A long, deep and wide artificial neural net for robust speech recognition in unknown noise

Feipeng Li, Phani S. Nidadavolu, and Hynek Hermansky

Center for Language and Speech Processing
Johns Hopkins University, Baltimore, MD, 21218

Abstract

A long deep and wide artificial neural net (LDWNN) with multiple ensemble neural nets for individual frequency subbands is proposed for robust speech recognition in unknown noise. It is assumed that the effect of arbitrary additive noise on speech recognition can be approximated by white noise (or speech-shaped noise) of similar level across multiple frequency subbands. The ensemble neural nets are trained in clean and speech-shaped noise at 20, 10, and 5 dB SNR to accommodate noise of different levels, followed by a neural net trained to select the most suitable neural net for optimum information extraction within a frequency subband. The posteriors from multiple frequency subbands are fused by another neural net to give a more reliable estimation. Experimental results show that the subband ensemble net adapts well to unknown noise.

Index Terms: structured deep neural networks, multi-stream recognition of speech, ensemble neural net, noise robustness

1. Introduction

How to deal with unknown noise in realistic environments is a long-lasting problem for automatic speech recognition (ASR). Today most ASR systems use artificial neural net (ANN) for front-end processing. It works well when the test condition matches the training condition, but it deteriorate quickly when the speech is corrupted by noise. A common approach to improve system robustness is to train the neural net in all expected conditions [8, 4], which generally gives a system of better generalization ability but more mediocre performance in matched conditions. For example, a neural net trained on both clean speech and noisy speech at 0 dB SNR white noise may show better performance for an unseen condition of noisy speech at 10 dB SNR white noise than two neural nets trained in only one of the two conditions, but it usually performs worse on the matched conditions. In many realistic environments, both the noise level and spectral shape may change with time. It is impractical to train a single neural net and expect it to show good performance for all noises.

The human auditory system, in contrast, show superb performance as well as adaptation ability in almost all noisy conditions. According to [1], the cochlear frequency range spans of about 20 articulatory bands from 0.3 to 8 kHz that act as a processing channels for speech recognition. Speech events are extracted by hundreds of inner hair cells in every articulatory band, and this information is then assembled in the central auditory system. The capacity of each processing channel is determined by the signal-to-noise ratio within the channel.

Inspired by the parallel processing in human speech perception, a multi-stream speech recognition system has been proposed, in which the full band speech is divided into multiple subbands [9, 10, 16]. Motivations and history of the multi-stream processing can be found in [18].

In the current study, a neural net is trained within each subband to classify the speech sound given the partial acoustic evidence. The speech information from multiple subbands is integrated by a set of fusion neural nets. To maximize the information being extracted from individual subbands in both clean and noisy conditions, we have conducted several experiments to optimise the training of subband neural nets. It is shown that a neural net trained in clean condition shows poor generalization in noise, while a neural nets trained in noise generally show poor performance in clean signs. A multi-style trained neural net maintains a balance between the two by sacrificing performance for generalization ability.

In this study we develop a LDWNN for robust speech recognition by replacing the single subband neural net of multi-stream system with an ensemble neural nets using relatively long spans of speech signal as their inputs, trained on clean signal and signal with various levels of noise. The subband ensemble neural nets appear to generalize well in unknown noise.

2. Approach

The LDWNN is developed based on the multi-stream framework for parallel processing. It is assumed that the effect of non-stationary additive noise on neural net training or testing can be simulated by an artificial noise of similar spectral shape. Fig. 1 shows how the arbitrary spectral shape of unknown noise is approximated by using white noise of the same level across multiple frequency subbands.

