Multichannel Signal Processing with Deep Neural Networks for Automatic Speech Recognition

Tara N. Sainath, *Senior Member, IEEE,* Ron J. Weiss, *Member, IEEE,*
Kevin W. Wilson, *Member, IEEE*, Bo Li, *Member, IEEE*, Arun Narayanan, *Member, IEEE*, Ehsan Variani *Member, IEEE*, Michiel Bacchiani, *Member, IEEE*, Izhak Shafran *Member, IEEE*, Andrew Senior *Member, IEEE*, Kean Chin *Member, IEEE*, Ananya Misra *Member, IEEE*, Chanwoo Kim *Member, IEEE*

***Abstract*—Multichannel ASR systems commonly separate speech enhancement, including localization, beamforming and postfiltering, from acoustic modeling. In this paper, we perform multichannel enhancement jointly with acoustic modeling in a deep neural network framework. Inspired by beamforming, which leverages differences in the fine time structure of the signal at different microphones to filter energy arriving from different directions, we explore modeling the raw time-domain waveform directly. We introduce a neural network architecture which performs multichannel filtering in the first layer of the network and show that this network learns to be robust to varying target speaker direction of arrival, performing as well as a model that is given oracle knowledge of the true target speaker direction. Next, we show how performance can be improved by *factoring* the first layer to separate the multichannel spatial filtering operation from a single channel filterbank which computes a frequency decomposition. We also introduce an adaptive variant, which updates the spatial filter coefficients at each time frame based on the previous inputs. Finally we demonstrate that these approaches can be implemented more efficiently in the frequency domain. Overall, we find that such multichannel neural networks give a relative word error rate improvement of more than 5% compared to a traditional beamforming-based multichannel ASR system and more than 10% compared to a single channel waveform model.**

***Index Terms*—Beamforming, Deep Learning, Noise-robust speech recognition**

I. Introduction

While state-of-the-art automatic speech recognition (ASR) systems perform reasonably well in close-talking microphone conditions, performance degrades in conditions when the microphone is far from the user. In such farfield cases, the speech signal is degraded by reverberation and additive noise. To improve recognition in such cases, ASR systems often use signals from multiple microphones to enhance the speech signal and reduce the impact of reverberation and noise [1], [2], [3].

Multichannel ASR systems often use separate modules to perform recognition. First, microphone array speech enhancement is applied, typically broken into localization, beamforming and postfiltering stages. The resulting single channel enhanced signal is passed to an conventional acoustic model [4], [5]. A commonly used enhancement technique is filter-and-sum beamforming [2], which begins by aligning signals from different microphones in time (via localization) to adjust for the propagation delay from the target speaker to each microphone. The time-aligned signals are then passed through a filter (different for each microphone) and summed to enhance the signal from the target direction and to attenuate noise coming from other directions. Commonly used filter design criteria are based on Minimum Variance Distortionless Response (MVDR) [3], [6] or multichannel Wiener filtering (MWF) [1].

When the end goal is to improve ASR performance, tuning the enhancement model independently from the acoustic model might not be optimal. In an early effort to address this issue [7] proposed a likelihood-maximizing beamforming (LIMABEAM) which optimizes beamformer parameters jointly with those of the acoustic model. This technique was shown to outperform conventional techniques such as delay-and-sum beamforming (i.e. filter-and-sum where the filters consist of impulses). Like most enhancement techniques, LIMABEAM is a model-based scheme and requires an iterative algorithm that alternates between acoustic model inference and enhancement model parameter optimization. Contemporary acoustic models are generally based on neural networks, optimized using a gradient learning algorithm. Combining model-based enhancement with an acoustic model that uses gradient learning can lead to considerable complexity, e.g. [8].

In this paper we extend the idea of performing beamforming jointly with acoustic modeling from [7], but do this within the context of a deep neural network (DNN) framework by training an acoustic model directly on the raw signal. DNNs are attractive because they have been shown to be able to perform feature extraction jointly with classification [9]. Previous work has demonstrated the possibility of training deep networks directly on raw, single channel, time domain waveform samples [10], [11], [12], [13], [14], [15]. The goal of this paper is to explore a variety of different joint enhancement/acoustic modeling DNN architectures that operate on multichannel signals. We will show that jointly optimizing both stages is more effective than techniques which cascade independently tuned enhancement algorithms with acoustic models.

Since beamforming takes advantage of the fine time structure of the signal at different microphones, we begin by modeling the raw time-domain waveform directly. In this model, introduced in [14], [16], the first layer consists of multiple time convolution filters, which map the multiple microphone signals down to a single time-frequency representation. As we will show, this layer learns bandpass filters which are spatially selective, often learning several filters with nearly identical

frequency response, but with nulls steered toward different directions of arrival. The output of this spectral filtering layer is passed to an acoustic model, such as a convolutional long short-term memory, deep neural network (CLDNN) acoustic model [17]. All layers of the network are trained jointly.

As described above, it is common for multichannel speech recognition systems to perform spatial filtering independently from single channel feature extraction. With this in mind, we next investigate explicitly factorizing these two operations to be separate neural network layers. The first layer in this "factored" raw waveform model consists of short-duration multichannel time convolution filters which map multichannel inputs down to a single channel, with the idea that the network might learn to perform broadband spatial filtering in this layer. By learning several filters in this "spatial filtering layer", we hypothesize that the network can learn filters for multiple different spatial look directions. The single channel waveform output of each spatial filter is passed to a longer-duration time convolution "spectral filtering layer" intended to perform finer frequency resolution spectral decomposition analogous to a time-domain auditory filterbank as in [15]. The output of this spectral filtering layer is also passed to an acoustic model, and all layers of the network are trained jointly.

One issue with the two architectures above is that once weights are learned during training, they remain fixed for each test utterance. In contrast, some beamforming techniques, such as the generalized sidelobe canceller [18], update weights adaptively within each utterance. We explore an adaptive neural net architecture, where an LSTM is used to predict spatial filter coefficients that are updated at each frame. These filters are used to filter and sum the multichannel input, replacing the "spatial filtering layer" of the factored model described above, before passing the enhanced single channel output to a waveform acoustic model.

Finally, since convolution between two time domain signals is equivalent to the element-wise product of their frequency domain counterparts, we investigate speeding up the raw waveform neural network architectures described above by consuming the complex-valued fast Fourier transform of the raw input and implementing filters in the frequency domain.

II. Experimental Setup

A. Data

We conduct experiments on a dataset comprised of about 2,000 hours of noisy training data consisting of 3 million English utterances. This data set is created by artificially corrupting clean utterances using a room simulator to add varying degrees of noise and reverberation. The clean utterances are anonymized and hand-transcribed voice search queries, and are representative of Google's voice search traffic. Noise signals, which include music and ambient noise sampled from YouTube and recordings of "daily life" environments, are added to the clean utterances at SNRs ranging from 0 to 20 dB, with an average of about 12 dB. Reverberation is simulated using the image method [19] – room dimensions and microphone array positions are randomly sampled from 100 possible room configurations with T_{60}s ranging from 400 to 900 ms, with an average of about 600 ms. The simulation uses an 8-channel uniform linear microphone array, with inter-microphone spacing of 2 cm. Both noise source location and target speaker locations change between utterances; the distance between the sound source and the microphone array varies between 1 to 4 meters.

The primary evaluation set consists of a separate set of about 30,000 utterances (over 20 hours), and is created by simulating similar SNR and reverberation settings to the training set. Care was taken to ensure that the room configurations, SNR values, T_{60} times, and target speaker and noise positions in the evaluation set differ from those in the training set. The microphone array geometry between the training and simulated test sets is identical. Most of the results we report will be on this test set.

We obtained a second "rerecorded" test set by playing the evaluation set and the noises separately using a mouth simulator and a speaker, respectively, in a living room setting. The signals are recorded using a 7-channel circular microphony array with a radius of 3.75 cm. Assuming an x-axis that passes through two diagonally opposite mics along the circumference of the array, the angle of arrival of the target speaker ranges from 0 to 180 degrees. Noise originates from locations different from the target speaker. The distance of the sources to the target ranges from 1 to 6 meters. To create noisy rerecorded eval sets, the rerecorded speech and noise signals are mixed artificially after scaling noise to obtain SNRs ranging from 0 to 20 dB. The distribution of the SNR matches the distribution used to generate the simulated evaluation set. The average T_{60} for this set is around 200ms. We create 4 versions of the rerecorded sets to measure generalization performance of our models. The first two have rerecorded speech w/o any added noise. The mic array is placed at the center of the room and closer to the wall, respectively, to capture reasonably different reverberation characteristics. The remaning two subsets correspond to the noisy versions of these sets.

B. Baseline Acoustic Model

We compare the models proposed in this paper to a baseline CLDNN acoustic model trained using log-mel features [17] computed with a 25ms window and a 10ms hop. Single channel models are trained using signals from channel 1, $C = 2$ channel models use channels 1 and 8 (14 cm spacing), $C = 4$ channel models use channels 1, 3, 6, and 8 (14 cm array span, with adjacent microphone spacing of 4cm-6cm-4cm).

The baseline CLDNN architecture is shown in the `CLDNN` module of Figure 1. First, the `fConv` layer performs convolution across the frequency dimension of the input log-mel time-frequency feature to gain some invariance to pitch and vocal tract length. The architecture used for this convolutional layer is similar to that proposed in [20]. Specifically, a single convolutional layer with 256 filters of size 1×8 in time-frequency is used. Our pooling strategy is to use non-overlapping max pooling along the frequency axis, with a pooling size of 3. The pooled output is given to a 256-dimensional linear low-rank layer.

