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

Complex Linear Projection (CLP): A Discriminative Approach to Joint Feature Extraction and Acoustic Modeling

Ehsan Variani, Tara N. Sainath, Izhak Shafran, Michiel Bacchiani

Google Inc., USA
{variani, tsainath, izhak, michiel}@google.com

Abstract

State-of-the-art automatic speech recognition (ASR) systems typically rely on pre-processed features. This paper studies the time-frequency duality in ASR feature extraction methods and proposes extending the standard acoustic model with a complex-valued linear projection layer to learn and optimize features that minimize standard cost functions such as cross-entropy. The proposed Complex Linear Projection (CLP) features achieve superior performance compared to pre-processed Log Mel features.

Index Terms: feature extraction, complex neural network, speech recognition

1. Introduction

The most common ASR feature extraction method applies an auditory-inspired Mel filter bank to the squared magnitude of the short-time Fourier transform of the speech window [1]. The filter parameters are set based on knowledge about human speech perception. The filter bank outputs, commonly referred to as Mel features, are then used to train an acoustic model (AM). This *separation* of perceptually motivated filters from the AM, is not always the best choice in statistical modeling frameworks such as ASR, where the end goal is word error rate minimization. This motivates the essence of data driven learning schemes for *joint* learning of filter bank parameters and the acoustic model.

There have been numerous efforts in the ASR community looking at data-driven filter bank design. Statistical tools such as Independent Component Analysis [2, 3, 4], Linear Discriminant Analysis (LDA) [5] were explored to design filter bank which extract higher order statistical characteristics of the speech signal. Joint optimization of the filter bank parameters and classification error were investigated in [6] which led to superior results over the baseline Log Mel features. While these statistical methods have shown some improvement in small ASR tasks, most of these methods were explored within *shallow* acoustic model architectures such as AMs based on Gaussian Mixture Models (GMMs).

The emergence of *deep* neural networks for acoustic modeling [7] has recently sparked a resurgence in data-driven techniques which jointly estimate filter bank and AM parameters [8, 9, 10, 11, 12, 13, 14, 15]. In these models, the first layers of AM network is designed to learn filter bank directly from raw waveform. The architecture typically consist of a convolution layer followed by a pooling layer. The success of the resulting model highly depends on the choice of convolutional and pooling layer parameters. While most of these techniques still lag behind Log Mel trained AMs, recently proposed Raw waveform model has shown promising results for both single and multi-channel systems [14, 16, 17].

This paper introduces a technique which performs both filtering and pooling in the frequency domain, avoiding the complexity of convolution in the time domain and the extra parameter tuning required for convolutional layers. We propose to process frequency-domain features with a complex linear projection (CLP) layer. This layer does both filtering and pooling in one operation, and produces set of ASR features which can be fed to the backend neural network AM. For single channel and multi-channel settings, the model obtained via joint training of CLP and AM achieves superior performance compared to both Log Mel and Raw waveform model using Convolutional Neural Networks (CNNs). In addition, it is computationally more efficient than CNN based models.

2. Time Frequency Duality in ASR Feature Extraction

The ultimate goal of a feature extraction method in ASR is to represent a window of speech samples with a feature vector that encodes the maximum information about the signal. Time domain feature extraction involves *convolution* followed by a *pooling* process. Each filter is *convolved* with the input signal to extract a *revertible* representation of the signal. *Pooling* is then applied to each filter response to further compress and remove the *redundant* information and also to induce *invariance* against small noises in the input signal. To avoid the complexity of the convolution operation, the above process can be implemented in the frequency domain using duality theorem.

Depending on the complexity of the analytical expression for the filters and the pooling operation, feature extraction methods are preferred to be implemented in either time or frequency domain. The Mel features are derived by element-wise multiplication of the magnitude spectrum with positive Mel filter weights followed by ${\ell_2}^2$ pooling [1]. Gammatone features are computed by convolution of the time signal with Gammatone filters followed by an average pooling [18].

