INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Feature Learning with Raw-Waveform CLDNNs for Voice Activity Detection

Ruben Zazo[*1], *Tara N. Sainath*[2], *Gabor Simko*[2], *Carolina Parada*[2]
[1] ATVS Biometric Recognition Group - Universidad Autonoma de Madrid
[2] Google Inc., Mountain View, CA
ruben.zazo@uam.es, tsainath@google.com, gsimko@google.com, carolinap@google.com

Abstract

Voice Activity Detection (VAD) is an important preprocessing step in any state-of-the-art speech recognition system. Choosing the right set of features and model architecture can be challenging and is an active area of research. In this paper we propose a novel approach to VAD to tackle both feature and model selection jointly. The proposed method is based on a CLDNN (Convolutional, Long Short-Term Memory, Deep Neural Networks) architecture fed directly with the raw waveform. We show that using the raw waveform allows the neural network to learn features directly for the task at hand, which is more powerful than using log-mel features, specially for noisy environments. In addition, using a CLDNN, which takes advantage of both frequency modeling with the CNN and temporal modeling with LSTM, is a much better model for VAD compared to the DNN. The proposed system achieves over 78% relative improvement in False Alarms (FA) at the operating point of 2% False Rejects (FR) on both clean and noisy conditions compared to a DNN of comparable size trained with log-mel features. In addition, we study the impact of the model size and the learned features to provide a better understanding of the proposed architecture.

1. Introduction

Voice Activity Detection (VAD) refers to the process of identifying segments of speech in an audio utterance [1]. This task is often a pre-processing stage of an automatic speech recognition (ASR) system to both reduce computation and to guide the user interface. A typical VAD system uses a frame-level classifier with acoustic features to make speech/non-speech decisions for each audio frame (every 10ms) [2]. Significant research has been devoted to finding the optimal features for this task [3, 4, 5], as well as the best classifier or model to use [6, 7, 8].

Deep Neural Networks (DNNs) are a commonly used model for VAD [6]. However, inspired by the advancements in acoustic modeling for speech, we explore alternative deep learning architectures for VAD. Convolutional Neural Networks (CNNs) [9] and Long Short-Term Memory (LSTM) recurrent neural networks [10] are popular choices since they have shown improvements over DNNs for several speech recognition tasks [11, 12]. The modeling capabilities of these different architectures are complementary. CNNs are good at reducing frequency variations, LSTMs are good at sequence modeling, and DNNs mapping features into a more separable space. To exploit the complementary traits of these systems, [13] introduced "Convolutional, Long Short-Term Memory, Fully Connected Deep Neural Networks" (CLDNNs), obtaining better performance than any of those architectures individually. Since VAD is a sequence task, we believe that architectures which model the temporal structure, such as a CLDNN, are better than a DNN for this task.

Even though the mel-scale was designed to mimic the critical bands of hearing in the human ear and has been successful in several speech related tasks, the problem of VAD and noise detection slightly different than speech processing. Thus, since neural networks are good at learning features, we explore if learning directly from the raw waveform has benefit for VAD.

In this work, we propose using raw waveform CLDNNs for VAD [14]. We compare the performance of this system against a standard DNN system, an LSTM system, and a CLDNN system all trained with log-mel features. We demonstrate that this approach achieves over a 78% relative improvement (in terms of FA when fixing FR at 2%) on both clean and noisy conditions when compared to a standard DNN trained using log-mel filterbank energies as input. Furthermore, we analyze these results and show the benefit of temporal modeling for VAD with the CLDNN, as well as the importance of learning features for the task at hand with the raw waveform.

The rest of the paper is as follows. In Section 2 we describe the CLDNN and raw waveform architectures proposed, as well as the LSTM and DNN reference architectures. The experimental setup is described in Section 3. Section 4 is devoted to presenting and analyzing the results and, finally, Section 5 concludes the paper.

2. Neural Network Architectures for Voice Activity Detection

In this section we describe the neural network architectures we compare for VAD, namely DNNs, LSTMs, CLDNNs and raw waveform CLDNNs. For each architecture we explore the impact of model size and select their configurations to obtain models of size $\sim$30k, $\sim$100k, and $\sim$200k parameters.

