knowledge about the type of noise the trained model would encounter. Thus here we examine the noise robustness of models that were trained exclusively on clean speech.

László Tóth was supported by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences and the UNKP-18-4 New National Excellence Programme of the Ministry of Human Capacities.

Spectro-temporal processing is one technique used in our combination. It was motivated by experiments confirming the sensitivity of auditory neurons to localised spectro-temporal modulation in grassfrogs [11] as well as in other animals [12, 13]. Later, experiments also indicated that auditory processing works with spectro-temporal patterns, rather than performing steps of spectral and temporal filtering consecutively [14]. Another motivation for this method was purely practical: unlike in spectrally global analysis techniques, like MFCC, in spectro-temporal processing a band-limited noise does not contaminate all the features [15]. These results lead to the idea of spectro-temporal processing [16], which seeks to capture spectro-temporal modulations in speech by processing it using filters that are localised in both the time domain and the frequency-domain. One popular spectro-temporal processing method is extracting spectro-temporal features from speech by applying the two-dimensional discrete cosine transform (2D DCT) on localised patches of its spectral representation [15, 17]. Another common method for spectro-temporal processing is the application of the real part of Gabor filters (a product of a two-dimensional Gaussian and an oriented sinusoid) on patches of the spectrogram [18]. In an earlier study [19] we experimented with both methods concerning multi-band processing. Our results on the TIMIT corpus indicated that Gabor filters are more suitable in the given framework for the task of ASR than 2D DCT coefficients. Because of this, in this study we will carry out our experiments using a set of Gabor filters (see Fig. 1) introduced by Kovács et al. in 2015 [20].

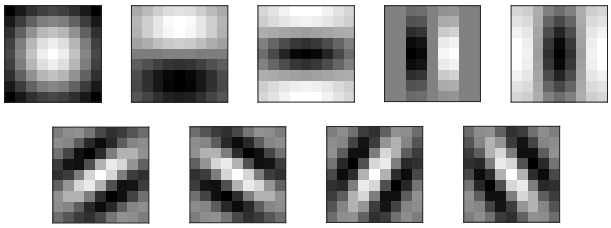

Fig. 1. A set of Gabor filters with size 9 by 9

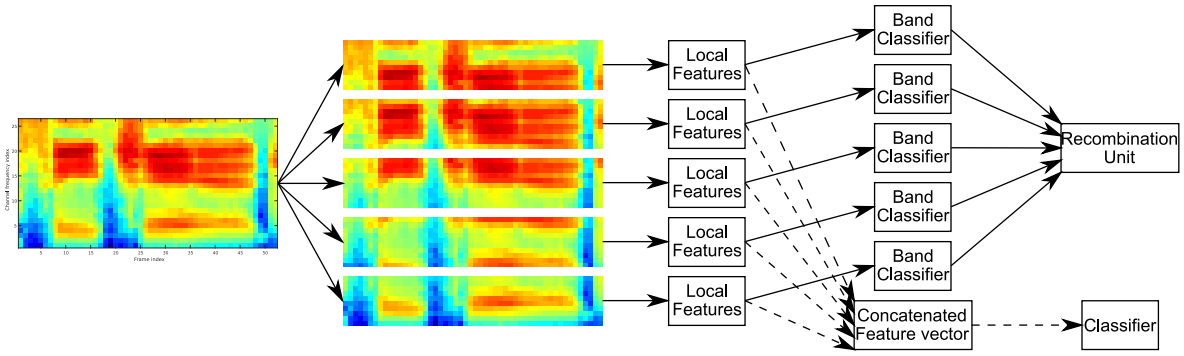

Fig. 2. An illustration of the multi-band processing approach (where local features from different bands are processed separately), and the feature recombination approach (where local features from different bands are processed together).

Multi-band processing was also motivated by auditory processing, and the pursuit of noise-robust ASR [21, 22, 23]. In this paradigm the input is decomposed into separate bands, which are then processed independently (this usually entails a partial recognition being performed [22, 24, 25, 26]). Then, to produce an overall recognition result, information from the separate bands is recombined. Thus two key issues in multi-band speech processing are the method used for separating the input into separate bands, and the method used for combining the information from these bands. Several methods exist for addressing the latter issue, from simple fixed linear combinations [22, 24] to sophisticated methods that try to dynamically assess the reliability of the bands [21, 23, 27, 28, 29], including neural nets [30, 31]. As this straightforward solution proved to be successful in our earlier experiments [19], here we applied deep neural nets (DNNs) for this task, without performing any explicit reliability assessment for the bands. Regarding the formation of bands we should mention that spectro-temporal processing already provides us with an answer to this issue. When processing our input with spectro-temporal feature extraction methods (e.g. Gabor filters), in each time-interval we get our features from various frequency ranges. Hence, by separating the features based on the frequency range they originate from, they naturally form separate bands (see Fig. 2).