The output of frequency convolution is passed to a stack of LSTM layers, which model the signal across long time scales. We use 3 LSTM layers, each comprised of 832 cells, and a 512 unit projection layer for dimensionality reduction following [21]. Finally, we pass the final LSTM output to a single fully connected DNN layer comprised of 1,024 hidden units. Due to the high dimensionality of the 13,522 context-dependent state output targets used by the language model, a 512-dimensional linear output low rank projection layer is used prior to the softmax layer to reduce the number of parameters in the overall model [22]. Some experiments in the paper do not use the frequency convolution layer, and we will refer to such acoustic models as LDNNs. It is important to note that all methods presented in this paper use an LSTM-type architecture, which has been shown to work much better with larger amounts of data [15] and is much more well-suited for our task.

During training, the CLDNN is unrolled for 20 time steps and trained using 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 helps to better predict the current frame [17].

All neural networks are trained using asynchronous stochastic gradient descent (ASGD) optimization [23] to optimize a cross-entropy (CE) criterion. Unless otherwise indicated, all word error rate (WER) numbers in this paper are presented with the CE criterion. Additional sequence training experiments also use distributed ASGD [24]. All networks have 13,522 context-dependent (CD) output targets. The weights for all CNN and DNN layers are initialized using the Glorot-Bengio strategy [25], while those of all LSTM layers are randomly initialized using a uniform distribution between -0.02 and 0.02. We use an exponentially decaying learning rate, initialized to 0.004 and decaying by 0.1 over 15 billion frames.

III. Multichannel Raw Waveform Neural Network

A. Motivation

The proposed multichannel raw waveform CLDNN is related to filter-and-sum beamforming, a generalization of delay-and-sum beamforming which filters the signal from each microphone using a finite impulse response (FIR) filter before summing them. Using similar notation to [7], filter-and-sum enhancement can be written as follows:

$$y[t] = \sum_{c=0}^{C-1} \sum_{n=0}^{N-1} h_c[n] x_c[t-n-\tau_c] \quad (1)$$

where $h_c[n]$ is the nth tap of the filter associated with microphone c, $x_c[t]$, is the signal received by microphone c at time t, τ_c is the steering time difference of arrival induced in the signal received by microphone c used to align it to the other array channels, and $y[t]$ is the output signal. C is the number of microphones in the array and N is the number of FIR filter taps.

B. Multichannel filtering in the time domain

Enhancement algorithms implementing Equation 1 generally depend on an estimate of the steering delay τ_c obtained using a separate localization model, and they compute filter parameters $h_c[n]$ by optimizing an objective such as MVDR. In contrast, our aim is to allow the network to jointly estimate steering delays and filter parameters by optimizing an acoustic modeling classification objective. The model captures different steering delays using a bank of P multichannel filters. The output of filter $p \in \{0, \dots, P-1\}$ can be written as follows:

$$y^p[t] = \sum_{c=0}^{C-1} \sum_{n=0}^{N-1} h_c^p[n] x_c[t-n] = \sum_{c=0}^{C-1} x_c[t] * h_c^p \quad (2)$$

where the steering delays are implicitly absorbed into the filter parameters $h_c^p[n]$. In this equation, '*' denotes the convolution operation.

The first layer in our raw waveform architecture [26] implements Equation 2 as a multichannel convolution (in time) with a FIR spatial filterbank $h_c = \{h_c^1, h_c^2, \dots h_c^P\}$ where $h_c \in \Re^{N \times P}$ for $c \in 0, \cdots, C-1$. Each filter h_c^p is convolved with the corresponding input channel x_c, and the overall output for filter p is computed by summing the result of this convolution across all channels $c \in \{0, \cdots, C-1\}$. The operation within each filter is equivalent to an FIR filter-and-sum beamformer, except that it does not explicitly shift the signal in each channel by an estimated time difference of arrival. As we will show, the network learns parameters for a fixed set of filters that give good speech recognition performance.

The output signal remains at the same sampling rate as the input signal, which contains more information than is typically relevant for acoustic modeling. In order to produce an output that is invariant to perceptually and semantically identical sounds appearing at different time shifts we pool the outputs in time after filtering [14], [15], in an operation that has an effect similar to discarding the phase in the short-time Fourier transform. Specifically, the output of the filterbank is max-pooled across time to give a degree of short term shift invariance, and then passed through a compressive non-linearity.

As shown in [14], [15], single channel time convolution layers similar to the one described above implement a conventional time-domain filterbank. Such layers are capable of implementing, for example, a standard gammatone filterbank, which consists of a bank of time-domain filters followed by rectification and averaging over a small window. Given sufficiently large P, the corresponding multichannel layer can (and as we will show, does in fact) similarly implement a frequency decomposition in addition to spatial filtering. We will therefore subsequently refer to the output of this layer as a "time-frequency" feature representation.

A schematic of the multichannel time convolution layer is shown in the tConv block of Figure 1. First, we take a small window of the raw waveform of length M samples for each channel C, denoted as $\{x_0[t], x_1[t], \dots, x_{C-1}[t]\}$ for $t \in 1, \cdots, M$. The signal from each channel x_c is convolved with a bank of P filters with N taps $h_c[n] = \{h_c^1[n], h_c^2[n], \dots, h_c^P[n]\}$. When the convolution is strided by 1 in time across M samples, the output from the convolution in each channel is $y_c[t] \in \Re^{(M-N+1) \times P}$. After summing

$y_c[t]$ across channels c, we max pool the filterbank output in time (thereby discarding short term phase information), over the entire time length of the output signal $M - N + 1$, to produce $y[t] \in \Re^{1\times P}$. Finally, we apply a rectified non-linearity, followed by a stabilized logarithm compression[1], to produce $z[l]$, a P dimensional frame-level feature vector at frame l. We then shift the window around the waveform by 10ms and repeat this time convolution, producing a sequence of feature frames at 10ms intervals.

To match the time-scale of the log-mel features, the raw waveform features are computed with an identical filter size of 25ms, or $N = 400$ at a sampling rate of 16kHz. The input window size is 35ms ($M = 560$) giving a 10ms fully overlapping pooling window. Our experiments explore varying the number of time-convolutional filters P.

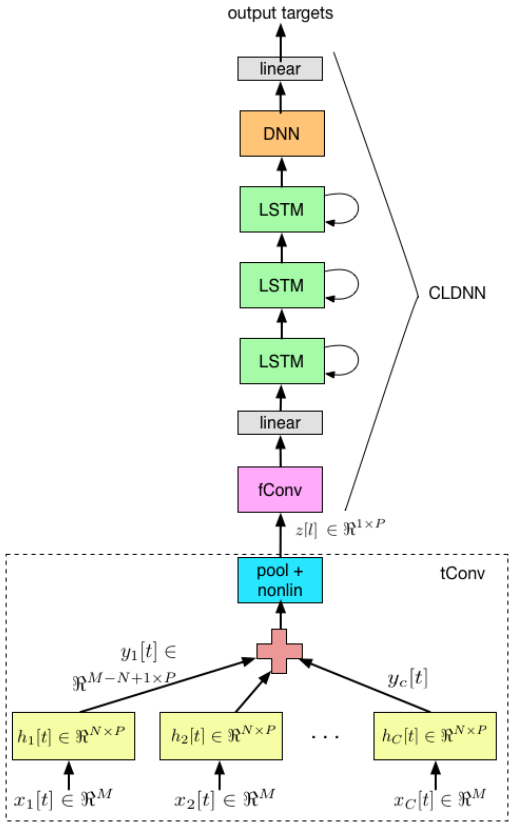

Fig. 1: Multichannel raw waveform CLDNN architecture.

As shown in the CLDNN block of Figure 1, the output of the time convolutional layer (`tConv`) produces a frame-level feature, denoted as $z[l] \in \Re^{1\times P}$. This feature is then passed to a CLDNN model [17] described in Section II, which predicts context dependent state output targets.

C. Filterbank spatial diversity

Figure 2 plots example multichannel filter coefficients and their corresponding spatial responses, or beampatterns, after training for `tConv`. The beampatterns show the magnitude response in dB as a function of frequency and direction of arrival, i.e. each horizontal slice of the beampattern corresponds to the filter's magnitude response for a signal coming from a particular direction, and each vertical slice corresponds to the filter's response across all spatial directions in a particular frequency band. Lighter shades indicate regions of the frequency-directions space which are passed through the filter, while darker shades indicate regions which are filtered out. Within a given beampattern, we refer to the frequency band containing the maximum overall response as the filter's *center frequency* (since the filters are primarily bandpass in frequency), and the direction corresponding to the minimum response in that