2.1. Baseline - Deep Neural Network

DNNs have been shown to give good performance for VAD [6]. The baseline DNN model used in this paper is a standard feedforward fully connected neural network with k hidden layers and n hidden units per layer. For each hidden layer, a rectified linear unit (ReLU) function is used. The output layer of our DNN model is a softmax with 2 units to predict speech and non-speech. The input into this model is a 40-dimensional log-mel feature, surrounded by a context of 5 past frames and 5 future frames. We experimented with larger input context windows, but it did not affect performance significantly. More details on this topology for different model sizes can be seen in Table 1.

[*] The author performed the work as an intern at Google, Mountain View

http://dx.doi.org/10.21437/Interspeech.2016-268

DNN model details			
# hidden layers	2	3	4
# hidden units per layer	64	128	208
Total number of parameters	32,384	89,344	221,728

Table 1: Details of the different parameters used for the DNN log-mel model.

2.2. Long Short-Term Memory Recurrent Neural Networks

Since LSTMs are good at sequential tasks, we explore if using a recurrent architecture will help with the sequential nature of a VAD task. The LSTM VAD architecture is unidirectional and similar to the architecture described in [15] with *k* hidden layers and *n* hidden units per layer. Note we do not use a projection layer with the LSTM. More details on this topology can be seen in Table 2. The input into the LSTM is a single 40-dimensional log-mel feature. The LSTM is unrolled for 20 time steps for training with truncated backpropagation through time (BPTT). In addition, the output state label is delayed by 5 frames, as we have observed that information about future frames improves prediction of the current frame.

LSTM model details			
# hidden layers	3	3	3
# hidden units per layer	32	64	96
Total number of parameters	29,506	106,114	229,826

Table 2: Details of the different parameters used for the LSTM log-mel model.

2.3. CLDNN

Convolutional, long short-term memory deep neural networks (CLDNNs) are a new type of sequential model that have shown improvements over LSTMs for LVCSR tasks [13]. The idea is that convolutional layers are good at modeling frequency variations, LSTMs are good at modeling temporal variations and DNN layers are good at mapping features to a more separable space. The CLDNN architecture uses all 3 layers in a unified framework, helping to combine the benefits of individual layers.

A diagram of the CLDNN architecture is shown in the `CLDNN` module in Figure 1. The input to the CLDNN is a 40-dimensional log-mel feature. The first layer of the CLDNN architecture consists of a frequency convolution layer, with filters of size 1 × 8 in `time ×frequency`. Consistent with [9] our pooling strategy is to use non-overlapping max pooling along the frequency axis, with a pooling size of 3. The output from the convolutional layer is passed to a few LSTM layers, and then to one DNN layer, before predicting 2 output targets. Details for each layer of the CLDNN for different model sizes can be seen in Table 3.

2.4. Raw Waveform CLDNN

Recently, [14] introduced a raw waveform CLDNN architecture, and showed that it was possible to learn directly from the raw waveform rather than using log-mel features. The network was found to learn a frequency representation very similar to log-mel, but learned for the task at hand. The authors found improvements with raw waveform for noisy tasks. Motivated by this work, and given the large body of research into the appropriate features for VAD, we wanted to see if there was benefit to learning the features directly in the network.

A block diagram of the raw waveform CLDNN is shown in Fig. 1. The input into the raw waveform CLDNN is a raw signal spanning roughly M samples, where $M = 35$ms . A convolution layer with P filters is convolved against the input, with each filter spanning a length of N. Typically we use between 40-128 filters for P and $N = 25$ms filters. After that, we pool over the entire length of the convolution ($M-N+1$). Finally, we apply a rectified non-linearity, followed by a stabilized logarithm compression, to produce a P-dimensional time-frequency representation. The output of this is passed to a CLDNN, as described in the previous section. The time convolution and CLDNN layers are trained jointly. Details for each layer of the raw waveform CLDNN for all model sizes can be seen in Table 4.