Here, following our earlier study [19], we examine the combination of multi-band processing and spectro-temporal feature extraction with Gabor filters for improved noise-robustness of ASR. In our investigation of the combined process, we examine the effect of various parameters on the recognition results, like the overlap of the bands and the number of bands used. We also investigate another approach of multi-stream processing, where each stream is formed using features from all but one band. Then, we combine the best performing method with channel dropout [32], which drops a random number of bands during the training of the recombinational net, preventing it from relying too much on certain bands.

2. EXPERIMENTAL SETUP

2.1. TIMIT corpus

Despite its relatively small size, the TIMIT corpus [33] still has an important role in testing new ideas, as improvements achieved on it can be scaled up to larger corpora [34]. When using it, we followed the standard partitioning, namely the train set of 3696 sentences, the core test set of 192 sentences, and the remaining sentences taken from the full test set as the development set. To create a phoneme recogniser, we used a HMM/ANN hybrid and a simple bigram language model. The labels here were 858 triphone tri-state phoneme models that had been collapsed into 39 categories for evaluation, as has become the standard [35].

As our main goal was to evaluate the noise robustness of our models (trained by using just clean speech), we also evaluated them on the noise contaminated versions of the core test set we created using the FaNT tool [36]. For this, we created bandlimited noise by filtering white noise with a bandpass filter active between 3 kHz and 5 kHz. We also used three noise samples from the NOISEX-92 database [37], namely babble noise, volvo noise, and the factory-1 noise sample.

2.2. Aurora-4 corpus

We also evaluated our proposed methods on the Aurora-4 corpus [38]. This database contains two training sets, namely the clean set and the multi-condition set, both consisting of 7138 utterances. As our whole concept is based on the assumption that we have no noisy training data, and our goal was to create a noise-robust model under these conditions, here we just use the clean training set, consisting of clean data collected from the Sennheiser microphone. The test set of Aurora-4 consists of 4620 utterances, with a subset recorded using a Sennheiser close-talking microphone, and a subset recorded using a set of secondary microphones. Both of these subsets contain a clean subset and a subset consisting of the noise-corrupted versions

of the same utterances. The final subsets are called test set A (clean recordings with the Sennheiser microphone), set B (noise-corrupted version of set A), set C (clean recordings with secondary microphones), and set D (noise-corrupted version of set C).

When working with the Aurora-4 corpus we first used Kaldi's [39] recipe to train a HMM/GMM model. We then performed forced alignment with this model, and utilised the acquired 1997 frame-level context-dependent state labels as training targets for our in-house DNN implementantion. We trained these DNNs with backpropagation using the frame-level cross-entropy error function. A random 10% of the training set was held out as the development set used for early stopping. Decoding was performed with Kaldi, using the standard tri-gram language model and the 5k word vocabulary.

2.3. Time-frequency processing

As the initial time-frequency representation we chose the log-mel scaled spectrogram with 45 channels that we computed with the HTK Toolkit [40] using 400 samples (25 ms) per frame at 160 sample (10 ms) hops, and applying a 1024-point FFT on the frames. Next, each sentence was normalised so as to give a zero mean and unit variance.

2.4. Band formation

We applied our set of Gabor filters on the resulting spectrograms to acquire spectro-temporal features. Then we calculated the Δ and $\Delta\Delta$ coefficients for each of these features, and separated the resulting features into sets based on the requirements of the current method. This process had three important parameters, namely:

- Multi-band or Leave-1-band-out: As an alternative to multi-band processing, where each band classifier has access to features from just one band, similar classifiers can be used in a way where each one has access to features from all but one band.
- Overlap of filters: In our earlier experiments [17, 20] we applied Gabor filters on the spectral representation with an overlap of 55%. Here, we validate whether this is indeed beneficial in the multi-band framework.
- Number of bands: When applying Gabor filters with no overlap, features originate from five separate positions in the frequency domain, defining five bands. However, when applying Gabor filters with a 55% overlap, we get features from ten positions in the frequency domain. Thus we can separate them into five equal bands as we did earlier (by grouping together features from different positions), but we can also separate them into ten bands. We will examine which setting is more advantageous based on our results on the TIMIT corpus.