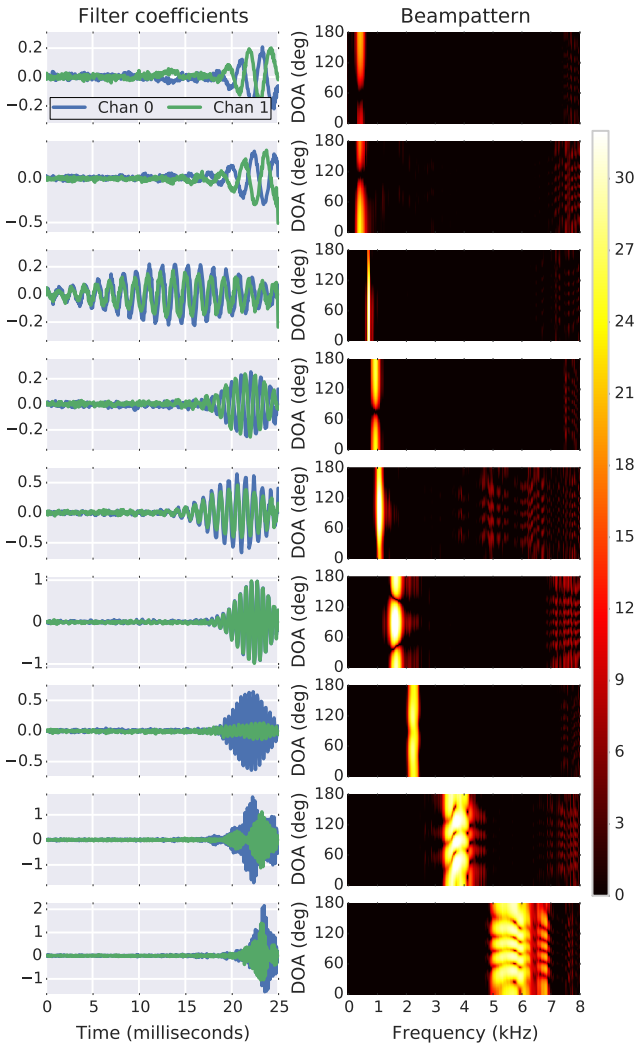

Fig. 2: Example filter coefficients and corresponding spatial response beampatterns learned in a network with 128 `tConv` filters trained on 2 channel inputs. Some filters learned by this network have nearly-identical center frequencies but different spatial responses. For example, the top two example filters both have center frequencies of about 440Hz, but the first filter has a null at a direction of arrival of about 60 degrees, while the second has a null at about 120 degrees. The corresponding phase difference between the two channels of each filter is visible in the time domain filter coefficients plotted on the left.

[1]We use a small additive offset to truncate the output range and avoid numerical instability with very small inputs: $\log(\cdot + 0.01)$.

frequency as the filter's *null direction*.

The network tends to learn filter coefficients with very similar shapes in each channel except they are slightly shifted relative to each other, consistent with the notion of a steering delay τ_c described in Section III. Most filters have a bandpass response in frequency, with bandwidths that increase with center frequency, and many are steered to have stronger response for signals arriving from a particular direction. Approximately two-thirds of the filters in the model shown in Figure 2 demonstrate a significant spatial response, i.e. show a difference of at least 6dB between the direction with the minimum and maximum response at the filter center frequency. Such strong spatial responses are clearly visible in the null near 120 degrees in the second filter, and a similar null near 60 degrees in the fourth filter shown in Figure 2.

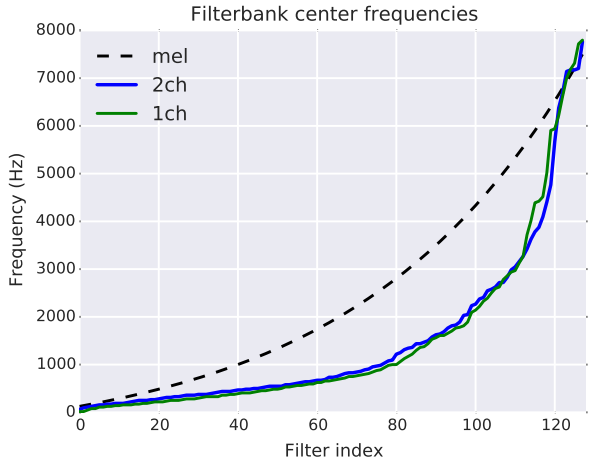

Fig. 3: Comparison of the peak response frequencies of waveform CLDNN filterbanks trained on one- and two-channel inputs to the standard mel frequency scale.

Figure 3 plots the peak response frequencies of filterbanks from networks trained on a one- and two-channel networks of the form shown in 2. The two networks converge to similar frequency scales, both consistently allocating many more filters to low frequencies compared to the mel scale. For example, the learned filterbanks have roughly 80 filters with peak responses below 1000Hz, while a 128-band mel scale has only 40 bands with center frequencies below 1000Hz. The network furthermore tends to learn subsets of filters with the same overall shape and frequency response but tuned to have nulls in different directions, as illustrated by the top two example filters in Figure 2. Such diversity in spatial response gives upstream layers information that can be used to discriminate between signals arriving from different directions.

The ability of the network to exploit directional cues is constrained by the number of filters it uses. By increasing the number of filters, we can potentially improve the spatial diversity of the learned filters and therefore allow the network to better exploit directional cues.

To see how the distribution of null directions learned by the filters corresponds to the direction of arrival (DOA) of noise, Figure 4 plots the distribution of noise DOA for the training and test sets. Notice there is a strong correlation between the noise DOA distrubtion and the learned filter null directions in Figure 5, illustrating that the learned filters are learning to filter out noise.

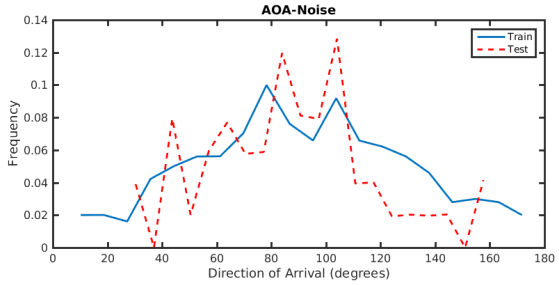

Fig. 4: Distribution of the Direction of Arrival of Noise in Training and Test Set.

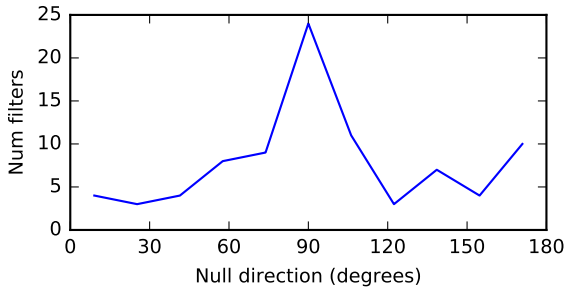

Fig. 5: Histogram of null directions of spatial filters from a 2 channel model with 128 filters. "Null direction" is computed as the direction of minimum response for filters where the minimum response is at least 6dB below the maximum. (Filters for which minimum and maximum directional responses differ by less than 6dB are not included in the plot.)

Table I demonstrates the effect of increasing the number of filters on overall word error rate (WER). Improvements saturate at 128 filters for networks trained on 2 channel inputs, while 4 and 8 channels networks continue to improve with 256 filters. With additional input channels the `tConv` filters are able to learn more complex spatial responses (even though the total array span is unchanged), enabling the network to make use of additional filterbank capacity to improve performance.

Filters	2ch (14cm)	4ch (4-6-4cm)	8ch (2cm)
128	21.8	21.3	21.1
256	21.7	20.8	20.6
512	-	20.8	20.6

TABLE I: WER for raw waveform multichannel CLDNNs with different number of input channels. The inter-microphone spacing is given in parentheses.

Multi-microphone signal processing can help to enhance the signal and suppress noise. Therefore, we should expect to see improvements in WER as the number of microphones increases, especially under more challenging conditions. A breakdown of WER in Figure 6 shows that this is indeed the

case – the performance improvement of the 8 channel system over the 1 and 2 channel systems is largest in low SNR and high reverberation time. Notice also that there is very little difference in performance going from 4 to 8 channels.

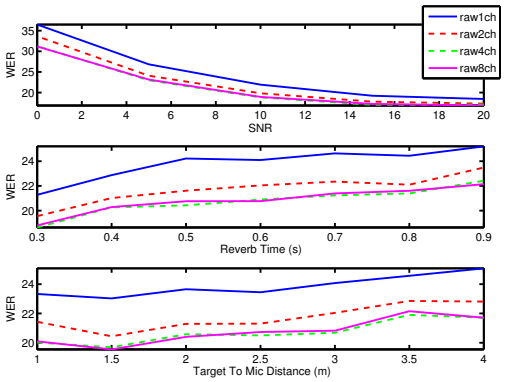

Fig. 6: WER breakdown for multichannel models.

D. Comparison to log-mel

We train baseline multichannel log-mel CLDNNs by computing log-mel features for each channel, and treating these as separate feature maps into the CLDNN. Since the raw waveform model improves as we increase the number of filters, we perform the same experiment for log-mel. Table II shows that for log-mel, neither increasing the number of filters (frequency bands) nor increasing the number of microphone channels has a strong effect on word error rate. Since log-mel features are computed from the FFT magnitude, the fine time structure (stored in the phase), and therefore information about inter-microphone delays, is discarded. Log-mel models can therefore only make use of the weaker inter-microphone level difference cues. However, the multichannel time-domain filterbanks in the raw waveform models utilize the fine time structure and show larger improvements as the number of filters increase.

Filters	2ch (14cm)	4ch (4-6-4cm)	8ch (2cm)
128	22.0	21.7	22.0
256	21.8	21.6	21.7

TABLE II: WER for log-mel multichannel CLDNNs.

Comparing Tables I and II we can see that raw waveform models consistently outperform log-mel, particularly for larger number of channels where more spatial diversity is possible.

E. Comparison to oracle knowledge of speech TDOA

Note that the models presented in the previous subsection do not explicitly estimate the time delay of arrival of the target source arriving at different microphones, which is commonly done in beamforming [2]. Time delay of arrival (TDOA) estimation is useful because time aligning and combining signals steers the array such that the target speech signal is enhanced relative to noise sources coming from other directions.

