Neural networks for distant speech recognition

Citation for published version:
Renals, S & Swietojanski, P 2014, Neural networks for distant speech recognition. in *Proceedings 2014 Workshop on Hands-Free Speech Communication and Microphone Arrays* . Institute of Electrical and Electronics Engineers (IEEE), pp. 172-176. https://doi.org/10.1109/HSCMA.2014.6843274

Digital Object Identifier (DOI):
10.1109/HSCMA.2014.6843274

Published In:
Proceedings 2014 Workshop on Hands-Free Speech Communication and Microphone Arrays

NEURAL NETWORKS FOR DISTANT SPEECH RECOGNITION

Steve Renals and Pawel Swietojanski

Centre for Speech Technology Research, University of Edinburgh, Edinburgh EH8 9AB, UK
{s.renals,p.swietojanski}@ed.ac.uk

ABSTRACT

Distant conversational speech recognition is challenging owing to the presence of multiple, overlapping talkers, additional non-speech acoustic sources, and the effects of reverberation. In this paper we review work on distant speech recognition, with an emphasis on approaches which combine multichannel signal processing with acoustic modelling, and investigate the use of hybrid neural network / hidden Markov model acoustic models for distant speech recognition of meetings recorded using microphone arrays. In particular we investigate the use of convolutional and fully-connected neural networks with different activation functions (sigmoid, rectified linear, and maxout). We performed experiments on the AMI and ICSI meeting corpora, with results indicating that neural network models are capable of significant improvements in accuracy compared with discriminatively trained Gaussian mixture models.

Index Terms— convolutional neural networks, distant speech recognition, rectifier unit, maxout networks, beamforming, meetings, AMI corpus, ICSI corpus

1. INTRODUCTION

Distant conversational speech recognition [1] is highly challenging for several reasons. A typical recording may include multiple overlapping talkers, as well as additional non-speech acoustic sources, and the recording environment may have significant reverberation. During the 1990s a number of pioneering studies investigated the development of DSR systems based on a microphone array (e.g. [2, 3, 4]), and an evaluation framework for speech recognition based on multichannel recordings of Wall Street Journal sentences [5] has enabled some comparability in this area. In practice, the effect of speaker and channel adaptation has been found to have a much greater effect on speech recognition word error rates, compared with changes to the beamforming algorithm and postfilter [6]. On the other hand, a number of techniques have been developed to address specific challenges such as reverberation and overlapping talkers [7, 8].

Over the past decade, there has been an increased focus on the recognition of multiparty conversational speech. Much of the work has been in meeting transcription: the ICSI Meeting Project resulted in the first major corpus in the area. The ICSI Meeting Corpus [9] used individual headmounted microphones (IHM), as well as 4 boundary microphones placed about 1m apart along the tabletop. One limitation of this corpus was the fact that the distant microphones were widely spaced and not in known postions. Subsequently, the AMI meeting corpus [10] was recorded using one or two 8-element circular microphone arrays, in addition to headset and lapel microphones. From 2004–2009, the NIST RT evaluations focused on the problem of meeting transcription, and enabled comparison between various automatic meeting transcription systems (e.g. [11, 12]), in the IHM, SDM (single distant microphone), and MDM (multiple distant microphone) cases. In the MDM systems, the microphone array processing part was usually distinct from the speech recognition part. For instance, the AMIDA MDM system of Hain et al [12] processed the multi-channel microphone array data using a Wiener noise filter, followed by beamforming based on time-delay-of-arrival (TDOA) estimates, postprocessed using a Viterbi smoother. In practice the beamformer tracked the direction of maximum energy, passing the beamformed signal onto a conventional ASR system – in the case of [12], a Gaussian mixture model / hidden Markov model (GMM/HMM) trained using the discriminative minimum phone error (MPE) criterion [13], speaker adaptive training [14], and the use of bottleneck features [15] derived from a neural network trained as a phone classifier. The resulting system employed a complex multi-pass decoding scheme, including substantial cross-adaptation and model combination.