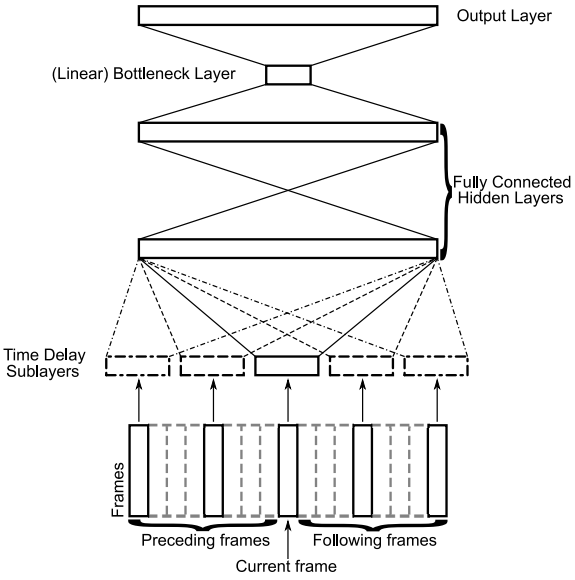

Fig. 3. Structure of a band classifier TDNN

2.5. Band classifier

To process the individual bands, we employed time delay neural nets [41] extended with sub-sampling [42] (see Fig. 3), the meta-parameters of which were based on findings of our earlier experiments [43, 44] [1]. Here, the input is first processed by a smaller layer at 5 positions, using subsampling with a gap of 3 frames. This layer is depicted in Fig. 3 as five sub-layers, as it takes its input from five positions, and the dimensionality of its output is five times its size. This layer is followed by traditional fully connected hidden layers with rectifier units. This is followed by a bottleneck layer providing the input for the recombinational net (its size is given in such a way that the number of inputs for the recombinational net is always 200). In this layer the neurons apply a linear activation function, as with a rectifier activation function many input features of the subsequent neural net had a value of zero for all examples from the training set. Lastly, we had an output layer with 858 (TIMIT) or 1997 (Aurora-4) softmax neurons.

2.6. Recombinational network

For the combination of information from separate bands we used simple fully connected Deep Rectifier Neural Nets (DRNs) with three hidden layers, each containing a thousand neurons. The output layer contained 858 (TIMIT) or 1997 (Aurora-4) softmax neurons (more details on the parameters of the neural nets can be found in tables 1 and 4). For each method we trained 10 such networks in our experiments, and reported their average performance.

[1] For comparability reasons, this model was also applied when we did not invoke multi-band processing. We treated this case as if we had just one band.

Table 1. Key meta-parameters of the neural nets used.

		No overlap			Overlap				
		FC	L1	MB	FC	L1	MB	L1-10	MB-10
Band classifier layers	Time-delay	200	200	200	200	200	200	140	140
	Hidden	$3 \cdot 2000$	$2 \cdot 1000$	$2 \cdot 1000$	$3 \cdot 2000$	$2 \cdot 1000$	$2 \cdot 1000$	$2 \cdot 700$	$2 \cdot 700$
	Bottleneck	200	40	40	200	40	40	20	20
	No. of bands	1	5	5	1	5	5	10	10
	Param. no.	10.4M	10.3M	10.2M	10.5M	10.4M	10.3M	10.3M	10.0M
Recomb. net	Neigh. no.	9	9	9	9	9	9	9	9
	Param. no.	3.8M	3.8M	3.8M	3.8M	3.8M	3.8M	3.8M	3.8M
Max. param. no.		14.2M	14.1M	14.0M	14.2M	14.2M	14.0M	14.1M	13.8M

3. EXPERIMENTS ON THE TIMIT CORPUS

Here, based on their phoneme recognition accuracy scores on the TIMIT corpus, we compare three different approaches, namely the Feature reCombination approach (FC), where spectro-temporal features are concatenated, and treated as one band; the Multi-Band (MB) approach, where each band is processed separately, and the Leave-1-out approach (L1), where from each stream only one band is excluded. To present the best available baseline, for the FC approach we evaluated both the band classifier and the recombinational network, and reported the lower error rates of the two.