In speech recognition a neural net is trained to estimate a posteriori probability of sound classes given the acoustic evidence [15]. It performs the best when the distribution of testing data matches that of training data. When the speech sound is corrupted by noise, the mis-match between the training and testing data can cause the performance to drop quickly. Adding noise to training data helps to improve the generalization of neural net, but it reduces performance in matched conditions. The higher the noise level, the less a neural net can learn. In multi-style training the speech feature is mixed with different types of noise at various levels. It generally produces a system of mediocre performance, especially when the training noises dif-
subband classifiers at the current time and times
Karhunen-Loeve Transform (KLT). Reduced vectors from all
of subband ensemble nets is reduced to 25 by applying the
technique [3]. The dimensionality of the pre-sigmoid outputs
class Neural Network - Hidden Markov Model - (ANN - HMM)
various levels of noise.
is divided into multiple subbands, each has an ensemble net for
fusion in a sequence, and is "wide" because the full band speech
stages of neural nets for feature extraction, stream selection, and
probability of each speech sound, is "deep" as it includes three
using 500 ms signal spans for deriving the estimated posterior
from the first classification stage over 200 ms, thus effectively
for its initial subband classification and concatenates outputs
takes 300 ms long segments of the input signal as the evidence
to give a more reliable estimation. The system is "long" as it
is trained to fuse the speech information from multiple subbands
for its initial subband classification and concatenates outputs
for optimal information extraction. ANN-Q, ANN-L, ANN-M, and ANN-H are trained
in quiet, low, middle, and high level of noise to accommodate
time-varying unknown noise

2.2. Subband Ensemble Net

Figure 2 depicts the block diagram of subband ensemble net
for optimal information extraction from a frequency subband. A set of \( N \) neural nets, including ANN-Q trained in quiet and several others ANN-L/M/H trained in various levels of noise, are employed to cover all possible conditions. In this study, we use the training speech signal at clean, 20, 10, and 5 dB SNR because the neural net trained in noise seem to generalize well within \( \pm 5 \) dB SNR. Instead of using white noise, which has a much bigger masking effect in the high frequency than in the low frequency, babble noise is used because it has the
same long-term spectral shape as speech signal, and therefore
provides a balanced masking across all frequency subbands. For
any unknown noise, the posteriors of the neural net gives the
best performance is selected for each frequency subband.

Suppose the subband ensemble net includes \( SNR_s = \{ Q, L, M, H \} \) dB for the quantization of noise in a frequency subband. The training of subband ensemble net takes two steps.

Step 1: Train feed-forward ANN-Q/L/M/H

For each \( snr \) in \( SNR_s \)
- add speech-shaped noise to all utterances in clean training
  set at \( snr \) dB SNR;
- compute FDLP2 feature \( X \) for all utterances in the noisy training set;
- train ANN-$snr$s with FDLP2 feature \( X \) using back-propagation;

In order for the ANN selector learn to pick the best posteriors produced by the ANN of matched condition, a training set of noisy speech is created by adding noise to each utterance at Q/L/M/H level in a random order, so that at any time there is always an ANN working on the matched condition, while the location of ANN to be selected shift randomly across the four positions. To make sure that the prior probability is equal for all noises, a sequence of pseudo-random number with uniform distribution between 1 and \( N \) is used as the index of noise type to be applied at any given time during the training. Next,
the synthesized noisy speech are forward-passed through ANN-Q/L/M/H. The output of the four neural nets are concatenated and used as the feature for the training of ANN selector, a 3-layer MLP in this study.

Step 2: Train ANN selector

For each utterance \( u \) in the clean training set
- randomly pick a \( \text{snr} \) in SNRs with equal prob.;
- add speech-shaped noise to \( u \) at \( \text{snr} \) dB SNR;
- compute the FDLP2 feature \( x \) for utterance \( u \);
- forward-pass \( x \) through ANN-Q/L/M/H;
- concatenate ANN outputs \( y = [y_Q, y_L, y_M, y_H] \);
- pad temporal context \( z(t) = [y(t+10), y(t+5), y(t), y(t-5), y(t-10)] \);
- train a three-layer perceptron that maps \( z \) to \( l \) using back-propagation;

3. Experiments

Three experiments are conducted to test the proposed approach using the TIMIT speech corpus. The first experiment evaluates the maximum performance and generalization ability of subband ensemble net as the noise level increases from clean, 20, 15, 10, 5, and 0 dB SNR. The second experiment evaluates the performance of overall system after fusion. In the third experiment the LDWNN system is assessed by unknown noise.