In this section, we analyze the behavior of raw waveform CLDNNs when the signals are time aligned using the true TDOA calculated using the room geometry. For the delay-and-sum (D+S) approach, we shift the signal in each channel by the corresponding TDOA, average them together, and pass the result into a 1-channel raw waveform CLDNN. For the time-aligned multichannel (TAM) approach, we align the signals in time and pass them as separate channel inputs to a multichannel raw waveform CLDNN. Thus the difference between the multichannel raw waveform CLDNNs described in Section 2 and TAM is solely in how the data is presented to the network (whether or not they are first explicitly aligned to "steer" toward the target speaker direction); the network architectures are identical.

Feature	1ch	2ch (14cm)	4ch (4-6-4cm)	8ch (2cm)
oracle D+S	23.5	22.8	22.5	22.4
oracle TAM	23.5	21.7	21.3	21.3
raw, no tdoa	23.5	21.8	21.3	21.1

TABLE III: WER with oracle knowledge of the true target TDOA. All models use 128 filters.

Table III compares the WER of D+S, TAM, and raw waveform models when we do not shift the signals by the TDOA. First, notice that as we increase the number of channels, D+S continues to improve, since finer spatial sampling reduces the sidelobes of the spatial response, leading to increased suppression of noise and reverberation energy arriving from other directions. Second, notice that TAM always has better performance than D+S, as TAM is more general than D+S because it allows individual channels to be filtered before being combined. But notice that the raw waveform CLDNN, without any explicit time alignment or localization (TDOA estimation), performs as well as TAM with the time alignment. This shows us that the trained un-aligned network is implicitly robust to varying TDOA.

F. Summary

Model	Filters	WER - CE	WER - Seq
raw, 1ch	128	23.5	19.3
D+S, 8ch, oracle	128	22.4	18.8
MVDR, 8ch, oracle	128	22.5	18.7
raw, unfactored, 2ch	128	21.8	18.2
raw, unfactored, 4ch	256	20.8	17.2
raw, unfactored, 8ch	256	20.6	17.2

TABLE IV: Raw waveform model WER after sequence training.

To conclude this section, we show the results after sequence training in Table IV. We compare to results for 8 channel oracle D+S, where the true target speech TDOA is known, and to oracle MVDR [6] where the true noise covariance is known in addition to the target TDOA. Oracle MVDR has been shown to perform well, especially on simulated spatialized data [27], [28] similar to what we use here. Note that because we are using the oracle noise covariance, our oracle MVDR does not

suffer from target cancelation due to reverberation, a common failure mode for non-oracle MVDR. Table IV shows that the raw unfactored model, even using only 2 channel inputs and no oracle information, outperforms the single channel and oracle signal processing methods. Using 4 channel inputs, the raw-waveform unfactored model achieves between an 8-10% relative improvement over single channel, D+S and MVDR.

IV. Factoring Spatial and Spectral Selectivity

A. Architecture

In multichannel speech recognition systems, multichannel spatial filtering is often performed separately from single channel feature extraction. However, in the unfactored raw-waveform model, spatial and spectral filtering are done in one layer of the network. In this section, we factor out spatial and spectral filtering into separate layers [26], as shown in Figure 7.

The motivation for this architecture is to design the first layer to be spatially selective, while implementing a frequency decomposition shared across all spatial filters in the second layer. Thus the combined output of the second layer will be the Cartesian product of all spatial and spectral filters.

The first layer, denoted by `tConv1` in the figure, again models Equation 2 and performs a multichannel time-convolution with a FIR spatial filterbank. The operation of each filter $p \in \{0, \ldots, P-1\}$, which we will refer to as a spatial look direction in the factored model, can again be interpreted as a filter-and-sum beamformer, except that any overall time shift is implicit in the filter coefficients rather than being explicitly represented as in Equation 1. The main differences between the unfactored and factored approaches are as follows. First, both the filter size N and number of filters P are much smaller in order to encourage the network to learn filters with a broadband response in frequency that span a small number of spatial look directions needed to cover all possible target speaker locations. The shorter filters in this layer will have worse frequency resolution than those in the unfactored model, but that will be dealt with in the next layer. We hope that this poor frequency resolution will encourage the network to use this first layer to focus on spatial filtering, with a limited spectral response. To make the combination of the first two layers of the factored model conceptually similar to the first layer of the unfactored model (i.e., a bank of bandpassed beamformers), the multi-channel (first) filter layer is not followed by any non-linear compression (i.e. ReLU, log), and we do not perform any pooling between the first and second layers.

The second time-convolution layer, denoted by `tConv2` in the figure, consists of longer-duration single-channel filters. It therefore can learn a decomposition with better frequency resolution than the first layer but is incapable of doing any spatial filtering. Given the P feature maps from the first layer, we perform a time convolution on each of these signals, very similar to the single-channel time-convolution layer described in [15], except that the time convolution is shared across all P feature maps or "look directions". We denote this layer's filters as $g \in \Re^{L \times F \times 1}$, where 1 indicates sharing across the P input feature maps. The "valid" convolution produces an output $w[t] \in \Re^{(M-L+1) \times F \times P}$. The output of the spectral convolution layer, for each look direction p and each filter f, is given by Equation 3.

$$w_f^p[t] = y^p[t] * g_f \tag{3}$$

Next, we pool the filterbank output in time thereby discarding short-time (i.e. phase) information, over the entire time length of the output signal, to produce an output of dimension $1 \times F \times P$. Finally, we apply a rectified non-linearity, followed by a stabilized logarithm compression, to produce a frame-level feature vector at frame l, i.e., $z_l \in \Re^{1 \times F \times P}$, which is then passed to a CLDNN model. We then shift the window of the raw waveform by a small (10ms) hop and repeat this time convolution to produce a set of time-frequency-direction frames at 10ms intervals.

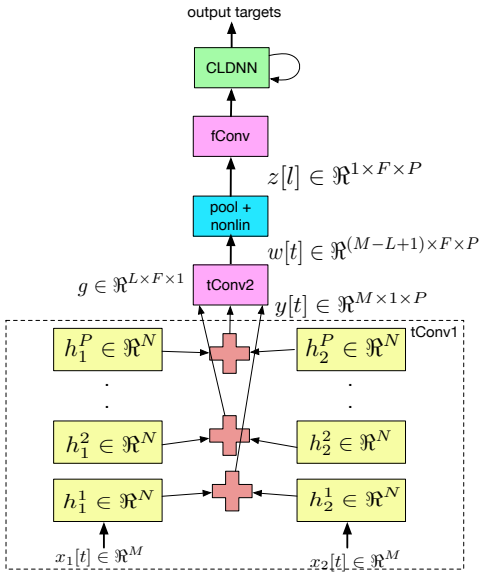

Fig. 7: Factored multichannel raw waveform CLDNN architecture for P look directions. The figure shows two channels for simplicity.

The output of the time convolutional layer (`tConv2`) produces a frame-level feature, denoted as $z[l] \in \Re^{1 \times F \times P}$, which is then passed to a CLDNN acoustic model.

B. Number of Spatial Filters

We first explore the behavior of the proposed factored multichannel architecture as the number of spatial filters P varies. Table V shows that we get good improvements up to 10 spatial filters. We did not explore above 10 filters due to the computational complexities of passing 10 feature maps to the `tConv2` layer. Furthermore, we later found that after sequence training, there was no difference in performance between 5 and 10 spatial filters [29].

The factored network, with 10 spatial filters, achieves a WER of 20.4%, a 6% relative improvement over the 2 channel unfactored multichannel raw-waveform CLDNN. It is important to note that since the `tConv2` layer is shared across all look directions P, the total number of parameters is actually less than the unfactored model.

# Spatial Filters P	WER
baseline 2 ch, raw [16]	21.8
1	23.6
3	21.6
5	20.7
10	20.4

TABLE V: WER when varying the size of the spatial filters in `tConv1`. All models use 128 filters for `tConv2` and results are presented for 2 channels.

C. Filter Analysis

To better understand what the `tConv1` layer learns, Figure 8 plots two-channel filter coefficients and the corresponding spatial responses, or beampatterns, after training.

Despite the intuition described in Section IV, the first layer filters appear to perform both spatial and spectral filtering. However, the beampatterns can nevertheless be categorized into a few broad classes. For example, filters 2, 3, 5, 7, and 9 in Figure 8 only pass through some low frequency subbands below about 1.5 kHz, where most vowel energy occurs, but steered to have nulls in different directions. Very little spatial filtering is done in high-frequency regions, where many fricatives and stops occur. The low frequencies are most useful for localization because they are not subject to spatial aliasing and because they contain much of the energy in the speech signal; perhaps that is why the network exhibits this structure.

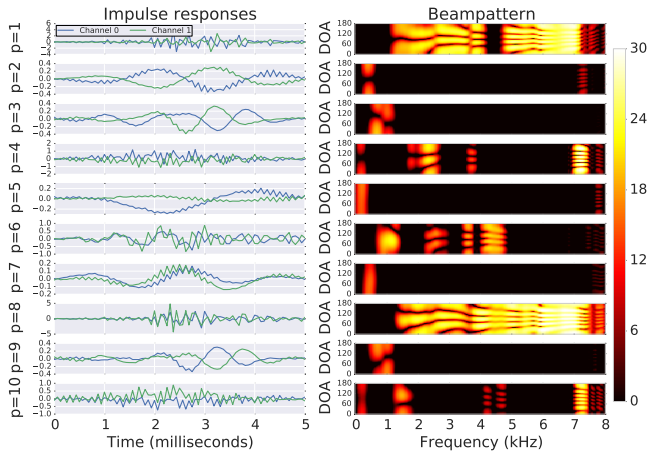

Fig. 8: Trained filters and spatial responses for 10 spatial directions.