Previous efforts on joint learning of ASR filter bank and AM are in time domain, typically using CNNs [14]. The m point filter h_i is convolved with a segment of speech signal x containing n samples. The filter support size m is determined through *extensive* experimentation. Because of the duality theorem between time and frequency domain, the frequency domain equivalence of the circular convolution is $X \odot H_i$ where X and H_i are the Fourier transforms of x and h_i, respectively, and $\odot$ is the element-wise multiplication. Here, H_i has the same FFT size as X and can be learned as a neural network layer without any constraint on its time domain equivalent support size.

The $m + n - 1$ dimensional output of the convolution is projected to a single scalar through a pooling function $f_p : \mathbb{R}^{m+n-1} \to \mathbb{R}$. In CNN literature, the deterministic functions

http://dx.doi.org/10.21437/Interspeech.2016-1459

such as max [19], average [20], or ℓ_p [21] as well as learnable pooling functions using a small multi-layer perceptron [22] are used to aggregate the multi-dimensional convolution output into a scalar. In a few cases there is a clear duality between time and frequency domain pooling operations. For example, the ${\ell_2}^2$ pooling in the time domain is equivalent to ${\ell_2}^2$ pooling in the frequency domain due to Parseval's theorem [23], ${\ell_2}^2(x) = {\ell_2}^2(\mathscr{F}(x))$. For pooling operations such as max or $\ell_{p>2}$ norm, the Fourier transform derivation is not straight forward. Instead this paper proposes to use summation in the frequency domain for pooling. This simple operation has the advantage that the final model can be expressed as a projection of X into a lower dimensional space, through a simple matrix vector multiplication which can intuitively explain the ultimate goal of feature extraction, *projection*.

3. Complex Linear Projection Model

We propose Complex Linear Projection (CLP) feature extraction for ASR:

$$x \xrightarrow{\mathscr{F}} X \xrightarrow{W} Y = WX \xrightarrow{|.|} |Y| \tag{1}$$

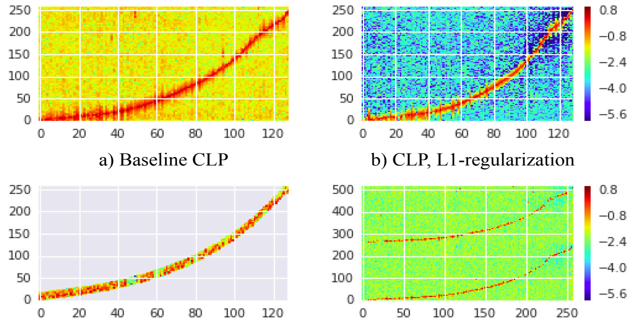
W is a complex matrix in $\mathbb{C}^{p\times(N+1)}$ where p is the projection size which can also be interpreted as number of filters. N is half of the FFT size of $\mathscr{F}(x)$. Since x is a real valued signal, the projection matrix W is only applied to the first half of FFT vector i.e. $X \in \mathbb{C}^{N+1}$. For ease of notations, we will notate this complex vector as $X = X_R + jX_I$ and the projection matrix with $W = W_R + jW_I$ where the real and imaginary parts are the real valued matrices of same dimension as W. Since the rest of network is real, we take element-wise norm of vector Y followed by a logarithmic compression and pass it to the rest of network.

The complex linear layer is embedded as the first layer of the neural network AM. The complex weights are jointly learned with the rest of the network parameters using the ASR optimization cost. While the above process involves a complex matrix-vector multiplication, to ensure that all the values and gradients are in the real domain, the complex layer is implemented by four linear matrix vector multiplications. Since norm of $Y = WX$ is passed to the next layers, we can directly compute $|Y|$ in the real domain:

$$\begin{aligned} |Y| &= \left[\Re\{Y\}^2 + \Im\{Y\}^2\right]^{1/2} \\ \Re\{Y\} &= W_R X_R - W_I X_I \\ \Im\{Y\} &= W_R X_I + W_I X_R \end{aligned} \tag{2}$$