CLDNN model details			
Freq convolution			
# filter outputs	32	64	64
filter size (freq x time)	8x1	8x1	8x1
pooling size (freq x time)	3x1	3x1	3x1
LSTM layers			
# lstm hidden layers	1	2	3
# hidden units per layer	32	64	80
DNN layer			
# hidden units	32	64	80
Total number of parameters	37,570	131,642	218,498

Table 3: Details of the different parameters used for the CLDNN log-mel model.

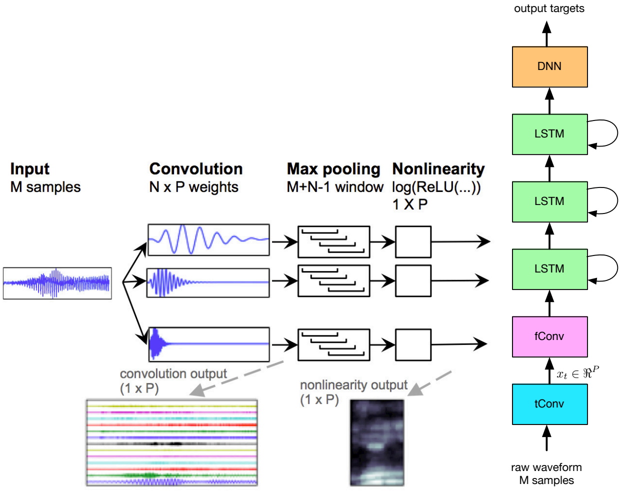

Figure 1: Modules of the raw waveform CLDNN: a) Time-domain Convolution Layer, b) Time convolution and CLDNN layers

3. Experimental Details

3.1. Dataset

Our experiments are conducted on a noisy training set consisting of 3,800 hours (3 million uterances), where roughly 50% of the frames correspond to speech and the remaining frames correspond to background noise. This data set was created by artificially adding noise to clean utterances using a room simulator.

Raw waveform CLDNN model details			
Time convolution			
# filter outputs	40	84	128
filter size: 1 x 25ms	1x401	1x401	1x401
pooling size: 1 x 10ms	1x161	1x161	1x161
Freq convolution			
# filter outputs	16	64	64
filter size (freq x time)	8x1	13x1	21x1
pooling size (freq x time)	3x1	6x1	9x1
LSTM layers			
# lstm hidden layers	2	2	3
# hidden units per layer	16	48	64
DNN layer			
# hidden units	16	48	64
Total number of parameters	35,794	124,738	221,938

Table 4: Details of the different parameters used for the raw waveform CLDNN model.

We added varying degrees of noise and reverberation, such that the overall SNR is between 5dB and 30dB. The noise sources are from YouTube, daily life noisy environmental recordings, car noise, cafeteria noise, and music.

We evaluate our system on a *clean* and *noisy* test-sets, consisting of anonymized voice-search queries. Our *clean* test set consists of 30 hours of audio, including about 50% speech frames. Our *noisy* test set consists of a 20 hours of audio, but only 15% (about 3 hours) of noisy speech, while the remaining 15 hours including noisy background such as music, car, cafeteria noise. These are meant to represent two different use-cases or applications.

All training and test-sets are anonymized and hand-transcribed, and are representative of Google's voice search traffic.

3.2. Neural Network Training and VAD Evaluation

The proposed LSTM, CLDNN and raw waveform CLDNN systems will be compared to a reference DNN VAD system. The input feature for all models but the raw waveform CLDNN are 40-dimensional log-mel filterbank features, computed every 10ms. All neural networks are trained using the asynchronous stochastic gradient descent (ASGD) optimization strategy described in [16] with the cross-entropy criterion. The CNN and DNN layers are initialized using the Glorot-Bengio strategy described in [17] while the LSTM layers are uniform randomly initialized to be between -0.02 and 0.02. The learning rates are exponentially decayed and independently chosen for each model, and are chosen to be the largest value such that training remains stable.

4. Results

In this section we present and compare the performance of the different models. All models are trained on the noisy training set and results are reported for both clean and noisy test sets. We have explored different sizes for each model, and the used parameters are summarized in Tables 1, 2, 3 and 4.