To further understand the benefit of the spatial and spectral filtering in `tConv1`, we enforce this layer to only perform spatial filtering by initializing the filters to be an impulse centered at a delay of zero for channel 0, and offset from zero in channel 1 by different delays for each filter. By not training this layer, this amounts to performing delay-and-sum filtering across a set of fixed look directions. Table VI compares performance when fixing vs. training the `tConv1` layer. The results demonstrate that learning the filter parameters, and therefore performing some spectral decomposition, improves performance over keeping this layer fixed.

# Spatial Filters P	`tConv1` Layer	WER
5	fixed	21.9
5	trained	20.7

TABLE VI: WER for training vs. fixing the `tConv1` layer, 2 channel.

D. Results Summary

To conclude this section, we show the results after sequence training, comparing the factored and unfactored models. Notice that the 2 channel factored model provides 6% relative improvement over the unfactored model, while the 4 channel model provides 5% relative improvment. We do not go above 4 channels, as results from Table IV in Section III-F show that there is no difference between 4 and 8 channels.

Method	WER - CE	WER - Seq
raw, unfactored, 2ch	21.8	18.2
raw, factored, 2ch	20.4	17.2
raw, unfactored, 4ch	20.8	17.2
raw, factored, 4ch	19.6	16.3

TABLE VII: Factored Model WER after sequence training, simulated

V. Adaptive Beamforming

While the unfactored model improves over the factored model, the model also suffers from a few drawbacks. First, the learned filters in this model are fixed during decoding, which potentially limits the ability of these models to adapt to previously unseen or changing conditions. In addition, since the factored model must perform spectral filtering for every look direction, this comes with a large computational complexity.

A. NAB Model

To address the limited adaptability and reduce the computational complexity of the models from [16], [26], we propose a neural network adaptive beamforming (NAB) model [30] which re-estimates a set of spatial filter coefficients at each input frame using a neural network.

The NAB model is depicted in Figure 9. At each time frame l, it takes in a small window of M waveform samples for each channel c from the C channel inputs, denoted as $x_0(l)[t], x_1(l)[t], \cdots, x_{C-1}(l)[t]$ for $t \in \{0, \cdots, M-1\}$. Additional to previous notations, the frame index l is explicitly used in this section to emphasize the frame dependent filtering coefficients. For simplicity, the figure shows an NAB model with $C = 2$ channels. We will describe the different modules of this model in subsequent subsections.

Note that our NAB model is similar to the model proposed in [31], although filtering was performed in the frequency domain, as opposed to our model which processes time domain signals. We will show in Section VI-D that performing NAB in the time domain requires estimation of many fewer filter coefficients, and results in better WER compared to frequency domain filter prediction.

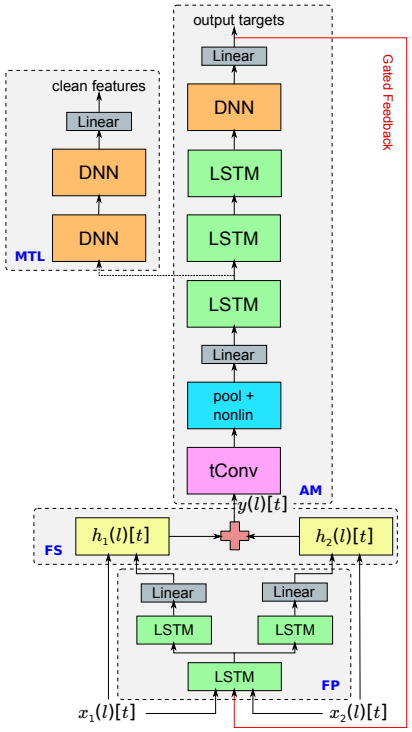

Fig. 9: *Neural network adaptive beamforming (NAB) model architecture. It consists of filter prediction (FP), filter-and-sum (FS) beamforming, acoustic modeling (AM) and multi-task learning (MTL). The figure shows only two channels just for simplicity.*

1) Adaptive Filters: The adaptive filtering layer is given by Equation 4, where $h_c(l)[n]$ is the estimated filter for channel c at time frame l. This model is very similar to the FS model from Equation 1, except now the steering delay τ_c is implicitly absorbed into the estimated filter parameters.

$$y(l)[t] = \sum_{c=0}^{C-1} \sum_{n=0}^{N-1} h_c(l)[n] x_c(l)[t-n] \quad (4)$$

In order to estimate $h_c(l)[t]$, we train a filter prediction (FP) module with one shared LSTM layer, one layer of channel-dependent LSTMs and linear output projection layers to predict N filter coefficients for each channel. The input to the *FP module* is a concatenation of frames of raw input samples $x_c(l)[t]$ from all the channels, and can also include features typically computed for localization such as cross correlation features [32], [31], [33]. The estimation of *FP module* parameters are jointly done with AM parameters by directly minimizing a cross-entropy or sequence loss function. Following Equation 4 the estimated filter coefficients $h_c(l)[t]$ are convolved with input samples $x_c(l)[t]$ for each channel. The outputs of the convolution are summed across channels to produce a single channel signal $y(l)[t]$.

After adaptive FS, the single channel enhanced signal $y(l)[t]$ is passed to an *AM module* (Figure 9). We adopt the single channel raw waveform CLDNN model [15] for acoustic modeling, except that we now skip the frequency convolution layer as it has recently been shown in [34] to not help for noisier tasks. During training, the AM and FP (Figure 9) are trained jointly.

2) Gated Feedback: Augmenting the network input at each frame with the prediction from the previous frame has been shown to improve performance [35]. To investigate the benefit of feedback in the NAB model, we pass the AM prediction at frame $l-1$ back to the FP model at time frame l (red line in Figure 9). Since the softmax prediction is very high dimensional, we feed back the low-rank activations preceding the softmax to the *FP module* to limit the increase of model parameters [36].

This feedback connection gives the *FP module* high level information about the phonemic content of the signal to aid in estimating beamforming filter coefficients. This feedback is comprised of model *predictions* which may contain errors, particularly early in training, and therefore might lead to poor model training [35]. A gating mechanism [37] is hence introduced to the connection to modulate the degree of feedback. Unlike conventional LSTM gates, which control each dimension independently, we use a global scalar gate to moderate the feedback. The gate $g^{\texttt{fb}}(l)$ at time frame l, is computed from the input waveform samples $\boldsymbol{x}(l)$, the state of the first FP LSTM layer $\boldsymbol{s}(l-1)$, and the feedback vector $\boldsymbol{v}(l-1)$, as follows:

$$g^{\texttt{fb}}(l) = \sigma(\boldsymbol{w}_x^T \cdot \boldsymbol{x}(l) + \boldsymbol{w}_s^T \cdot \boldsymbol{s}(l-1) + \boldsymbol{w}_v^T \cdot \boldsymbol{v}(l-1)) \quad (5)$$

where $\boldsymbol{w}_x$, $\boldsymbol{w}_s$ and $\boldsymbol{w}_v$ are the corresponding weight vectors and σ is an elementwise non-linearity. We use a logistic function for σ which outputs values in the range $[0, 1]$, where 0 cuts off the feedback connection and 1 directly passes the feedback through. The effective FP input is hence $\left[\boldsymbol{x}(l), \quad g^{\texttt{fb}}(l)\boldsymbol{v}(l-1)\right]$.

3) Regularization with MTL: Multi-task learning has been shown to yield improved robustness [26], [38], [39]. We adopt an *MTL module* similar to [26] during training by configuring the network to have two outputs, one recognition output which predicts CD states and a second denoising output which reconstructs 128 log-mel features derived from the underlying clean signal. The denoising output is only used in training to regularize the model parameters; the associated layers are discarded during inference. In the NAB model the *MTL module* branches off of the first LSTM layer of the *AM module*, as shown in Figure 9. The *MTL module* is composed of two fully connected DNN layers followed by a linear output layer which predicts clean features. During training the gradients back propagated from the two outputs are weighted by α and $1-\alpha$ for the recognition and denoising outputs respectively.

B. NAB Filter Analysis

The best NAB model found in [30] has the following configurations:

1) the *FP module* has one shared 512-cell LSTM layer across channels, one layer of channel-dependent 256-cell LSTMs and one layer of channel-dependent 25-dimensional linear projection layer;
2) the *FP module* takes in the concatenation of raw waveform samples from each channel;

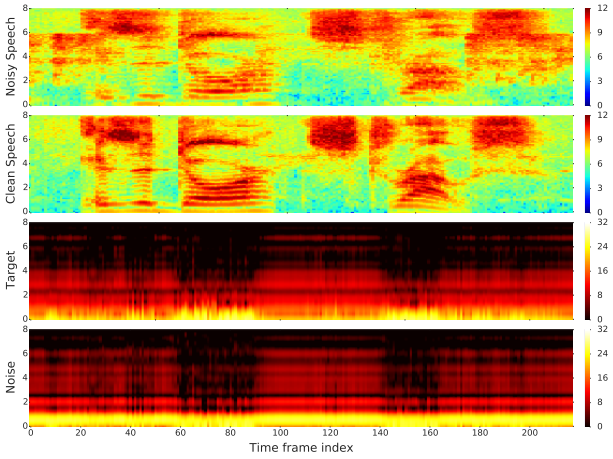

Fig. 10: *Visualizations of the predicted beamformer responses at different frequency (Y-axis) across time (X-axis) at the target speech direction (3rd) and interfering noise direction (4th) with the noisy (1st) and clean (2nd) speech spectrograms.*

3) the *FP module* outputs a 1.5ms filter (25-dimensional vector) for each channel;
4) the *AM module* is a single channel raw waveform LDNN model [15] with 256 `tConv` filters and without the frequency convolution layer [34], which is also similar to other multichannel models discussed in this paper;
5) 128-dimensional clean log-mel features are used as the secondary reconstruction objectives with a weight of 0.1 for MTL;
6) per-frame gated feedback connection from the bottleneck layer right before the *AM module*'s softmax layer is appended to the *FP module*'s input.