Seltzer [16] has argued that the above approach, which may be viewed as using a speech enhancement framework for microphone array processing, is sub-optimal and that it would be preferable to optimise all system components using a common objective function related to the overall task, i.e. minimising the speech recognition word error rate (WER). LIMABEAM [17, 18] is an example of such an approach, in which the parameters of the microphone array beamformer are estimated so as to maximise the likelihood of the correct utterance model. Marino and Hain [19] explored removing the beamforming component entirely,

This research was supported by EPSRC Programme Grant grant, no. EP/I031022/1 (Natural Speech Technology).

and driving an HMM/GMM system with concatenated feature vectors from the different microphones. Whereas the LIMABEAM approach retains explicit beamforming parameters, but optimises them according to a criterion related to speech recognition accuracy, the concatenated approach makes the beamforming parameters implicit.

Building on [18, 19], our goal in this paper is to explore ways in which deep neural networks can learn suitable representations for distant speech recognition based on multichannel input. Deep neural network (DNN) acoustic models [20] now define the state-of-the-art in acoustic modelling for automatic speech recognition (ASR), typically using a hybrid configuration [21, 22, 23, 24, 25, 26] in which the neural network is used to estimate HMM output probabilities. We have recently demonstrated that hybrid neural network systems can significantly increase the accuracy of distant conversational speech recognition [27], by conducting experiments using the AMI corpus. A benefit of using neural network acoustic models is the possibility to use frequency domain feature vectors with no extra cost (unlike GMM-based systems which require a full covariance model); experiments indicate that log spectral domain features result in a small, but consistent, reduction in WER over cepstral domain features [28].

This paper extends our previous work to the ICSI corpus, and investigates the use of piecewise-linear activation functions which have shown promise for clean speech recognition [29, 30, 31, 32]. By producing highly sparse hidden activations, we believe that some of these activation functions are well suited to distant speech recognition. In each case we also experiment with convolutional layers [33] and their recent variant for modelling speech by convolution and pooling along frequency [34].

2. CONVOLUTIONAL NEURAL NETWORKS

A fully-connected feed-forward neural network implements a cascade of $L-1$ non-linear transformations in which the l-th layer computes $\mathbf{h}^l = f\left(\mathbf{W}^l\mathbf{h}^{l-1} + \mathbf{b}^l\right)$. $\mathbf{W}^l \in \mathbb{R}^{A\times B}$ and $\mathbf{b}^l \in \mathbb{R}^B$ are a trainable matrix of connection weights between two consecutive layers and vector of additive biases, respectively. The activation function $f(\cdot)$ applies some non-linearity to the hidden units. The topmost L-th layer estimates posterior probability of a tied context-dependent phonetic HMM state s given an observation vector $\mathbf{o}_t$ at time t: $P(s|\mathbf{o}_t) = \exp(a\{s\})/\sum_{s'}\exp(a\{s'\})$, where $a\{s\} = \mathbf{w}_s^L\mathbf{h}^{L-1} + b_s^L$ is a linear activation at the s-th output of the top layer.

This architecture may be enriched by constraining one or more of the lower layers to have local connectivity and to share parameters – such a model is referred to as a Convolutional Neural Network (CNN). CNNs have defined the state of the art on many vision tasks [35] and recently have been found to reduce the speech recognition word error rate (WER) when applied to acoustic modelling [34, 36]. The major conceptual difference between recent CNN structures for speech modelling and previous trials in the form of both CNNs [35] and the closely-related time-delay neural networks [37] lies in performing convolution and/or sharing parameters across frequency rather than time.

The input to a CNN comprises of (log) mel-spectral features within an acoustic context window $\mathbf{V} \in \mathbb{R}^{B\times Z}$ re-ordered in a way such that each of B frequency bands contain all the Z related coefficients (statics and dynamics). The hidden activations are then generated by a linear valid convolution of a local frequency region, i.e. $[\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3]$ with J weight vectors (filters), $\mathbf{w}_{1\ldots J}$. The same set of filters is then applied across different frequency regions to form a complete set of convolutional activations which can be subsampled, for instance by using the maxpooling operator [33], to further limit the variability across different frequencies.