The results are compared with those of the single-stream (SS) baseline using MFCC, PNCC, and FDLP feature and a multi-stream (MS) baseline system [12] with and without a performance monitor, denoted as MS and MS-PM respectively. Both MS systems have only one neural net in every subband. The subband neural nets are trained in either clean (denoted as clean) or both clean and noisy conditions (denoted as mixed). For simplicity the SS, MS, MS-PM, and LDWNN systems are built as single-state monophone systems without using any context information. All neural nets are three-layer MLPs with a hidden layer of 1000 units.

The TIMIT corpus has 5040 sentences, of which 3696 sentences are used for training, and the rest 1344 are used for testing. The target phoneme set contains 40 phonemes. The speech sounds are sampled at 16k Hz.

4. Results

4.1. Generalization Ability of Subband Ensemble Net

Results of the first experiment indicate that the subband ensemble net works well for a wide range of SNR across multiple frequency subbands. Fig. 3 depicts the phoneme error rate of subband ensemble net for the 2, 4, and 6\( ^{th} \) subband, centered around 0.5, 2, and 4 kHz, in clean and various levels of babble noise. For comparison the phoneme error rates of single neural nets trained in clean, babble noise at 20, 15, 10, 5, 0 dB SNR, or multi-style training (i.e., all the conditions mixed) are also plotted on the same figure. The neural net trained in clean works very well in clean, but the performance drops quickly as the noise level increases. The neural nets trained in various levels of babble noise show better generalization ability, but their maximum performance decreases noticeably even in mild noise at 20 dB SNR. Multi-style training helps improve the performance in the range of 10 to 20 dB SNR, but it still performs worse in clean conditions and in heavy noise (\( \leq 10 \) dB SNR) than the neural nets of matched conditions. The subband ensemble net significantly outperforms all single nets on matched conditions in every frequency subband.
Table 1: Phoneme Error Rate (PER) of variants SS and MS system in unknown noise

<table>
<thead>
<tr>
<th>Noise (dB SNR)</th>
<th>SS MFCC</th>
<th>PNCC</th>
<th>FDLP</th>
<th>MS clean</th>
<th>mixed</th>
<th>MS-PM clean</th>
<th>mixed</th>
<th>LDWNN</th>
</tr>
</thead>
<tbody>
<tr>
<td>clean</td>
<td>33.50</td>
<td>33.51</td>
<td>31.35</td>
<td>30.04</td>
<td>31.70</td>
<td>29.45</td>
<td>30.99</td>
<td>32.29</td>
</tr>
<tr>
<td>babble (15)</td>
<td>65.14</td>
<td>58.93</td>
<td>57.10</td>
<td>47.05</td>
<td>39.15</td>
<td>45.16</td>
<td>38.48</td>
<td>36.03</td>
</tr>
<tr>
<td>subway (15)</td>
<td>57.42</td>
<td>49.48</td>
<td>46.62</td>
<td>38.28</td>
<td>35.27</td>
<td>36.28</td>
<td>34.35</td>
<td>33.93</td>
</tr>
<tr>
<td>factory1 (10)</td>
<td>74.65</td>
<td>72.05</td>
<td>68.10</td>
<td>63.74</td>
<td>56.46</td>
<td>61.56</td>
<td>55.26</td>
<td>45.91</td>
</tr>
<tr>
<td>restaurant (10)</td>
<td>72.48</td>
<td>67.14</td>
<td>63.14</td>
<td>58.19</td>
<td>49.65</td>
<td>55.69</td>
<td>48.60</td>
<td>42.96</td>
</tr>
<tr>
<td>street (5)</td>
<td>80.98</td>
<td>70.26</td>
<td>67.26</td>
<td>62.56</td>
<td>54.72</td>
<td>59.37</td>
<td>53.92</td>
<td>44.41</td>
</tr>
<tr>
<td>exhali (5)</td>
<td>79.90</td>
<td>74.97</td>
<td>70.67</td>
<td>69.47</td>
<td>58.04</td>
<td>64.46</td>
<td>57.76</td>
<td>48.30</td>
</tr>
<tr>
<td>f16 (0)</td>
<td>77.32</td>
<td>84.31</td>
<td>86.10</td>
<td>86.54</td>
<td>89.31</td>
<td>83.83</td>
<td>85.07</td>
<td>71.27</td>
</tr>
<tr>
<td>car (0)</td>
<td>90.24</td>
<td>52.58</td>
<td>54.32</td>
<td>40.79</td>
<td>35.78</td>
<td>35.48</td>
<td>34.56</td>
<td>34.41</td>
</tr>
</tbody>
</table>