Figure 10 illustrates the frequency responses of the predicted beamforming filters at the target speech and interfering noise directions. The SNR for this utterance is 12dB. The responses in the target speech direction have relatively more speech-dependent variations than those in the noise direction. This may indicate that the predicted filters are attending to the speech signal. Besides, the responses at high speech-energy regions are generally lower than others, which suggests the automatic gain control effect of the predicted filters.

C. Result Summary

Finally, to conclude this section, we show the results after sequence training compared to the factored model. To understand the impact of WER with respect to each factor discussed in Section V, we refer the reader to [30]. Instead, in this paper we show the results for the best NAB setup, which uses the factors outlined in Section V-B. Since the NAB model is trained without frequency convolution (i.e., LDNN), we do the same for the factored model. In addition, we show results for the factored model with MTL to be comparable to the NAB model. Table VIII shows that the factored model can potentially handle different directions by enumerating many look directions in the spatial filtering layer and can achieve slightly better performance compared to the adaptive model. However, the adaptive model has less computational complexity, as measured by the number of multiplies and additions (MultAdd) of the model, as shown in the Table.

Model	WER (%) CE	WER (%) Seq	Param	MultAdd
factored (w/MTL)	20.1	16.9	18.9M	35.1M
NAB (w/MTL)	20.5	17.2	24.0M	28.8M

TABLE VIII: Comparison between 2-channel factored and adaptive models, in terms of WER at cross-entropy (CE) and sequence (Seq) training, as well as the total number of multiplies and adds (MultAdd) in millions.

VI. Filtering in the Frequency Domain

Until now, we have presented three multichannel models in the time domain. However, it is well know that "circular" convolution between two time domain signals is equivalent to the element-wise product of their frequency domain counterparts [40], [41]. A benefit of operating in the complex FFT space is that element-wise products are much faster to compute compared to convolutions, particularly when the convolution filters and input size is large as in our multichannel raw waveform models. In this section, we describe how we can implement both the factored Model from Section IV and the NAB Model from Section V, in the frequency domain [29].

A. Factored Model

In this section, we describe the factored model in the frequency domain.

1) Spatial Filtering: For frame index l and channel c, we denote $X_c[l] \in \mathbb{C}^K$ as the result of an M-point Fast Fourier Transform (FFT) of $x_c[t]$ and $H_c^p \in \mathbb{C}^K$ as the FFT of h_c^p. Note that we ignore negative frequencies because the time domain inputs are real, and thus our frequency domain representation of an M-point FFT contains only $K = M/2+1$ unique complex-valued frequency bands. The spatial convolution layer in Equation 2 can be represented by Equation 6 in the frequency domain, where $\cdot$ denotes element-wise product. We denote the output of this layer as $Y^p[l] \in \mathbb{C}^K$ for each look direction p:

$$Y^p[l] = \sum_{c=0}^{C-1} X_c[l] \cdot H_c^p \tag{6}$$

In this paper, we explore two different methods for implementing the "spectral filtering" layer in the frequency domain.

2) Spectral Filtering: Complex Linear Projection: It is straightforward to rewrite the convolution in Equation 3 as an element-wise product in frequency, for each filter f and look direction p:

$$W_f^p[l] = Y^p[l] \cdot G_f \tag{7}$$

In the above equation, $W_f^p[l] \in \mathbb{C}^K$ and $G_f \in \mathbb{C}^K$ is the FFT of the time domain filter g_f in Equation 3. There is no frequency domain equivalent to the max-pooling operation in the time domain. Therefore to mimic max-pooling exactly requires taking the inverse FFT of $W_f^p[l]$ and performing the

pooling operation in the time domain, which is computationally expensive to do for each look direction p and filter output f.

As an alternative [42] recently proposed the Complex Linear Projection (CLP) model which performs average pooling in the frequency domain and results in similar performance to a single channel raw waveform model. Similar to the waveform model the pooling operation is followed by a pointwise absolute-value non-linearity and log compression. The 1-dimensional output for look direction p and filter f is given by:

$$Z_f^p[l] = \log \left| \sum_{k=1}^{N} W_f^p[l,k] \right| \quad (8)$$

Collecting the output for all look directions and filters, the feature at frame index l is denoted as $Z[l] \in \{Z_1^1[l], \dots Z_F^P[l]\} \in \Re^{1 \times F \times P}$, with the same dimensions as the corresponding feature frame in the waveform model, $z[l]$.

3) Spectral Filtering: Linear Projection of Energy: We also explore an alternative decomposition that is motivated by the log-mel filterbank. Given the complex-valued FFT for each look direction, $Y^p[l]$, we first compute the energy at each time-frequency bin (l,k):

$$\hat{Y}^p[l,k] = |Y^p[l,k]|^2 \quad (9)$$

After applying a power compression with $\alpha = 0.1$, $\hat{Y}^p[l]$ is linearly projected down to an F dimensional space, in a process similar to the mel filterbank, albeit with learned filter shapes:

$$Z_f^p[l] = G_f \times (\hat{Y}^p[l])^\alpha \quad (10)$$

As in the other models, the projection weights $G \in \Re^{K \times F}$, are shared across all look directions.

The main difference between the CLP and LPE models is that the former retains phase information when performing the filterbank decomposition with matrix G. In contrast, LPE operates directly on the energy in each frequency band with the assumption that phase not important for computing features.

B. NAB Model

In the frequency-domain NAB setup, we have an LSTM which predicts complex FFT (CFFT) inputs for both channels. Given a 512-pt FFT input, this amounts to predicting 4×257 frequency points for real and imaginary components for 2 channels, which is much more than the predicted filter size in the time domain (i.e., $1.5ms = 25$ taps). After the complex filters are predicted for each channel, element-wise product is done with the FFT of the input for each channel, mimicking the convolution in Equation 4 in the frequency domain. The output of this is given to a single channel LDNN in the frequency domain, which does spectral decomposition, using either CLP or LPE, and acoustic modeling.

C. Results: Factored Model

1) Performance: First, we explore the performance of the frequency domain factored model. Note this model does not have any frequency convolution layer. We explore this for a similar setting to most efficient raw-waveform factored setup [29], namely $P = 5$ look directions in the spatial layer and $F = 128$ filters in the spectral layer. The input is 32ms instead of 35ms like raw-waveform, as this allows us to take a $D = 512$-point DFT without zero-padding at a sampling rate of 16kHz. A 35-ms input would have required us to take a 1024-point DFT, and we have not found any big difference in performance between 32 and 35ms inputs for raw-waveform.

Table IX shows that the WER performance of both the CLP and LPE factored models are similar. The table also indicates the total number of multiplication and addition operations (MultAdd) for different layers of the model. Both models reduce the number of operations by a factor of 1.9x over the best waveform model, with a small degradation in WER.

Model	Spatial MultAdd	Spectral MultAdd	Total MultAdd	WER CE	WER Seq
time	525.6K	15.71M	35.1M	20.4	17.1
CLP	10.3K	655.4K	19.6M	20.5	17.2
LPE	10.3K	165.1K	19.1M	20.7	17.2

TABLE IX: Frequency Domain Factored Model Performance, in terms of WER at cross-entropy (CE) and sequence (Seq) training, as well as the total number of multiplies and adds (MultAdd) in millions.

However, given that the frequency models are more computationally efficient, we explore improving WER by increasing the window size (and therefore computational complexity) of the factored models. Specifically, since longer windows typically help with localization [6], we explore using 64ms input windows for both models. With a 64ms input, the frequency models require a 1024-point FFT. Table X shows that the frequency models improve the WER over using a smaller 32ms input, and still perform roughly the same. However, the frequency model now has an even larger computational complexity savings of 2.7x savings compared to the time domain model.

Feat	Spatial MultAdd	Spectral MultAdd	Total MultAdd	WER Seq
time	906.1K	33.81M	53.6M	17.1
CLP	20.5K	1.3M	20.2M	17.1
LPE	20.5K	329.0K	19.3M	16.9

TABLE X: Results with a 64ms Window Size, in terms of WER at cross-entropy (CE) and sequence (Seq) training, as well as the total number of multiplies and adds (MultAdd) in millions.

2) Comparison between learning in time vs. frequency: Figure 11 shows the spatial responses (i.e., beampatterns) for both the time and frequency domain spatial layers. Since the LPE and CLP models have the same spatial layer and we have found the beampatterns to look similar, we only plot the CLP model for simplicity. The beampatterns show the magnitude response in dB as a function of frequency and direction of arrival, i.e. each horizontal slice of the beampattern corresponds to the filter's magnitude response for a signal coming from a particular direction. In each frequency band (vertical slice),

lighter shades indicate that sounds from those directions are passed through, while darker shades indicate directions whose energy is attenuated. The figures show that the spatial filters learned in the time domain are band-limited, unlike those learned in the frequency domain. Furthermore, the peaks and nulls are aligned well across frequencies for the time domain filters. One hypothesis we have for the band-limited nature of the raw waveform spatial filters is that they are short (i.e., 5ms), compared to the frequency-domain spatial filters which span over the entire 32 or 64-ms input.