The most frequent choice for the hidden layer activation function $f(\cdot)$ until recently was sigmoid $f(x) = 1/(1+\exp(-x))$, or the closely related $\tanh(x)$. The reason for this is that smooth and continuously differentiable non-linearities were considered to be a crucial component of training DNNs, allowing for a smooth flow of back-propagated gradients and the discovery of highly non-linear features. However, it has been shown experimentally that semi-hard functions which break many of these conventional design mainstays can be not only very accurate but also easy and fast to learn. An example of such activation functions are rectified linear units (ReLU) [38] implementing the lower bounded operation $f(x) = \max(0, x)$ and maxout units [39] computing $f(x_i \ldots x_{i+K}) = \max_{j=i}^{i+K} x_j$ over a group of K units. Unbounded piece-wise linear activation functions prevent the network from saturating and mitigate the vanishing gradients problem in deeper networks.

Stochastic gradient descent training is carried out by minimising a negative log posterior probability cost function $\mathcal{L}(\theta) = -\sum_{t=1}^{T}\log P(s_t|\mathbf{o}_t;\theta)$, over the set of training examples $\mathcal{O} = \{\mathbf{o}_1, \ldots, \mathbf{o}_T\}$; where s_t is the most likely state at time t obtained by a forced-alignment of the acoustics with the transcript, and $\theta = \{\mathbf{W}_1, \ldots, \mathbf{W}_L, \mathbf{b}_1, \ldots, \mathbf{b}_L\}$ is the set of parameters of the network. Decoding is carried out using scaled log-likelihoods $\log p(\mathbf{o}_t|s) \propto \log P(s|\mathbf{o}_t) - \log P(s)$, where $P(s)$ is a prior probability of state s calculated from training data.

3. EXPERIMENTAL SETUP

We have performed experiments using the AMI[1] [10] and ICSI[2] [9] meeting corpora. We used training and test split defined in the 100h AMI corpus release, as in our previous work [27, 40]. In the 72h ICSI corpus we used 5 complete meetings for testing and defined `dev` and `eval` sets[3]. For simplicity

[1] `http://corpus.amiproject.org`

[2] `http://catalog.ldc.upenn.edu/LDC2004S02`

[3] `dev` {Bmr021 and Bns001}, `eval` {Bmr013, Bmr018 and Bro021}

of exposition, we report results using all segments, including those with simultaneous speakers. The WERs from scoring non-overlapped segments only are around 10-12% absolute lower for both AMI and ICSI corpora[4], and results using this scoring can be found in [40].

We used a 50,000 word pronunciation dictionary [12]. For the AMI experiments we used the language model (LM) described in [27] which was built using both in-domain AMI training transcripts (0.8M words) as well as Fisher (22M words) and Switchboard (3M words) text data. For the ICSI experiments we further interpolate the AMI LM with in-domain 3-gram LM estimated from ICSI training transcripts. The AMI LM gives a perplexity of 78 on the AMI `dev` set; the ICSI LM gives perplexity of 110 on ICSI `dev` set.

All neural networks were trained using 40-dimensional log Mel filterbank (FBANK) features appended with the first and the second time derivatives [28]. Our distant microphone systems within this work remain unadapted to both speakers and sessions. Based on previous experiments, DNNs with sigmoid or ReLU hidden unit activation functions had 6 hidden layers with 2048 hidden units per layer. Maxout networks were tuned to have a similar number of parameters with six hidden layers, resulting in 1150 maxout units and a group size $K = 3$. Convolutional layers were configured to have $J = 128$ filters. Experiments were performed using the Kaldi speech recognition toolkit [42], and the `pylearn2` machine learning library [43].

For each neural network we sample initial weights from a uniform distribution with range $\pm r$. For the ReLU and maxout models we use $r = 0.005$, while the sigmoid networks make use of a normalised initialisation with $r = 4\sqrt{6/(n_l + n_{l+1})}$, where n_l denotes the input dimensionality of the l-th layer [44]. All models are finetuned with the exponentially decaying "newbob" learning rate schedule[5] staring from an initial learning rate of 0.08 (for sigmoids) and 0.01 for piece-wise linear activations. We have not used unsupervised pre-training [45] in these experiments. Although pretraining can be beneficial we have observed its effect to lessen as the amount of training data increases. Restricted Boltzmann machine [45] pretraining is well-matched to sigmoid activation functions, and can also be used for convolutional layers [46]. For activation functions such as ReLUs and maxout it would be possible to used stacked autoencoder pretraining [47] which is not limited to a specific form of activation function.