3.1. Experiments with no overlap

The results are listed in Table 2. We observe that in clean speech both the MB and L1 approaches significantly outperform the baseline. We can see the same results in band-limited noise, which is hardly surprising, given that one motivation behind the use of multi-band processing was its supposed robustness to this specific noise type. What is of more interest is the performance with real-life noise. Here, on average the MB framework's performance is significantly better than that of the two other methods.

3.2. Experiments with overlap

When examining the results of applying the same methods attained with overlapping filters (methods FC, L1, and MB in Table 2), we observe that the overlap we used in the extraction of spectro-temporal features was beneficial in both clean speech and speech contaminated with real-life noise types. When comparing the results of these three methods with each other, we can see that the MB approach performs the best in both clean speech, and noise contaminated speech (regardless of whether the noise was artificial or real-life).

In the case of the L1 and the MB approaches it is also interesting to see, how the error rates change when using 10 bands instead of the 5 we used previously. We observe in the rightmost columns of Table 2 that in the Leave-1-out approach using 10 bands in most cases leads to higher error rates. The opposite is true for the multi-band approach, where in most cases this change decreases the error rates obtained, and in many of these cases this decrease is significant. Overall we can say that in noise contaminated speech the results got with the MB-10 approach were significantly better (or at least not significantly worse) than results got with any other approach used. For clean speech the MB approach performed significantly better than the FC and the Leave-1-out schemes.

Table 2. Phoneme Error Rates (PER %) on TIMIT. On each line the best result for the overlapping and the non-overlapping scenarios (and the results not significantly different from it for $p < 0.005$) is shown in bold.

Settings	Noise type	SNR	No overlap			Overlap				
			FC	L1	MB	FC	L1	MB	L1-10	MB-10
Clean	–	–	20.0	**19.4**	**19.6**	19.7	19.3	**18.7**	20.0	19.4
Artificial Noise	Band-limited	10 dB	45.4	43.5	**33.7**	45.9	45.1	34.9	44.6	**32.6**
		20 dB	32.2	31.0	**27.2**	32.2	31.1	27.0	31.5	**25.7**
Real-life Noise	Babble noise	10 dB	48.9	49.4	**47.0**	48.3	48.9	48.5	49.5	**45.2**
		20 dB	28.8	28.4	**27.8**	28.0	27.8	27.4	28.9	**26.8**
	Factory noise	10 dB	**49.9**	50.6	50.8	**48.8**	**49.2**	49.4	50.0	**49.0**
		20 dB	31.1	30.4	**30.0**	30.2	29.9	**29.0**	31.2	**29.1**
	Volvo noise	10 dB	25.2	**24.5**	**24.6**	24.0	23.3	**22.9**	23.8	**22.9**
		20 dB	21.7	**21.2**	**21.3**	21.1	**20.6**	**20.5**	20.9	**20.6**
	Average		34.3	34.1	**33.6**	33.4	33.3	32.9	34.1	**32.3**

Table 3. Phoneme Error Rates (PER %) on TIMIT. On each line the best result (and the results not significantly different from it for $p < 0.005$) is shown in bold.

Settings	Noise type	SNR	Overlap		
			[43]	[19]	MB-10++
Clean	–	–	23.0	22.8	**19.0**
Artificial Noise	Band-limited	10 dB	49.8	35.9	**32.5**
		20 dB	38.1	29.6	**25.5**
Real-life Noise	Babble noise	10 dB	49.7	46.7	**44.7**
		20 dB	31.8	31.1	**26.4**
	Factory noise	10 dB	54.9	53.2	**49.0**
		20 dB	34.8	33.8	**28.8**
	Volvo noise	10 dB	27.5	25.8	**22.4**
		20 dB	24.8	23.9	**20.1**
Average			37.3	35.7	**31.9**

3.3. Comparison with earlier results

Meta-parameters of the recombinational network (i.e. number of neighbouring frames used, and phoneme insertion penalty) so far were also based on our earlier studies [43, 44]. This was sufficient when the aim was simply to compare the results of our methods with each other. However, when comparing our results with those reported by other studies, it is reasonable that we should optimise these meta-parameters as well. Because of this, we used the development set of the TIMIT corpus to optimise the number of neighbouring frames used in our MB-10 settings, as well as the phoneme insertion penalty applied during the decoding phase. We found that the best results can be obtained using 13 neighbouring frames, and a phoneme insertion penalty of -0.5. We will refer to this setting as the MB-10++ approach. We also evaluated this approach using the various noise-contaminated versions of the core test set, and compared the resulting scores with results found in the literature for the same noise types that we used in our experiments here.