3.1. CLP Filter bank

The filter bank parameters of the complex projection model are jointly learnt with the AM network. The AM is a Long Short Term Memory (LSTM) model and shared CLP layer is integrated into time steps of the first LSTM layer. Figure 1 presents the logarithm of the magnitude weights of the CLP projection matrix for single channel and multi-channel setting. Each column corresponds to a row of the complex matrix W. The filter bank converges to set of narrow-band bandpass filters which nonuniformly spaced across frequency. Due to the sparse nature of filters in the frequency domain we use $L1$ regularization to assist training, Figure 1-b. Same effect can be obtained by *constraining* the weight matrix to only learn a frequency range around Bark scale Figure 1-c. Our experiments show that this model achieves similar performance as $L1$ regularized model with fewer number of parameters. Finally Figure 1-d shows the learned filters for 2-channel setting. Without applying any signal processing beamforming technique, the model is able to learn best combination rule from the channels FFT vectors.



a) Baseline CLP

b) CLP, L1-regularization

c) CLP, hard zero constraint

d) 2-channel CLP with L1

Figure 1: The trained CLP filters on english task. (a,b,c) are single channel (d) 2-channel. The columns are sorted based on the center frequency.

3.2. Comparison with Mel

The following presents the main steps toward extraction of the Mel features on spectrum energy:

$$x \xrightarrow{\mathscr{F}} X \xrightarrow{|.|^2} |X|^2 \xrightarrow{Mel} Y = M|X|^2 \tag{3}$$

where M is the Mel filter bank matrix for which all the elements are positive. The i^{th} element of Mel feature vector is $Y_i^{\text{Mel}} = \sum_j \left({M'}_i \odot |X|\right)_j^{\,2}$ where M' is a matrix which each element is simply square root of the corresponding element in M, ${M'}_{ij} = {M_{ij}}^{0.5}$.

The output of the i^{th} filter in CLP model is $Y_i^{\text{CLP}} = \sum_j \left(W_i \odot X\right)_j$. The CLP model is different from Mel in terms of their pooling operation; Mel uses ${\ell_2}^2$ pooling while CLP uses a simple summation pooling. The Mel filter bank operates on $|X|$ features while the CLP model operates on $X = |X|e^{j\theta_X}$. This means that the phase information in X is removed in the Mel model while it is preserved in CLP model. In speech recognition, there is a long-standing debate about the importance of phase for single microphone recognition. However, it is generally agreed that phase information is necessary for multi-channel processing. The phase information preserves relative delay of the speech signal at each microphone which is the main component of many enhancement techniques [24, 25, 16].

3.3. Time Domain Equivalent

Lemma 1. *Summation in the frequency domain is equivalent to weighted average pooling in the time domain. More precisely, if X is the Fourier transform of the $2N$ point signal x, then*

$$\sum_{k=0}^{N} X_k = \sum_{n=0}^{2N-1} \alpha_n x_n \tag{4}$$

where,

$$\alpha_n = \begin{cases} N+1 & n = 0 \\ \coth\left(j\frac{\pi n}{2N}\right) & \operatorname{mod}(n,2) = 1 \\ 1 & \operatorname{mod}(n,2) = 0, \;\; n \neq 0 \end{cases} \tag{5}$$

Proof. This can easily be proved using the geometric sum:

$$
\begin{aligned}
\sum_{k=0}^{N} X_k &= \sum_{k=0}^{N}\sum_{n=0}^{2N-1} x[n]\exp\left(-j\frac{2\pi k}{2N}n\right) \\
&= \sum_{n=0}^{2N-1} x[n]\left(\sum_{k=0}^{N}\exp\left(-j\frac{\pi n}{N}k\right)\right) \\
&= (N+1)\,x[0] \\
&+ \sum_{n=1}^{2N-1} x[n]\left(\frac{\exp\left(-j\frac{\pi n}{N}(N+1)\right)-1}{\exp\left(-j\frac{\pi n}{N}\right)-1}\right) \\
&= \sum_{n=0}^{2N-1}\alpha_n x[n]. \qquad (6)
\end{aligned}
$$