Our aim in developing these experimental setups is to enable our experiments to be reproducible by other researchers by using readlily available data for acosutic and language model training.

[4]We use `asclite` tool for scoring overlapped speech [41] following the NIST RT recommendations (`http://nist.gov/speech/tests/rt/2009`). Scoring for non-overlapped segments only is obtained by using `asclite` with the `-overlap-limit 1` option.

[5]Developed as part of ICSI QuickNet: `http://www.icsi.berkeley.edu/Speech/qn.html`

Table 1. WER (%) on AMI and ICSI – SDM.

System	AMI	ICSI
BMMI GMM-HMM (LDA+STC)	63.2	56.1
DNN – Sigmoid	53.1	47.8
DNN – ReLU	51.1	46.3
DNN - Maxout	50.8	45.9
CNN – Sigmoid	51.3	46.5
CNN – ReLU	50.3	45.6
CNN – Maxout	50.5	-

4. RESULTS

In this section we report on speech recognition experiments using the AMI and ICSI corpora, with two distant speech conditions (SDM and MDM) and one close-talking speech condition (IHM). We have three baseline acoustic models:

- a GMM-based system, discriminatively trained using boosted maximum mutual information (BMMI) [48], with mel-frequency cepstral coefficient (MFCC) features post-processed with linear discriminant analysis (LDA) and decorrelated using a semi-tied covariance (STC) transform [49];
- a DNN using 6 hidden layers, with sigmoid activation functions, using 40-dimension log mel spectral features (plus 1st and 2nd derivatives)[27];
- a deep CNN comprising one convolutional layer with 128 filters, followed by 5 fully-connected layers, using the same acoustic features as the DNN [40].

Results for AMI are on the `dev` set (for comparability with [27, 40]), results for ICSI are on the `eval` set.

4.1. SDM – Single Distant Microphone

The SDM experiments used the first microphone from the AMI circular array and the second tabletop boundary microphone from the ICSI recordings. Our results are shown in Table 1, with the three baseline systems in line 1 (BMMI GMM), line 2 (DNN – Sigmoid), and line 5 (CNN-Sigmoid). The DNN baseline has a 15% relative lower WER than the discriminative GMM baseline, with the CNN baseline improving over the DNN baseline by a further 3% relative. Comparing the ReLU and Maxout DNN and CNN systems, with the sigmoid baselines, shows a consistent improvement in WER of 1.5–4.5%. Comparing DNNs and CNNs with the same activation function, we see that networks with the sigmoid nonlinearity benefit the most from a convolutional layer (3–4% relative reduction in WER), although the ReLU and Maxout systems do benefit from the use of a convolutional layer (0.5–2% relative). We note that these experiments have been performed with a fixed number of filters, optimised for sigmoid-based systems; further experiments are needed to ascertain if the ReLU and Maxout systems would give large decreases in WER if there were more convolutional filters.

Table 2. WER (%) on AMI and ICSI – MDM with beamforming

System	AMI	ICSI
BMMI GMM-HMM (LDA+STC)	54.8	46.8
DNN – Sigmoid	49.5	41.0
DNN – ReLU	46.3	38.7
DNN – Maxout	46.4	39.0
CNN – Sigmoid	46.3	39.5
CNN – ReLU	46.0	37.6
CNN – Maxout	45.9	38.1

Table 3. WER (%) on AMI – MDM with multi-channel input

System	AMI	ICSI
CNN – Sigmoid (conventional)	50.4	43.3
CNN – Sigmoid (channel-wise)	49.5	40.1
CNN – ReLU (channel-wise)	48.7	37.5
CNN – Maxout (channel-wise)	48.4	37.8

4.2. MDM – Multiple Distant Microphones

For the MDM systems we consider: (1) *beamforming* the signal into a single channel (using all 8 microphones for AMI and 4 tabletop boundary microphones for ICSI) and following the standard acoustic modelling approaches used for the SDM case [27]; (2) *cross-channel pooling* using a channel-wise convolutional layer for training on multiple microphone channels, in which the hidden activations are constructed from the maximum activations across the channels. The ICSI data is characterised by large distances between microphones, and picking the right microphone for a talker is crucial, which may be well-matched to cross-channel pooling.