As not many researchers have published results on noise contaminated versions of the TIMIT corpus (especially with the particular noise types and signal to noise ratios we used in our study), we chose two bases of comparisons here. One was the joint-training framework of [43], where the neurons implementing spectro-temporal filtering in the framework were initialised based on the same set of Gabor filters we applied here. The other basis of comparison was our earlier study on the combination of multi-band processing and Gabor filters [19]. The results reported in these studies are listed in Table 3, along with the results we got using the optimised version of the MB-10 approach (MB-10++). We observe here that both for the clean speech and for the speech contaminated with each kind of noise, the approach published here provided significantly better results than those reported earlier.

Table 4. Key meta-parameters of the neural nets used.

		Overlap		
		FC	MB-10-	MB-10⋆
Band classifier layers	Time-delay	200	100	200
	Hidden	$3 \cdot 2000$	$2 \cdot 500$	$2 \cdot 1000$
	Bottleneck	200	20	20
	No. of bands	1	10	10
	Param no.	10.9M	5.1M	20.3M
Recomb. net	Neigh. no.	–	1	13
	Param. no.	–	3.3M	5.7M
Param. no.		10.9M	8.4M	26.0M

4. EXPERIMENTS ON THE AURORA-4 CORPUS

We set out to repeat experiments using the best performing baseline (FC + overlap), and the best performing multi-band setup (MB-10++). Our early experiments on Aurora-4, however, revealed that in the FC setup the band classification net performs significantly better than the recombinational net. Because of this, we first compared the results of this network with those produced by a modified version of MB-10 (MB-10-) that more closely matches the band classifier net of FC regarding context size and total number of parameters used. For this we modified the MB-10 setup by decreasing the number of neurons in its band classifier nets, and the number of neighbouring frames used in its recombinatonal net. Then, to evaluate how well the method can perform without these constraints, we doubled the number of neurons in the layers of the band classifiers, and returned to the use of 13 neighbouring frames in the recombinational net.

The results of these experiments are listed in Table 5. We observe that the multi-band setup with the fewer parameters (MB-10-) already provides an overall 17% relative error rate reduction over our baseline (FC). And when examining the results broken down into the four subsets, we see that this approach significantly improves the results in 3 out of 4 subsets. And while with FC we get slightly lower error rates on test set A, after increasing the number of parameters in the multi-band framework we can match this performance, while further improving the performance on all other test sets. Once again, these improvements were significant with $p < 0.005$.

Table 5. Word Error Rates (WERs %) on the Aurora-4 corpus. The best result (and results not significantly different from it for $p < 0.005$) is shown in bold.

Data set	FC	MB-10-	MB-10⋆
Set A	**3.1**	3.7	**3.1**
Set B	20.2	15.9	**15.5**
Set C	35.3	29.4	**27.8**
Set D	49.7	41.9	**40.8**
Overall	33.9	28.0	**27.2**

Table 6. Word Error Rates (WERs %) on the Aurora-4 corpus. The best result (and results not significantly different from it – $p < 0.005$) is shown in bold.

		MB-10⋆	MB-10⋆	MB-10⋆
Additional layer		–	✓	✓
Channel dropout		–	–	✓
Data set	Set A	**3.1**	3.3	3.4
	Set B	15.5	15.3	**14.6**
	Set C	27.8	25.9	**23.9**
	Set D	40.8	39.9	**37.4**
Overall		27.2	26.6	**25.0**

Table 7. Comparison of our best result with some recent results given in the literature for Aurora-4, using the clean training set.

Method	WER
CNN with FBANK features [45]	28.9%
DNN with exemplar-based enhancement [46]	26.8%
CNN with channel dropout [32]	26.8%
CNN with data augmentation [47]	25.6%
GMM with auditory spectral enhancement [48]	25.5%
TDNN with Gabor filters, multi-band processing, and channel dropout	**25.0%**

4.1. Combining multi-band processing and channel dropout

Earlier, Kovács et al. [32] introduced channel dropout, a method based on input-dropout that drops out entire frequency bands during training, resulting in a network that is more robust to noise. It is apparent that this method can also be applied here in the training of our recombinational nets. By dropping out inputs resulting from certain band classifiers we expect that this would prevent the neural net model trained to rely heavily on a few preferred bands. In order to apply this method here, we first introduced an additional layer (with 2000 neurons) into the recombinatonal net, which is divided equally into 10 sublayers (each sublayer containing 200 neurons), where each sublayer processes the input from one band classifier net. We implemented the channel dropout method by applying the input dropout on the inputs of these sublayers. We should mention here that as the added layer is not fully connected, and while its addition increases the number of neurons in the network, the number of connections decrease, hence the number of trainable parameters in the network does not actually increase.