□

Proposition 1. *The projection in the frequency domain is equivalent to convolution followed by a weighted average pooling:*

$$\sum_{j=0}^{N} W_{ij}X_j \longleftrightarrow \sum_{j=0}^{2N-1}\alpha_j\left(w_i * x\right)[j] \qquad (7)$$

The i^{th} element of the CLP model is:

$$
\begin{aligned}
Y_i^{\text{CLP}} &= \sum_{j=0}^{N} W_{ij}X_j \\
&= \sum W_i \odot X \\
&\overset{\mathscr{F}}{\longleftrightarrow} \sum_{j=0}^{2N-1}\alpha_j\left(w_i * x\right)[j] \qquad (8)
\end{aligned}
$$

which is derived by of Lemma 1 and duality theorem.

4. Experiments

The Complex Linear Projection model is a *frequency* domain feature extraction method which can be *jointly* optimized with the acoustic model. To establish the effectiveness of this method, the experiments are designed to evaluate CLP in three aspects, (1) effectiveness of CLP model to achieve state-of-the-art performance using Log Mel and Raw Waveform model, (2) benefits of learning in frequency over time, and (3) benefits of joint training of features and AM over the separate feature extraction process like Log Mel.

Datasets. Two anonymized datasets were used: *Multi-channel American English* (en_us), consisting of 2000 hours of spontaneous speech from anonymized, human transcribed Voice Search data. This is denoted as clean dataset. The noisy set is created by artificially corrupting clean utterances using a room simulator, adding varying degrees of noise and reverberation. The simulation uses an 8-channel linear microphone array, with inter-microphone spacing of 2 cm were both noise and target speaker locations are changing between utterances. Channel one is used for single channel experiments, while experiments on 2-channel speech use channels 1 and 8, which are separated by 14cm. More details can be found in [16]. *Taiwanese Mandarin* (cmn_hant_tw) consisting of more than 30000 utterances, hand transcribed and anonymized. The test sets for each language are separately anonymized, human-transcribed Voice Search data of about 25 hours each. The noisy evaluation set includes simulated speech, synthesized similarly to the train set.

Models. The backend AM model, LDNN, is 3 LSTM layers followed by a Rectified Linear Unit (ReLu) layer and a Linear layer [16]. Each LSTM consist of 832 cells with 512 dimensional projection layer. The fully connected ReLu layer has 1024 nodes which is followed by a linear layer of 512 nodes and a softmax layer. The en_us and cmn_hant_tw models use 13522 and 10538 context-dependent state output targets, respectively. The Log Mel model consist of a LDNN model trained on separately extracted Log Mel features. The feature dimension per frame into LDNN is 128 for single channel and 256 for two channel experiment. The Raw Waveform model jointly train a feature extraction module and the LDNN model. The feature extraction module consists of a convolution, ReLu and max pooling layer plus logarithmic compression. The Raw model parameters are set as in [14]. The parameters are set such that feature extraction module extract 128 and 256 dimensional feature for single channel and two channel experiments, respectively. For two channel model, the 256 dimensional features extracted per channel are added so that same feature dimension as Log Mel model goes to time steps of the first LSTM layer. The CLP model also replace the separateley processed Log Mel features by a feature extraction module which is jointly trained with backend LDNN model. The feature extraction module follow the steps in Eq 1 followed by a logarithmic compression. Unlike the Raw model, the baseline CLP model has similar linear complexity as Log Mel model. All models trained with cross-entropy (CE) loss using ASGD training with 200 multi-core machines [26].