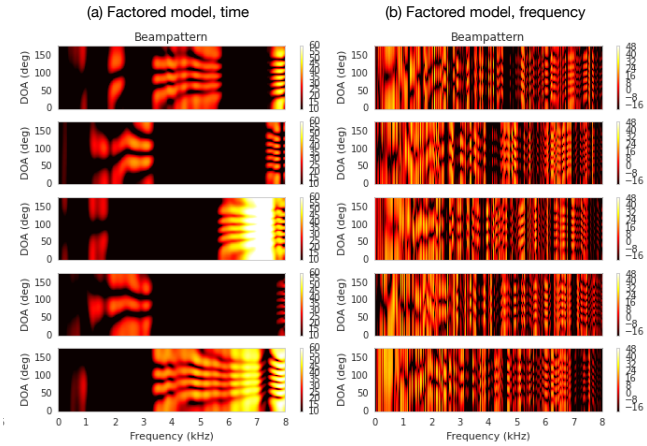

Fig. 11: Beampatterns of Time and Frequency Models

The differences between these models can further be seen in the magnitude responses of the spectral layer filters, as well as in the outputs of the spectral layers from different look directions plotted for an example signal. Figure 12 illustrates that the magnitude responses in both time and CLP models look qualitatively similar, and learn bandpass filters with increasing center frequency. However, because the spatial layers in time and frequency are quite different, we see that the spectral layer outputs in time are much more diverse in different spatial directions compared to the CLP model. In contrast to these models, the LPE spectral layer does not seem to learn bandpass filters.

At some level, time-domain and frequency-domain representations are interchangeable, but they result in networks that are parameterized very differently. Even though the time and frequency models all learn different spatial filters, they all seem to have similar WERs. In addition, even though the spatial layer of the CLP and LPE models are different, they too seem to have similar performance. There are roughly 18M parameters in the LDNN model that sits above the spatial/spectral layers, which accounts for over 90% of the parameters in the model. Any differences between the spatial layers in time and frequency are likely accounted for in the LDNN part of the network.

D. Results: Adaptive Model

Next, we explore the performance of the frequency domain NAB model. Table XI shows the WER at CE and computational complexity of the raw-waveform and CLP NAB

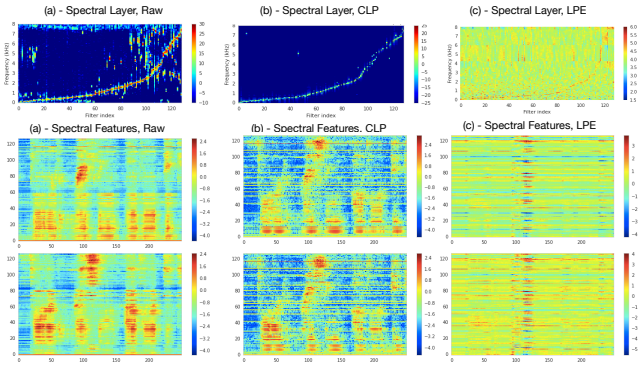

Fig. 12: The first row shows the magnitude response of the spectral layer for the (a) raw-waveform, (b) CLP and (c) LPE models. The second and third rows show the output of the spectral layers of these models for two different look directions.

models. While using CLP features greatly reduces computational complexity, the performance is worse than the raw-waveform model. One hypothesis we have is that frequency domain processing requires predicting a higher dimensional filter, which we can see from the table leads to a degradation in performance. Since the CLP and LPE models perform similarly for the factored model, and the CLP NAB model degraded performance over raw NAB, we did not repeat the experiment for LPE NAB.

Model	WER (%)	Param (M)	MultAdd (M)
raw	20.4	18.9	35.1
CLP	21.0	24.7	25.1

TABLE XI: Comparison between time and frequency NAB models.

VII. Final Comparison, Re-recorded Data

Finally, to show that a model trained on simulated data can handle real reverberation, we also evaluated the performance of different multichannel models presented in this paper on a real "Rerecorded" test set. Reverberation-I is when the microphone is placed on a coffee table, whereas Reverberation-II is when the mic is placed on a TV stand. Since this set contains a circular microphone geometry but our models are trained on a linear microphone geometry, we only report results with 2 channels to form a linear array with a 7.5cm spacing (in contrast to the 14cm spacing during training). In spite of these mismatches in reverberation and in microphone spacing, our technique performs well.

Table XII shows the results with different multichannel models after sequence training. All raw-waveform models are trained with 35-ms inputs and 128 spectral decomposition filters. The factored model has 5 look directions. The LPE factored model is trained with a 64-ms input, 5 look directions, and 128 spectral decomposition filters. All frontends use an LDNN architecture in the upper layers of the network.

Notice that the 2 channel raw factored model gives a 13% relative improvement over single channel, with larger

improvements in noisier test sets, which is to be expected. In addition, the LPE factored model performs similar to the raw factored model. Finally, the NAB model performs much worse than the factored model. We have not investigated this performance difference in detail, but one hypothesis is that because the NAB model adapts its filters over the course of an utterance, it may sometimes adapt its filters into a bad configuration from which it cannot recover when presented with mismatched test data.

We also include results on the noisier reverberation sets with oracle localization and a robust superdirective beamformer similar to [43], [44] with oracle knowledge of the speech and noise. Note that our superdirective beamformer is a fixed filter and sum beamformer similar in spirit to a maximum signal-to-noise ratio filter [2]. The table shows that for the noisier reverberation sets, our factored models, with just 2 channels and no oracle information, are able to match the performance of the oracle superdirective beamformer with 7 channels.

Model	Rev.-I	Rev.-II	Rev.-I Noisy	Rev.-II Noisy	Ave
1 channel raw	18.6	18.5	27.8	26.7	22.9
2 channel raw, unfactored	17.9	17.6	25.9	24.7	21.5
2 channel raw, factored	17.1	16.9	24.6	24.2	20.7
2 channel LPE, factored	17.4	16.8	25.2	23.5	20.7
2 channel raw, NAB	17.8	18.1	27.1	26.1	22.3
7 channel, superdirective	-	-	25.3	23.7	-

TABLE XII: WER on "Rerecorded" set after sequence training.

It is important to note that the motivation of the above results is to show the robustness of the neural network methods when there is slight mismatch in microphone array geometry and noise conditions between training and test, as this condition matches the real-world scenarios we currently deal with. For very severe microphone array geometry mismatch, we have noticed a larger degradation in performance, as shown in Section 4.4 of our paper [16]. In order to address severe mismatch, we have found that we must train the system with a variety of microphone spacings to make the model more robust.

VIII. Conclusions and Future Work

In this paper, we introduced a methodology to do multichannel enhancement and acoustic modeling jointly within a neural network framework. First, we developed a unfactored raw-waveform multichannel model, and showed that this model performed as well as a model given oracle knowledge of the true location. Next, we introduced a factored multichannel model to separate out spatial and spectral filtering operations, and found that this offered an improvement over the unfactored model. Next, we introduced an adaptive beamforming method, which we found to match the performance of the multichannel model with far fewer computations. Finally, we showed that we can match the performance of the raw-waveform factored model, with far fewer computations, with a frequency-domain factored model. Overall, the factored model provides between a 5-13% relative improvement over single channel and traditional signal processing techniques, on both simulated and rerecorded test sets.

Acknowledgements

Thank you to Yedid Hoshen and Arden Huang for discussions related to multichannel processing.

References

[1] M. Brandstein and D. Ward, *Microphone Arrays: Signal Processing Techniques and Applications*. Springer, 2001.

[2] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*. Springer, 2009.

[3] M. Delcroix, T. Yoshioka, A. Ogawa, Y. Kubo, M. Fujimoto, N. Ito, K. Kinoshita, M. Espi, T. Hori, T. Nakatani, and A. Nakamura, "Linear Prediction-based Dereverberation with Advanced Speech Enhancement and Recognition Technologies for the REVERB Challenge," in *REVERB Workshop*, 2014.

[4] T. Hain, L. Burget, J. Dines, P. Garner, F. Grezl, A. Hannani, M. Huijbregts, M. Karafiat, M. Lincoln, and V. Wan, "Transcribing Meetings with the AMIDA Systems," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 486–498, 2012.

[5] A. Stolcke, X. Anguera, K. Boakye, O. Çetin, A. Janin, M. Magimai-Doss, C. Wooters, and J. Zheng, "The SRI-ICSI Spring 2007 Meeting and Lecture Recognition System," *Multimodal Technologies for Perception of Humans*, vol. Lecture Notes in Computer Science, no. 2, pp. 450–463, 2008.

[6] B. D. Veen and K. M. Buckley, "Beamforming: A Versatile Approach to Spatial Filtering," *IEEE ASSP Magazine*, vol. 5, no. 2, pp. 4–24, 1988.

[7] M. Seltzer, B. Raj, and R. M. Stern, "Likelihood-maximizing Beamforming for Robust Handsfree Speech Recognition," *IEEE Trascations on Audio, Speech and Language Processing*, vol. 12, no. 5, pp. 489–498, 2004.

[8] J. R. Hershey, J. L. Roux, and F. Weninger, "Deep Unfolding: Model-Based Inspiration of Novel Deep Architectures," *CoRR*, vol. abs/1409.2574, 2014.

[9] A. Mohamed, G. Hinton, and G. Penn, "Understanding how Deep Belief Networks Perform Acoustic Modelling," in *ICASSP*, 2012.

[10] N. Jaitly and G. Hinton, "Learning a Better Representation of Speech Soundwaves using Restricted Boltzmann Machines," in *Proc. ICASSP*, 2011.

[11] D. Palaz, R. Collobert, and M. Doss, "Estimating Phoneme Class Conditional Probabilities From Raw Speech Signal using Convolutional Neural Networks," in *Proc. Interspeech*, 2014.