We first trained ten neural nets using this architecture without channel dropout, and then trained ten more, but using the same meta-parameters for channel dropout (in each batch dropping out a maximum of 6 channels with a probability of 60%) that had been used in [32]. The results of these experiments are listed in Table 6. Interestingly, by simply adding an extra layer where different bands are processed separately, we managed to significantly reduce the overall error rates, and also reduce the error rate on 3 out of the 4 subsets (two of these improvements being significant), while only slightly increasing the error rate in clean speech.

Furthermore, when we applied the channel dropout with the preset parameters, we managed to attain a further relative error rate reduction of 6% overall, leading to a relative error rate reduction compared with the original setting (i.e. with no additional layer) of 8%. And although there was again a slight increase in the error rates got for clean speech, in all other settings the improvement in the recognition scores was significant. Next, when compared with our baseline, we got an overall error rate reduction of more than 26%.

Lastly, we compare our results with those found in the literature for the same task (i.e. evaluation on the Aurora-4 when the models were trained on the clean training set, using no noise samples). As can be seen in Table 7, our method is competitive with most of the recently published solutions. We would also like to state, that while the channel dropout method already provided competitive results in itself, when combined with multi-band processing, we managed to improve its score by more than 6%. What is more, our method presented here outperformed the data agumentation method, which is an improved version of the original channel dropout method.

5. CONCLUSIONS AND FUTURE WORK

Here, we investigated several approaches for the combination of multi-band processing and spectro-temporal feature extraction using Gabor filters. Our experiments on the TIMIT corpus confirmed the benefits of multi-band processing. Using the results of these experiments we also found that having overlapping bands can also increase the accuracy of our framework, and increasing the number of bands had the same effect. We also examined an alternative approach to multi-band processing, where instead of training separate classifiers for each band, we train separate classifiers for ensembles that consist of all but one band. We found that while in some cases it can lead to an improvement over the baseline case of processing all features from all bands together, these improvements are less marked than those that can be achieved using the other alternative examined here. Experiments on the Aurora-4 corpus again confirmed the utility of the multi-band approach, and showed that when combined with channel dropout, it can produce competitive results.

In the future we would like to extend our experiments to other spectral representations (like the Power Normalised Spectrogram), and other speech corpora. We would also like to experiment with different structures for the recombinational neural net.

6. REFERENCES

[1] Bernd T. Meyer, Thorsten Wesker, Thomas Brand, Alfred Mertins, and Birger Kollmeier, "A human-machine comparison in speech recognition based on a logatome corpus," in *Proc. SRIV*, 2006.

[2] George Saon, Gakuto Kurata, Tom Sercu, Kartik Audhkhasi, Samuel Thomas, Dimitrios Dimitriadis, Xiaodong Cui, Bhuvana Ramabhadran, Michael Picheny, Lynn-Li Lim, Bergul Roomi, and Phil Hall, "English conversational telephone speech recognition by humans and machines," in *Proc. Interspeech*, 2017, pp. 132–136.

[3] Hari Krishna Maganti and Marco Matassoni, "Auditory processing-based features for improving speech recognition in adverse acoustic conditions," *EURASIP J AUDIO SPEE*, vol. 2014, no. 1, pp. 21, 2014.

[4] Bernd T. Meyer, "What's the difference? Comparing humans and machines on the Aurora2 speech database," in *Proc. Interspeech*, 2013, pp. 2634–2638.

[5] Jason J. Sroka and Louis D. Braida, "Human and machine consonant recognition," *Speech Communication*, vol. 45, no. 4, pp. 401 – 423, 2005.

[6] Bernd T. Meyer, Matthias Wächter, Thomas Brand, and Birger Kollmeier, "Phoneme confusions in human and automatic speech recognition," in *Proc. Interspeech*, 2007.