Baseline. Table 1 presents the baseline WERs for Log Mel and Raw waveform models across three different window sizes. Typically a window size of 25 msec with a 10 msec shift is used in ASR, which requires zero padding to perform the FFT. To avoid zero padding, a window size of 32 msec is also considered. In addition, for empirical investigation of longer window effects, Table 1 contains the WER results for 64 msec window. Longer window contains more temporal information as well as localization for multi-channel processing which result in WER improvement over shorter windows.

MODEL	1-CHANNEL		2-CHANNEL	
	LOG MEL	RAW	LOG MEL	RAW
25 MSEC	23.4	23.7	21.8	21.5
32 MSEC	22.8	23.4	21.3	21.2
64 MSEC	21.8	22.5	20.7	21.2

Table 1: WER for the baseline models on en_us dataset.

Baseline CLP. For the baseline CLP model, the frequency range between 125 Hz to 7500 Hz were kept, similar to the Log Mel model. Since the pre-emphasis filter used during FFT computation can harm the phase information, it is removed for the CLP models. The pre-emphasis was helping Log Mel performance, so it is kept for Log Mel models. To enforce sparsity, $L1$ regularization was used for learning the CLP component weights. Similar performance was achieved by constraining the frequency ranges covered by each filter, Figure 1-c. The baseline CLP model results are shown in Table 2. The CLP model performance is in par with the Log Mel model for single channel and yields a gain of about 4% over the 2-channel baseline models in Table 1.

Learning in Frequency versus Time. We argue learning in the frequency domain is more efficient compare to the time domain in three aspects, *parameter tuning*, *computation efficiency*, and *optimization*. The efficiency of the Raw waveform model is devoted to the proper choice of the convolution filter

MODEL	1-CHANNEL	2-CHANNEL
25 MSEC	23.2	21.5
32 MSEC	22.8	20.9
64 MSEC	22.0	20.5

Table 2: WER for CLP baseline models on en_us dataset.

size as well as pooling size. As Table 3 shows, to capture long temporal dependencies of speech, the best WER achieved using filter support size of 352 samples. However this leads to significant computation bottleneck during training and run time. The time implementation of p filters in the time domain with filter support size of d and stride of one for a $2N$ point signal and full convolution requires $2pd \times (2N + d + 1)$ operations. Similar operation can be done by $16pN$ operations in the frequency domain. Hence, there is a factor of kernel size difference between run-time of the same operation in the time and frequency domains. Table 4, compares the total number of parameters as well as Add and Multiplications for the CLP and Raw model. The CLP model brings 55-fold reduction in add and multiplication operations.

SUPPORT SIZE	112	192	272	352	432
WER	25.8	23.8	23.4	23.4	23.5

Table 3: WER for different convolution filter size in Raw model.

MODEL	RAW	CLP
NUM-PARAMS	45K	66K
NUM-ADD-MULT	14.51 M	263.17 K

Table 4: Computation efficiency comparison.

Defining the speech recognition filters in the frequency domain is appealing partly because acoustic filters are known to be narrow-band in the frequency domain [27, 28, 29] . This translate to very small non-zero support in the frequency domain. However, based on the Gabor limit [30], the dual of band limited frequency filters are wide band in the time domain and *vice versa*. So, the time representation is expected to have more non-zero entries. In other words, the frequency representation of the speech filters tend to be *sparser* in the frequency domain compared to the raw domain. The tendency of filters to be sparser in frequency representation greatly facilitates *optimization* [31].

Joint versus Separate Feature Extraction. While joint training of feature extraction module and AM network outperforms the separate model for multi-channel setting, the benefits of joint model is not clear for single channel task, Table 2 and Table 1. We argue that joint training is still beneficial for two reasons. First, it removes the extensive *feature engineering* in domain specific problems. For tonal languages like Taiwanese Mandarin, the usual practise in ASR is feature engineering by appending tonal features such as pitch to the Log Mel features. There is always a question of what is the best pitch extraction method and how to configure its parameters. As Table 5 shows, both joint models offer around 12% relative improvement over Log Mel baseline in single channel setting. Furthermore, Figure 2 shows how the CLP center frequencies are different for cmn_hant_tw versus en_us. It also compares the center frequency curves for noisy and clean en_us set.