4.1. Comparison of the different models

First, we establish a fair comparison of the proposed methods with the reference DNN and LSTM-based approaches (i.e., LSTM, CLDNN, rawCLDNN) with a comparable number of total parameters. The systems presented here have been designed to have roughly 100k parameters and the exact details can be seen in the second column of the Tables 1, 2, 3 and 4. As shown in Figs. 2 and 3 the LSTM-based architectures outperform the DNN for both clean and noisy tasks. Furthermore, we find these result hold even when using a state machine [7] to smooth outputs of the VAD, which shows the importance of having sequential modeling for VAD on top of a state machine. Moreover, using raw waveform modeling to learn the features offers improvements over the log-mel-based systems in noisy environments, being consistent with [18] which found more benefits with raw waveform CLDNNs in noisy environments. Overall, the raw waveform CLDNN provides a ∼78% and ∼84% relative improvement in FA over the DNN at the operating point of 2% FR on clean and noisy conditions respectively. In the next section, we provide a deeper analysis to better understand these gains.

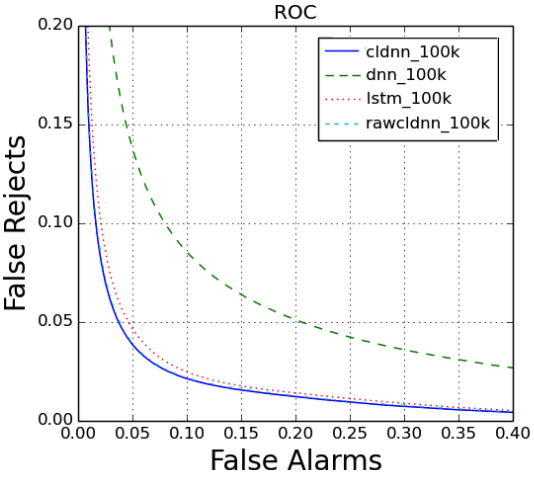

Figure 2: ROC curve for the different systems with ∼100k parameters on the clean test set. DNN performs worst, followed by LSTM while CLDNN and rawCLDNN fall one on top of the other

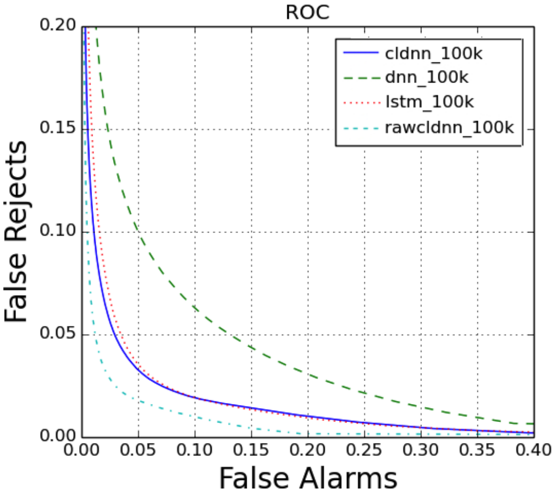

Figure 3: ROC curve for the different systems with ∼100k parameters on the noisy test set. DNN performs worst, followed by LSTM, then CLDNN and finally, performing the best, rawCLDNN

4.2. Analysis

4.2.1. *Benefit of Temporal Modeling*

In order to have a better insight of why LSTM-based architectures give improvements over the DNN, we recorded a single utterance with 3 speech segments in a noisy cafeteria environment. Fig 4 depicts the posterior probability of SPEECH, as provided by the different neural networks architectures. As can be seen in the figure, the output of the DNN system is very noisy compared to LSTM-based architectures, namely LSTM, CLDNN and rawCLDNN. Effectively, the LSTM layers in these architectures help smooth the output due to the recurrent connections in the LSTM which feed the activations from previous time steps as input to make the decision for the current input. Again, we should stress that while these figures show frame-level decisions, we do find that combining the VAD with a finite state machine still results in sequential models providing the best performance in terms of FA/FR.