Table 2 shows the results for the models trained on a single beamformed channel (using BeamformIt [50]). We observe similar reductions in WER for sigmoid CNNs over DNNs as in the SDM case. The gain of CNN variants using ReLUs and Maxout in place of sigmoid activation functions remains small. These trends can be observed for both the AMI and ICSI datasets. We note that the WERs obtained using the DNN or CNN models (table 1) are lower than the WERs obtained for the discriminative GMM systems in the MDM case trained on a beamformed signal.

Table 3 shows the results obtained for CNNs trained using multi-channel input without beamforming. The first row presents a "conventional" approach where convolutional activations are produced by a sum of filter activations from each channel. Since that was found to be especially harmful for less constrained microphone configurations (ICSI) the following rows present a channel-wise approach where only the maxium activations within the channels are considered [40]. For the AMI data the CNN architectures return similar WERs to DNNs using beamformed input; for the ICSI data CNNs using cross-channel pooling match the WERs obtained using beamforming, probably due to less accurate TDOA estimates from the uncalibrated microphone array.

Table 4. Word Error Rates (%) on AMI – IHM

System	AMI
BMMI GMM-HMM (LDA+STC, SAT)	29.6
DNN – Sigmoid	26.6
DNN – ReLU	25.5
DNN – Maxout	26.3
CNN – Sigmoid	25.6
CNN – ReLU	24.9
CNN – Maxout	25.0

4.3. IHM – Individual Headset Microphone

For comparison purposes we present WERs for the different architectures using close-talking IHM inputs, for the AMI data (Table 4). The WER trend is similar to the distant microphone cases, suggesting that the results for the different nonlinear activations generalise across signal qualities. BMMI-GMM models were estimated using speaker adaptive training.

5. DISCUSSION & CONCLUSIONS

The presented distant conversational speech recognition experiments have explored a number of different neural network architectures, using different nonlinear functions for the hidden layer activations. Our results, using the AMI and ICSI corpora, show that neural network acoustic models offer large reductions in WER compared with discriminatively trained GMM-based systems. Furthermore, we observed further significant reductions in WER by using a convolutional layer within a DNN architecture. Small, but consistent, reductions in WER were also obtained by using ReLU and Maxout activation functions in place of sigmoids.

These neural network based systems used log spectral input representations, which are potentially amenable to additional feature space transformations and modelling. In particular, our current experiments do not explicitly attempt to optimise the acoustic model for overlapping talkers, or for reverberation. The promising results using raw multiple channel input features in place of beamforming opens the possibilities to learning representations taking into account aspects such as overlapping speech.

6. REFERENCES

[1] M Wölfel and J McDonough, *Distant Speech Recognition*, Wiley, 2009.

[2] D Van Compernolle, W Ma, F Xie, and M Van Diest, "Speech recognition in noisy environments with the aid of microphone arrays," *Speech Commun.*, vol. 9, pp. 433–442, 1990.

[3] JE Adcock, Y Gotoh, DJ Mashao, and HF Silverman, "Microphone-array speech recognition via incremental MAP training," in *Proc IEEE ICASSP*, 1996, pp. 897–900.

[4] M Omologo, M Matassoni, P Svaizer, and D Giuliani, "Microphone array based speech recognition with different talker-array positions," in *Proc IEEE ICASSP*, 1997, pp. 227–230.

[5] M Lincoln, I McCowan, J Vepa, and HK Maganti, "The multi-channel Wall Street Journal audio visual corpus (MC-WSJ-AV): Specification and initial experiments," in *Proc IEEE ASRU*, 2005.

[6] E Zwyssig, F Faubel, S Renals, and M Lincoln, "Recognition of overlapping speech using digital MEMS microphone arrays," in *Proc IEEE ICASSP*, 2013.

[7] T Yoshioka, A Sehr, M Delcroix, K Kinoshita, R Maas, T Nakatani, and W Kellermann, "Making machines understand us in reverberant rooms: Robustness against reverberation for automatic speech recognition.," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 114–126, 2012.

[8] K Kumatani, J McDonough, and B Raj, "Microphone array processing for distant speech recognition: From close-talking microphones to far-field sensors," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 127–140, 2012.