SS – single-stream baseline trained in clean; MS – multi-stream system with only one stream including all subbands; “clean” and “mixed” refer to the training conditions for subband neural nets; MS-PM – multi-stream system with a performance monitor; LDWNN – proposed long, deep, and wide neural net (a multi-stream system with subband ensemble net).

4.2. Generalization Ability of LDWNN

Results of the second experiment indicate that the LDWNN show superior performance and generalization ability in both clean and unknown noise. Fig. 5 compares the LDWNN system with the MS system with and without performance monitor. The MS systems have a single subband neural nets trained in clean or multi-style condition for every subband. The three curves start at the same level at about 30% in clean condition. As the noise level increases, the MS system with clean-trained subband neural nets climbs quickly to about 60% at 10 dB SNR. The MS system with multi-style trained subband neural nets show much slower decrease in system performance, but still it is far behind the LDWNN system with subband ensemble net, which has a phoneme error rate of about 40% at 10 dB SNR.

4.3. Robustness to Unknown Noise

Results of the third experiment (refer to Tab. 1) show that the LDWNN system (i.e., a multi-stream system with subband ensemble net) substantially out-performs the MS system (both MS and MS-PM) with subband neural nets trained clean or multi-style condition for all types of unknown noise at various dB SNR. It is noted that the MS system with multi-style training substantially out-performs the MS system trained in clean condition. Stream selection with performance monitor reduces the phoneme error rate by about 5% relative for MS system with clean-trained subband neural nets, but the gain diminishes for MS system with multi-style training. The average phoneme error rate (PER) of the proposed LDWNN system is about 20% and 14% relative lower than the MS baseline with subband neural net trained in clean and multi-style condition respectively. It is on average about 40%, 30%, and 28% relative lower than the single-stream baseline using MFCC, PNCC, and FDLP respectively.

Figure 6 depicts the PER of LDWNN and the single-stream (SS) baseline using MFCC and FDLP feature in various types of unknown noise at different SNRs. As the noise level increases, the PER of SS(MFCC) climbs up quickly from about 30% in clean condition to more than 90% in car noise at 0 dB SNR. In contrast, the LDWNN degrades much slower in noise. For car noise at 0 dB SNR, a low-frequency noise that corrupts the speech components below 0.5 kHz, the PER of LDWNN is very close to that of clean condition, suggesting the the proposed approach is very effective in dealing with narrow-band noise.

5. Conclusion

In this study, we propose a long, deep and wide system named LDWNN for robust speech recognition based on the multi-stream framework, in which the full band speech is divided into multiple frequency subbands, each act as an independent channel for speech recognition. It is assumed that any unknown noise can be approximated by white noise (or speech-shaped noise) of similar level across multiple frequency subbands. Within each subband, an ensemble net, which consists of multiple neural nets trained for different noise levels, followed by a neural net to select the most suitable net for any unknown noise. Experimental results show that the subband ensemble net optimizes information extraction for every frequency subband in both clean and noisy conditions. The LDWNN based on subband ensemble nets show superior performance and generalization ability in both clean and unknown noise.
6. References