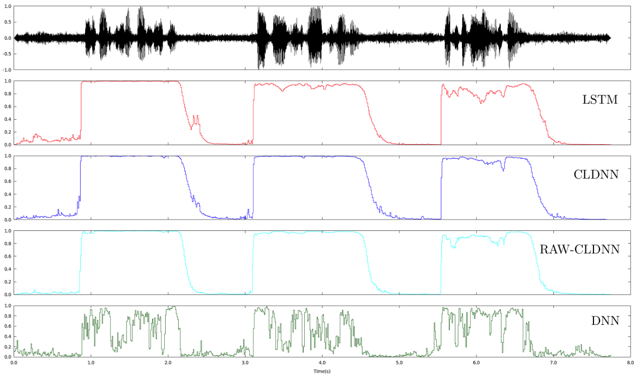

Figure 4: Frame by frame posterior of the different systems in a single utterance recorded in a noisy environment.

4.2.2. *Benefit of Learning Feature Representation*

To understand the improvements obtained with the raw waveform, Fig. 5 calculates the peak magnitude response for each filter in our biggest raw waveform CLDNN (details in the third column of Table 4), and sorts this based on increasing peak frequency. The figure highlights that filterbank learning devotes more filters to high frequencies compared to the standard mel filterbanks. This result is in contrast with our findings in [14] where using a similar architecture for acoustic modeling resulted in the learned filterbanks devoted more filters to lower frequencies. Since in this case the task is VAD and not acoustic modeling, it is possible that having more filters at high frequency regions helps to better discriminate between speech and noise. This further highlights the importance of learning features for the task at hand.

4.3. Impact of total number of parameters

In this section, we analyze the behavior of the 4 architectures with different number of parameters, namely ∼30k, ∼100k and ∼200k. Table 5 shows the false alarm rate when fixing the false rejects at 2% for all the four different models and the three sizes. The details of those models can be seen in Tables 1, 2, 3 and 4. The results show that the raw waveform-based system is consistently better in noisy conditions achieving 83% and 53% relative improvement (averaging the three model sizes) compared to DNN and CLDNN respectively. In addition the raw waveform-based system is less affected by the size of the model than the other systems. In clean conditions the performance of the raw waveform system is slightly worse than the CLDNN using log-mel as an input but the degradation is very small compared to the improvements in noisy conditions (3% average relative degradation on clean against 53% relative improvement on noisy).

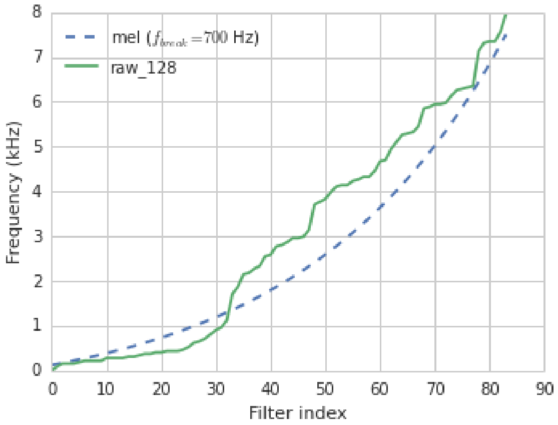

Figure 5: Center frequencies of learned filterbanks.

	FA% (clean/noisy) when fixing FR to 2%		
	∼30k	∼100k	∼200k
DNN	50.8 / 26.3	50.5 / 25.9	50.3 / 25.3
LSTM	14.6 / 20.4	13.1 / 9.3	12.6 / 10.2
CLDNN	**12.5** / 13.5	**10.9** / 9.5	**10.9** / 7.9
Raw CLDNN	13.1 / **4.7**	11.1 / **4.1**	11.3 / **4.2**

Table 5: False Alarm rates of the different sizes/models when fixing the False Reject rate to 2%

Finally, since the raw waveform CLDNN uses between 40 to 128 filters, whereas the log-mel based systems always use 40, we wanted to understand if the improvements in raw waveform were due to increased feature size. Note that the smallest raw waveform CLDNN system in Table 5 has also 40 time filters so that it can be compared more fairly to the other systems. The table shows that with the same number of filters, the raw waveform outperforms the other architectures, particularly in noisy conditions.