[9] A Janin, D Baron, J Edwards, D Ellis, D Gelbart, N Morgan, B Peskin, T Pfau, E Shriberg, A Stolcke, and C Wooters, "The ICSI meeting corpus," in *Proc IEEE ICASSP*, 2003, pp. I–364–I–367.

[10] J Carletta, "Unleashing the killer corpus: experiences in creating the multi-everything AMI Meeting Corpus," *Language Resources & Evaluation*, vol. 41, pp. 181–190, 2007.

[11] A Stolcke, X Anguera, K Boakye, O Cetin, A Janin, M Magimai-Doss, C Wooters, and J Zheng, "The SRI-ICSI Spring 2007 meeting and lecture recognition system," in *Multimodal Technologies for Perception of Humans*, R Stiefelhagen, R Bowers, and J Fiscus, Eds., number 4625 in LNCS, pp. 373–389. Springer, 2008.

[12] T Hain, L Burget, J Dines, PN Garner, F Grezl, AE Hannani, M Huijbregts, M Karafiat, M Lincoln, and V Wan, "Transcribing meetings with the AMIDA systems," *IEEE Trans. Audio, Speech, & Language Process.*, vol. 20, pp. 486–498, 2012.

[13] D Povey and PC Woodland, "Minimum phone error and I-smoothing for improved discriminative training," in *Proc IEEE ICASSP*, 2002, pp. 105–108.

[14] T Anastasakos, J McDonough, R Schwartz, and J Makhoul, "A compact model for speaker-adaptive training," in *Proc ICSLP*, 1996, pp. 1137–1140.

[15] F Grézl, M Karafiát, S Kontár, and J Černocký, "Probabilistic and bottle-neck features for LVCSR of meetings," in *Proc IEEE ICASSP*, 2007, vol. 4, pp. IV–757–IV–760.

[16] ML Seltzer, "Bridging the gap: Towards a unified framework for hands-free speech recognition using microphone arrays," in *Proc HSCMA*, 2008.

[17] M Seltzer, B Raj, and R Stern, "Likelihood-maximizing beamforming for robust hands-free speech recognition," *IEEE Trans. Speech, & Audio Process.*, vol. 12, pp. 489–498, 2004.

[18] M Seltzer and R Stern, "Subband likelihood-maximizing beamforming for speech recognition in reverberant environments," *IEEE Trans. Audio, Speech, & Lang. Process.*, vol. 14, pp. 2109–2121, 2006.

[19] D Marino and T Hain, "An analysis of automatic speech recognition with multiple microphones," in *Proc Interspeech*, 2011, pp. 1281–1284.

[20] G Hinton, L Deng, D Yu, GE Dahl, A-R Mohamed, N Jaitly, A Senior, V Vanhoucke, P Nguyen, TN Sainath, and B Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, 2012.

[21] H Bourlard and N Morgan, *Connectionist Speech Recognition: A Hybrid Approach*, Kluwer, 1994.

[22] S Renals, N Morgan, H Bourlard, M Cohen, and H Franco, "Connectionist probability estimators in HMM speech recognition," *IEEE Trans Speech & Audio Process.*, vol. 2, pp. 161–174, 1994.

[23] N Morgan and H Bourlard, "Neural networks for statistical recognition of continuous speech," *Proc IEEE*, vol. 83, pp. 742–772, 1995.

[24] AJ Robinson, GD Cook, DPW Ellis, E Fosler-Lussier, SJ Renals, and DAG Williams, "Connectionist speech recognition of broadcast news," *Speech Commun.*, vol. 37, pp. 27–45, 2002.

[25] TN Sainath, B Kingsbury, B Ramabhadran, P Fousek, P Novak, and A Mohamed, "Making deep belief networks effective for large vocabulary continuous speech recognition," in *Proc IEEE ASRU*, 2011.

[26] GE Dahl, D Yu, L Deng, and A Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Trans Audio, Speech & Lang. Process.*, vol. 20, pp. 30–42, 2012.

[27] P Swietojanski, A Ghoshal, and S Renals, "Hybrid acoustic models for distant and multichannel large vocabulary speech recognition," in *Proc IEEE ASRU*, 2013.