MODEL	LOGMEL	RAW	CLP
CMN_HANT_TW	17.2	16.8	16.6

Table 5: Joint vs. Separate: Language dependecy.

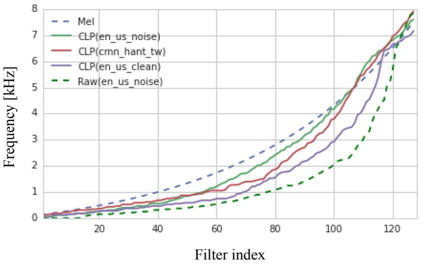


Figure 2: The center frequencies vs. filter index.

In addition, the joint models provide the possibility of investigating alternative feature extraction models such as high dimensional features. Table 6 shows how increasing number of filters can benefit the CLP model. These models obtain an additional 2-4% relative improvement over separate Log Mel and Raw models in Table 1. Increasing filter bank size is not possible for the Log Mel model with the same size FFT since there is a possibility to have one or no FFT tap per filter. On the other hand, increasing number of filters for Raw model is not feasible due to the computational cost and additional parameter tuning.

MODEL	1-CHANNEL		2-CHANNEL	
	$P = 128$	$P = 1280$	$P = 256$	$P = 2560$
32 MSEC	22.8	22.2	20.9	20.2
64 MSEC	22.0	21.6	20.5	19.7

Table 6: WER of CLP when increasing number of filters.

5. Conclusions

This paper casted feature extraction mechanism as a problem of learning a cascade of two stages, filtering followed by a pooling stage and identified these stages in the conventional Log Mel model and the recently proposed CNN based raw waveform models. A feature learning mechanism in the frequency domain was introduced consist of a complex-valued linear layer pre-pended to the acoustic model which embodying both filtering and pooling compactly and efficiently in the frequency domain. All the parameters of the resulting deep neural network architecture were then jointly optimized using cross-entropy cost function to learn features which are most suited for the recognition task. This joint optimization provides the capability of learning optimal features for different flavors of speech recognizers, e.g. as used for tonal language such as Taiwanese Mandarin. In addition, the properties of the complex-valued linear projection layer as an alternative of the convolutional neural networks were explored. It was shown that the complex-valued linear layer is computationally efficient compared to the CNN based time domain version. Furthermore, this model does not discard any information in the signal, including the time delay or phase information from multiple channels in a microphone array which makes it appropriate to automatically learn the optimal feature extraction parameters in the multi-channel setting without any signal processing based beamforming approach. Finally, the empirical comparisons Demonstrated the effectiveness of the proposed model to achieve state-of-the-art performance.

6. Acknowledgements

The authors would like to thanks all members of Google acoustic modeling team.

7. References

[1] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 28, no. 4, pp. 357–366, 1980.

[2] A. J. Bell and T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural computation*, vol. 7, no. 6, pp. 1129–1159, 1995.

[3] ——, "Learning the higher-order structure of a natural sound," *Network: Computation in Neural Systems*, vol. 7, no. 2, pp. 261–266, 1996.

[4] M. S. Lewicki, "Efficient coding of natural sounds," *Nature neuroscience*, vol. 5, no. 4, pp. 356–363, 2002.

[5] L. Burget and H. Heřmanskỳ, "Data driven design of filter bank for speech recognition," in *Text, Speech and Dialogue*. Springer, 2001, pp. 299–304.

[6] A. Biem, E. McDermott, and S. Katagiri, "A discriminative filter bank model for speech recognition." in *Eurospeech*. Citeseer, 1995.

[7] 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.

[8] N. Jaitly and G. Hinton, "Learning a better representation of speech soundwaves using restricted boltzmann machines," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 5884–5887.