[7] Bob Carpenter, "Human versus machine: Psycholinguistics meets ASR," in *Proc. ASRU*, 1999, pp. 225–228.

[8] Hynek Heřmanský, "Human speech perception: Some lessons from automatic speech recognition," in *Proc. TSD*, 2001, pp. 187–196.

[9] Bernd T. Meyer, *Human and automatic speech recognition in the presence of speech-intrinsic variations*, Ph.D. thesis, Carl-von-Ossietzky Universitt, 2009.

[10] Odette Scharenborg, "Reaching over the gap: A review of efforts to link human and automatic speech recognition research.," *Speech Communication*, vol. 49, no. 5, pp. 336–347, 2007.

[11] Ad M. Aertsen and Peter I. M. Johannesma, "Spectro-temporal receptive fields of auditory neurons in the grassfrog," *Biol. Cybern.*, vol. 38, no. 4, pp. 223–234, November 1980.

[12] Taishih Chi, Powen Ru, and Shihab A. Shamma, "Multiresolution spectrotemporal analysis of complex sounds," *J. Acoust. Soc. Am.*, vol. 118, no. 2, pp. 887–906, 2005.

[13] Frédéric E. Theunissen, Kamal Sen, and Allison J. Doupe, "Spectral-temporal receptive fields of nonlinear auditory neurons obtained using natural sounds," *J. Neurosci.*, vol. 20, pp. 2315–2331, March 2000.

[14] Didier A. Depireux, Jonathan Z. Simon, David J. Klein, and Shibab A. Shamma, "Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex," *J. Neurophysiol.*, vol. 85, pp. 12201234, 2001.

[15] Jake Bouvrie, Tony Ezzat, and Tomaso Poggio, "Localized spectro-temporal cepstral analysis of speech," in *Proc. ICASSP*, 2008, pp. 4733–4736.

[16] Michael Kleinschmidt, *Robust Speech Recognition Based on Spectro-Temporal Processing*, Ph.D. thesis, Carl-von-Ossietzky Universitt Oldenburg, 2002.

[17] György Kovács and László Tóth, "Phone recognition experiments with 2D-DCT spectro-temporal features," in *Proceedings of the International Symposium on Applied Computational Intelligence and Informatics*, 2011, pp. 143–146.

[18] Tony Ezzat, Jake V. Bouvrie, and Tomaso A. Poggio, "Spectro-temporal analysis of speech using 2D Gabor filters," in *Proc. Interspeech*, 2007, pp. 506–509.

[19] György Kovács, László Tóth, and Tamás Grósz, "Robust multi-band ASR using Deep Neural Nets and spectro-temporal features," in *Proc. SPECOM*, 2015, pp. 386–393.

[20] György Kovács, László Tóth, and Dirk Van Compernolle, "Selection and enhancement of Gabor filters for automatic speech recognition," *IJST*, vol. 18, no. 1, pp. 1–16, 2015.

[21] Hervé Bourlard and Stéphane Dupont, "A new ASR approach based on independent processing and recombination of partial frequency bands," in *Proc. ICSLP*, 1996, pp. 426–429.

[22] Hynek Heřmanský, Sangita Timbrewala, and Misha Pavel, "Towards ASR on partially corrupted speech," in *Proc. ICSLP*, 1996, pp. 464–465.

[23] Shigeki Okawa, Enrico Bocchieri, and Alexandros Potamianos, "Multi-band speech recognition in noisy environments," in *Proc. ICASSP*, 1998, pp. 641–644.

[24] Christophe Cerisara and Dominique Fohr, "Multi-band automatic speech recognition," *Computer Speech and Language*, vol. 15, no. 2, pp. 151–174, 2001.

[25] Astrid Hagen, Andrew Morris, and Hervé Bourlard, "Subband-based speech recognition in noisy conditions the full combination approach," Tech. Rep. Idiap-RR-15-1998, IDIAP, 1998.

[26] Adam Janin, Dan Ellis, and Nelson Morgan, "Multi-stream speech recognition: ready for prime time?," in *Proc. Eurospeech*, 1999, pp. 591–594.

[27] Astrid Hagen, Hervé Bourlard, and Andrew Morris, "Adaptive ML-weighting in multi-band recombination of Gaussian mixture ASR," in *Proc. ICASSP*, 2001, pp. 257–260.

[28] Nima Mesgarani, Samuel Thomas, and Hynek Heřmanský, "A multistream multiresolution framework for phoneme recognition," in *Proc. Interspeech*, 2010, pp. 318–321.