[28] J Li, D Yu, J-T Huang, and Y Gong, "Improving wideband speech recognition using mixed-bandwidth training data in CD-DNN-HMM," in *Proc IEEE SLT*, 2012, pp. 131–136.

[29] MD Zeiler, M Ranzato, R Monga, M Mao, K Yang, QV Le, P Nguyen, A Senior, V Vanhoucke, J Dean, and GE Hinton, "On rectified linear units for speech processing," in *Proc IEEE ICASSP*, 2013.

[30] M Cai, Y Shi, and J Liu, "Deep maxout neural networks for speech recognition," in *Proc ASRU*, 2013.

[31] Y Miao, F Metze, and S Rawat, "Deep maxout networks for low-resource speech recognition," in *Proc. IEEE ASRU*, 2013.

[32] P Swietojanski, J Li, and J-T Huang, "Investigation of maxout networks for speech recognition," in *Proc IEEE ICASSP*, 2014.

[33] Y LeCun and Y Bengio, "Convolutional networks for images, speech and time series," in *The Handbook of Brain Theory and Neural Networks*, pp. 255–258. MIT Press, 1995.

[34] O Abdel-Hamid, A-R Mohamed, J Hui, and G Penn, "Applying convolutional neural networks concepts to hybrid NN–HMM model for speech recognition," in *Proc IEEE ICASSP*, 2012, pp. 4277–4280.

[35] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc IEEE*, vol. 86, pp. 2278–2324, 1998.

[36] TN Sainath, B Kingsbury, A Mohamed, GE Dahl, G Saon, H Soltau, T Beran, AY Aravkin, and B Ramabhadran, "Improvements to deep convolutional neural networks for LVCSR," in *Proc IEEE ASRU*, 2013.

[37] KJ Lang, AH Waibel, and GE Hinton, "A time-delay neural network architecture for isolated word recognition," *Neural Networks*, vol. 3, pp. 23–43, 1990.

[38] V Nair and G Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc ICML*, 2010, pp. 131–136.

[39] IJ Goodfellow, D Warde-Farley, M Mirza, A Courville, and Y Bengio, "Maxout networks," in *Proc ICML*, 2013.

[40] P Swietojanski, A Ghoshal, and S Renals, "Convolutional neural networks for distant speech recognition," *IEEE Signal Process. Letters*, 2014, To appear.

[41] JG Fiscus, J Ajot, N Radde, and C Laprun, "Multiple dimension Levenshtein edit distance calculations for evaluating ASR systems during simultaneous speech," in *Proc LREC*, 2006.

[42] D Povey, A Ghoshal, G Boulianne, L Burget, O Glembek, N Goel, M Hannemann, P Motlíček, Y Qian, P Schwarz, J Silovský, G Stemmer, and K Veselý, "The Kaldi speech recognition toolkit," in *Proc IEEE ASRU*, 2011.

[43] IJ Goodfellow, D Warde-Farley, P Lamblin, V Dumoulin, M Mirza, R Pascanu, J Bergstra, F Bastien, and Y Bengio, "Pylearn2: a machine learning research library," *arXiv:1308.4214*, 2013.

[44] X Glorot and Y Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc AISTATS*, 2010.

[45] G Hinton, S Osindero, and Y Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, pp. 1527–1554, 2006.

[46] H Lee, P Pham, Y Largman, and A Ng, "Unsupervised feature learning for audio classification using convolutional deep belief networks," in *Proc NIPS 22*, 2009, pp. 1096–1104.

[47] Y Bengio, P Lamblin, D Popovici, and H Larochelle, "Greedy layer-wise training of deep networks," in *Proc NIPS 19*, 2007, pp. 153–160.

[48] D Povey, D Kanevsky, B Kingsbury, B Ramabhadran, G Saon, and K Visweswariah, "Boosted MMI for model and feature-space discriminative training," in *Proc IEEE ICASSP*, 2008, pp. 4057–4060.

[49] MJF Gales, "Semi-tied covariance matrices for hidden Markov models," *IEEE Trans Speech and Audio Process.*, vol. 7, pp. 272–281, 1999.

[50] X Anguera, C Wooters, and J Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE Trans. Audio, Speech, & Lang. Process.*, vol. 15, pp. 2011–2021, 2007.