[9] T. N. Sainath, B. Kingsbury, A.-r. Mohamed, and B. Ramabhadran, "Learning filter banks within a deep neural network framework," in *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*. IEEE, 2013, pp. 297–302.

[10] D. Palaz, R. Collobert, and M. M. Doss, "Estimating phoneme class conditional probabilities from raw speech signal using convolutional neural networks," *arXiv preprint arXiv:1304.1018*, 2013.

[11] Z. Tüske, P. Golik, R. Schlüter, and H. Ney, "Acoustic modeling with deep neural networks using raw time signal for lvcsr." in *INTERSPEECH*, 2014, pp. 890–894.

[12] Y. Hoshen, R. J. Weiss, and K. W. Wilson, "Speech acoustic modeling from raw multichannel waveforms," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4624–4628.

[13] P. Golik, Z. Tüske, R. Schlüter, and H. Ney, "Convolutional neural networks for acoustic modeling of raw time signal in lvcsr," in *Sixteenth Annual Conference of the International Speech Communication Association*, 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] D. Palaz, R. Collobert *et al.*, "Analysis of cnn-based speech recognition system using raw speech as input," in *Proceedings of Interspeech*, no. EPFL-CONF-210029, 2015.

[16] 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 *Proc. ASRU*, 2015.

[17] T. N. Sainath, R. J. Weiss, K. W. Wilson, A. Narayanan, and M. Bacchiani, "Factored Spatial and Spectral Multichannel Raw Waveform CLDNNs," in *to appear in Proc. ICASSP*, 2016.

[18] A. Aertsen, P. Johannesma, and D. Hermes, "Spectro-temporal receptive fields of auditory neurons in the grassfrog," *Biological Cybernetics*, vol. 38, no. 4, pp. 235–248, 1980.

[19] M. A. Ranzato, F. J. Huang, Y.-L. Boureau, and Y. LeCun, "Unsupervised learning of invariant feature hierarchies with applications to object recognition," in *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*. IEEE, 2007, pp. 1–8.

[20] B. B. Le Cun, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Handwritten digit recognition with a back-propagation network," in *Advances in neural information processing systems*. Citeseer, 1990.

[21] J. Feng, B. Ni, Q. Tian, and S. Yan, "Geometric p-norm feature pooling for image classification," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 2609–2704.

[22] M. Lin, Q. Chen, and S. Yan, "Network in network," *arXiv preprint arXiv:1312.4400*, 2013.

[23] M.-A. Parseval des Chłnes, "Mmoire sur les sries et sur l'intgration complte d'une quation aux diffrences partielles linaire du second ordre, coefficients constants." *Mmoires prsents lInstitut des Sciences, ettres et Arts, par divers savans, et lus dans ses assembles*, vol. 1, no. 1, pp. 638–648, 1806.

[24] T. Adali and S. Haykin, *Adaptive signal processing: next generation solutions*. John Wiley & Sons, 2010, vol. 55.

[25] J. Benesty, J. Chen, and Y. Huang, *Microphone array signal processing*. Springer Science & Business Media, 2008, vol. 1.

[26] 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.

[27] D. J. Field, "Relations between the statistics of natural images and the response properties of cortical cells," *JOSA A*, vol. 4, no. 12, pp. 2379–2394, 1987.

[28] ——, "What is the goal of sensory coding?" *Neural computation*, vol. 6, no. 4, pp. 559–601, 1994.

[29] R. Linsker, "Perceptual neural organization: some approaches based on network models and information theory," *Annual review of Neuroscience*, vol. 13, no. 1, pp. 257–281, 1990.

[30] D. Gabor, "Theory of communication. part 1: The analysis of information," *Electrical Engineers-Part III: Radio and Communication Engineering, Journal of the Institution of*, vol. 93, no. 26, pp. 429–441, 1946.

[31] D. Yu, F. Seide, G. Li, and L. Deng, "Exploiting sparseness in deep neural networks for large vocabulary speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 4409–4412.