5. Conclusions

We presented a novel raw waveform-based model for the VAD task that shows significant gains over previous models. We showed that using a sophisticated acoustic model (CLDNN) fed with the raw waveform trained on a big dataset we can obtain ≥78% relative improvement on clean and noisy conditions compared to a DNN-based system fed with log mel of equal size. The results prove that large improvements can be achieved using temporal modeling and learning the feature representation from the data.

6. Acknowledgements

Thank you to Rohit Prabhavalkar for help with experiments. Also thank you to Ron Weiss and Kevin Wilson for useful discussions regarding raw waveform modeling and VAD.

7. References

[1] J. Sohn, N. S. Kim, and W. Sung, “A statistical model-based voice activity detection,” *Signal Processing Letters, IEEE*, vol. 6, no. 1, pp. 1–3, 1999.

[2] S. Van Gerven and F. Xie, “A comparative study of speech detection methods.” in *Eurospeech*, vol. 97, 1997.

[3] E. Chuangsuwanich and J. R. Glass, “Robust voice activity detector for real world applications using harmonicity and modulation frequency.” in *INTERSPEECH*. Citeseer, 2011, pp. 2645–2648.

[4] A. Misra, “Speech/nonspeech segmentation in web videos.” in *INTERSPEECH*, 2012.

[5] P. Kumar, A. Tsiartas, and S. Narayanan, “Robust voice activity detection using long-term signal variability,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 3, pp. 600–613, 2011.

[6] Z. XL and D. Wang, “Boosted deep neural networks and multiresolution cochleagram features for voice activity detection,” in *Proc. Interspeech*, 2014, pp. 1534–1538.

[7] T. Hugues and K. Mierle, “Recurrent Neural Networks for Voice Activity Detection,” in *Proc. ICASSP*, 2013.

[8] S. Thomas, G. Saon, M. V. Segbroeck, and S. Naranyanan, “Improvements to the ibm speech activity detection system for the darpa rats program,” in *Proc. ICASSP*, 2015.

[9] T. N. Sainath, A. Mohamed, B. Kingsbury, and B. Ramabhadran, “Deep Convolutional Neural Networks for LVCSR,” in *Proc. ICASSP*, 2013.

[10] H. Sak, A. Senior, and F. Beaufays, “Long Short-Term Memory Recurrent Neural Network Architectures for Large Scale Acoustic Modeling,” in *Proc. Interspeech*, 2014.

[11] G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, “Deep Neural Networks for Acoustic Modeling in Speech Recognition,” *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[12] R. Zazo, A. Lozano-Diez, J. Gonzalez-Dominguez, D. T. Toledano, and J. Gonzalez-Rodriguez, “Language identification in short utterances using long short-term memory (lstm) recurrent neural networks,” *PloS one*, vol. 11, no. 1, 2016.

[13] T. N. Sainath, O. Vinyals, A. Senior, and H. Sak, “Convolutional, Long Short-Term Memory, Fully Connected Deep Neural Networks,” in *Proc. ICASSP*, 2015.

[14] T. N. Sainath, R. J. Weiss, K. W. Wilson, A. Senior, and O. Vinyals, “Learning the Speech Front-end with Raw Waveform CLDNNs,” in *Proc. Interspeech*, 2015.

[15] A. Graves, A. Mohamed, and G. Hinton, “Speech recognition with deep recurrent neural networks,” in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 6645–6649.

[16] J. Dean, G. Corrado, R. Monga, K. Chen, M. Devin, Q. Le, M. Mao, M. Ranzato, A. Senior, P. Tucker, K. Yang, and A. Ng, “Large Scale Distributed Deep Networks,” in *Proc. NIPS*, 2012.

[17] X. Glorot and Y. Bengio, “Understanding the Difficulty of Training Deep Feedforward Neural Networks,” in *Proc. AISTATS*, 2014.

[18] T. N. Sainath, R. J. Weiss, K. W. Wilson, A. Narayanan, M. Bacchiani, and A. Senior, “Speaker Localization and Microphone Spacing Invariant Acoustic Modeling from Raw Multichannel Waveforms,” in *to appear in Proc. ASRU*, 2015.