[29] Andrew Morris, Astrid Hagen, Hervé Glotin, and Hervé Bourlard, "Multi-stream adaptive evidence combination for noise robust ASR," *Speech Communication*, vol. 34, no. 1-2, pp. 25–40, 2001.

[30] Nikki Mirghafori, *A Multi-Band Approach to Automatic Speech Recognition*, Ph.D. thesis, International Computer Science Institute, 1999.

[31] Nikki Mirghafori and Nelson Morgan, "Combining connectionist multi-band and full-band probability streams for speech recognition of natural numbers," in *Proc. ICSLP*, 1998, pp. 743–746.

[32] György Kovács, László Tóth, Dirk Van Compernolle, and Sriram Ganapathy, "Increasing the robustness of CNN acoustic models using autoregressive moving average spectrogram features and channel dropout," *Pattern Recognition Letters*, vol. 100, pp. 44–50, 2017.

[33] Lori F. Lamel, Robert H. Kassel, and Stephanie Seneff, "Speech database development: design and analysis of the acoustic-phonetic corpus," in *Proc. DARPA Speech Recognition Workshop, Report no. SAIC-86/1546*, 1986.

[34] Tara N. Sainath, Bhuvana Ramabhadran, and Michael Picheny, "An exploration of large vocabulary tools for small vocabulary phonetic recognition," in *Proc. ASRU*, 2009, pp. 359–364.

[35] Carla Lopes and Fernando Perdigao, "Phoneme recognition on the TIMIT database," in *Speech Technologies*, Prof. Ivo Ipsic, Ed. InTech, 2011.

[36] Hans-Günter Hirsch, "Fant: Filtering and noise-adding tool," http://dnt.kr.hs-niederrhein.de/download.html.

[37] Andrew Varga and Herman J. M. Steeneken, "Assessment for automatic speech recognition ii: Noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Commun.*, vol. 12, no. 3, pp. 247–251, July 1993.

[38] Hans-Günter Hirsch and David Pearce, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *ASR2000-Automatic Speech Recognition: Challenges for the new Millenium ISCA Tutorial and Research Workshop (ITRW)*, 2000.

[39] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely, "The Kaldi speech recognition toolkit," in *Proc. ASRU*, 2011.

[40] Steve Young, Gunnar Evermann, Mark Gales, Thomas Hain, Dan Kershaw, Xunying A. Liu, Gareth Moore, Julian Odell, Dave Ollason, Dan Povey, Valtcho Valtchev, and Phil Woodland, *The HTK Book (for HTK Version 3.4)*, Cambridge University Engineering Department, 2006.

[41] Alexander Waibel, Toshiyuki Hanazawa, Geoffrey Hinton, Kiyohiro Shikano, and Kevin J. Lang, "Phoneme recognition using time-delay neural networks," *IEEE Trans. ASSP*, vol. 37, no. 3, pp. 328–339, 1989.

[42] Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Proc. Interspeech*, 2015, pp. 3214–3218.

[43] György Kovács and László Tóth, "Joint optimization of spectro-temporal features and Deep Neural Nets for robust automatic speech recognition," *Acta Cybernetica*, vol. 22, no. 1, pp. 117–134, 2015.

[44] György Kovács and László Tóth, "Multi-band noise robust speech recognition using Deep Neural Networks (in Hungarian)," in *Proc. MSZNY*, 2016, pp. 287–294.

[45] Jui-Ting Huang, Jinyu Li, and Yifan Gong, "An analysis of Convolutional Neural Networks for speech recognition," in *Proc. ICASSP*, 2015, pp. 4989–4993.

[46] Deepak Baby, Jort F. Gemmeke, Tuomas Virtanen, and Hugo Van hamme, "Exemplar-based speech enhancement for Deep Neural Network based automatic speech recognition," in *Proc. ICASSP*, 2015, pp. 4485–4489.

[47] László Tóth, György Kovács, and Dirk Van Compernolle, "A perceptually inspired data augmentation method for noise robust cnn acoustic models," 2018.

[48] Md Jahangir Alam, Patrick Kenny, and Douglas O'Shaughnessy, "Robust feature extraction based on an asymmetric level-dependent auditory filterbank and a subband spectrum enhancement technique," *Digital Signal Processing*, vol. 29, pp. 147 – 157, 2014.