[12] Z. Tüske, P. Golik, R. Schlüter, and H. Ney, "Acoustic Modeling with Deep Neural Networks using Raw Time Signal for LVCSR," in *Proc. Interspeech*, 2014.

[13] S. Dieleman and B. Schrauwen, "End-to-end learning for music audio," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 6964–6968.

[14] Y. Hoshen, R. J. Weiss, and K. W. Wilson, "Speech Acoustic Modeling from Raw Multichannel Waveforms," in *Proc. ICASSP*, 2015.

[15] 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.

[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, O. Vinyals, A. Senior, and H. Sak, "Convolutional, Long Short-Term Memory, Fully Connected Deep Neural Networks," in *Proc. ICASSP*, 2015.

[18] L. J. Griffiths and C. W. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Transactions on Antennas and Propagation*, vol. 30, no. 1, pp. 27–34, 1982.

[19] J. B. Allen and D. A. Berkley, "Image Method for Efficiently Simulation Room-Small Acoustics," *Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943 – 950, April 1979.

[20] T. N. Sainath, B. Kingsbury, A. Mohamed, G. Dahl, G. Saon, H. Soltau, T. Beran, A. Aravkin, and B. Ramabhadran, "Improvements to Deep Convolutional Neural Networks for LVCSR," in *Proc. ASRU*, 2013.

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

[22] T. N. Sainath, B. Kingsbury, V. Sindhwani, E. Arisoy, and B. Ramabhadran, ""Low-Rank Matrix Factorization for Deep Neural Network Training with High-Dimensional Output Targets," in *Proc. ICASSP*, 2013.

[23] 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.

[24] G. Heigold, E. McDermott, V. Vanhoucke, A. Senior, and M. Bacchiani, "Asynchronous Stochastic Optimization for Sequence Training of Deep Neural Networks," in *Proc. ICASSP*, 2014.

[25] X. Glorot and Y. Bengio, "Understanding the Difficulty of Training Deep Feedforward Neural Networks," in *Proc. AISTATS*, 2010.

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

[27] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'chime' speech separation and recognition challenge: Dataset, task and baselines," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 504–511.

[28] D. Bagchi, M. I. Mandel, Z. Wang, Y. He, A. Plummer, and E. Fosler-Lussier, "Combining spectral feature mapping and multi-channel model-based source separation for noise-robust automatic speech recognition," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 496–503.

[29] T. N. Sainath, A. Narayanan, R. J. Weiss, K. W. Wilson, M. Bacchiani, and I. Shafran, "Improvements to Factorized Neural Network Multichannel Models," in *in Proc. Interspeech*, 2016.

[30] B. Li, T. N. Sainath, R. J. Weiss, K. W. Wilson, and M. Bacchiani, "Neural Network Adaptive Beamforming for Robust Multichannel Speech Recognition," in *in Proc. Interspeech*, 2016.

[31] X. Xiao, S. Watanabe, H. Erdogan, L. Lu, J. Hershey, M. L. Seltzer, G. Chen, Y. Zhang, M. Mandel, and D. Yu, "Deep beamforming networks for multi-channel speech recognition," in *Proc. ICASSP*, 2016.

[32] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 24, no. 4, pp. 320–327, 1976.

[33] X. Xiao, S. Zhao, X. Zhong, D. L. Jones, E. S. Chng, and H. Li, "A learning-based approach to direction of arrival estimation in noisy and reverberant environments," in *Proc. ICASSP*, 2015, pp. 2814–2818.

[34] T. N. Sainath and B. Li, "Modeling Time-Frequency Patterns with LSTM vs. Convolutional Architectures for LVCSR Tasks," in *in Proc. Interspeech*, 2016.

[35] S. Bengio, O. Vinyals, N. Jaitly, and N. Shazeer, "Scheduled sampling for sequence prediction with recurrent neural networks," in *Advances in Neural Information Processing Systems*, 2015, pp. 1171–1179.

[36] Y. Zhang, E. Chuangsuwanich, and J. R. Glass, "Extracting deep neural network bottleneck features using low-rank matrix factorization." in *ICASSP*, 2014, pp. 185–189.

[37] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Gated feedback recurrent neural networks," *arXiv preprint arXiv:1502.02367*, 2015.

[38] R. Giri, M. L. Seltzer, J. Droppo, and D. Yu, "Improving speech recognition in reverberation using a room-aware deep neural network and multi-task learning," in *Proc. ICASSP*. IEEE, 2015, pp. 5014–5018.

[39] Z. Chen, S. Watanabe, H. Erdoğan, and J. R. Hershey, "Speech enhancement and recognition using multi-task learning of long short-term memory recurrent neural networks," in *Proc. Interspeech*. ISCA, 2015, pp. 3274–3278.

[40] Y. Bengio and Y. Lecun, "Scaling Learning Algorithms Towards AI," *Large Scale Kernel Machines*, 2007.

[41] R. Bracewell, *The Fourier Transform and Its Applications*, 3rd ed. McGraw-Hill, 1999.

[42] E. Variani, T. N. Sainath, I. Shafran, and M. Bacchiani, "Complex Linear Projection (CLP): A Discriminative Approach to Joint Feature Extraction and Acoustic Modeling," in *in Proc. Interspeech*, 2016.

[43] S. Doclo and M. Moonen, "Superdirective beamforming robust against microphone mismatch," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 2, pp. 617–631, 2007.

[44] K. D, K. Yao, and F. Lorenzelli, "Broadband maximum energy array with user imposed spatial and frequency constraints," in *Proc. ICASSP*, 1994.

Tara N. Sainath received her B.S (2004), M. Eng (2005) and Ph.D. (2009) in Electrical Engineering and Computer Science all from MIT. The main focus of her PhD work was in acoustic modeling for noise robust speech recognition. After her PhD, she spent 5 years at the Speech and Language Algorithms group at IBM T.J. Watson Research Center, before joining Google Research. She has co-organized a special session on Sparse Representations at Interspeech 2010 in Japan. She has also organized a special session on Deep Learning at ICML 2013 in Atlanta. In addition, she is a staff reporter for the IEEE Speech and Language Processing Technical Committee (SLTC) Newsletter. Her research interests are mainly in acoustic modeling and deep neural networks.

Ron J. Weiss is a software engineer at Google where he has worked on content-based audio analysis, recommender systems for music, and noise robust speech recognition. Ron completed his Ph.D. in electrical engineering from Columbia University in 2009 where he worked in the Laboratory for the Recognition of Speech and Audio. From 2009 to 2010 he was a postdoctoral researcher in the Music and Audio Research Laboratory at New York University.

Kevin W. Wilson is a software engineer at Google, where he works on audio content analysis, with applications to speech acoustic modeling and audio event detection. He received his B.S (1999), M. Eng (2000), and Ph.D. (2006) in Electrical Engineering and Computer Science, all from MIT. His Ph.D. work was on multi-microphone audio-visual source localization and tracking. From 2006 to 2010, he worked on multimedia content analysis as a Member Technical Staff at the Mitsubishi Electric Research Lab.

Bo Li is a research scientist at Google, where he works on acoustic modeling for robust speech recognition. He received his Ph.D. in School of Computing from National University of Singapore in 2014. His Ph.D. work was on noise robust speech recognition with deep neural networks. Bo received his B. Eng (2008) in School of Computer from Northwestern Polytechnical University in China. His research interests are mainly in robust acoustic modeling and deep neural networks.

Arun Narayanan received his M.S. and the Ph.D. degrees in computer science from the Ohio State University, Columbus, USA, in 2012 and 2014, respectively. Since 2014, he has been a Research Scientist at Google, Inc. His research interests include robust automatic speech recognition, speech separation, and machine learning.

Ehsan Variani received his M.S. and the Ph.D. degrees in Electrical and Computer Engineering from the Johns Hopkins University, Baltimore, Maryland, USA, in 2011 and 2015, respectively. Since 2015, he has been a Research Scientist at Google, Inc. His research interests include machine learning, information theory and automatic speech recognition.

Michiel Bacchiani has worked in various areas of speech recognition research for more than 20 years with an emphasis on acoustic modeling. He currently manages the acoustic modeling team at Google responsible for developing the technology backing all Google speech applications. At Google, he previously lead the efforts around voicemail transcription and YouTube automatic captioning. Before joining Google, Michiel Bacchiani worked as a member of technical staff at IBM Research. Before that he worked at AT&T Research Labs and ATR International in Kyoto Japan. At all these assignments he focused on various aspects of speech recognition algorithm research. Michiel Bacchiani received the "ingenieur" (ir.) degree from the Technical University of Eindhoven, The Netherlands and the Ph.D degree from Boston University. He has authored numerous scientific publications. He is elected to be the chair of the IEEE Speech and Language Processing Technical Committee. He is a board member and subject editor of Speech Communication. He has served on various conference and workshop technical committees and served as area chair for major international conferences (ICASSP, Interspeech).

Izhak Shafran is a speech and machine learning researcher, who has been working on acoustic modeling for speech recognition. Before joining Google in 2014, he was an Associate Professor and a member of the Center for Spoken Language Processing at OHSU, where he also focused on medical application of spoken language technology. He graduated from University of Washington in Seattle in 2001 and subsequently worked at AT&T Research Labs at Florham Park with the speech algorithms group. In summer of 2006, he was a visiting professor at University of Paris-South, working at LIMSI. Subsequently, he was a research faculty at the Center for Language and Speech Processing (CLSP) in Johns Hopkins University. He received an NIH Career Development Award in 2010.

Andrew Senior received his PhD from Cambridge University for his thesis "Recurrent Neural Networks for Offline Cursive Handwriting Recognition". He is currently a research scientist in deep learning at Google DeepMind in London. Previously he worked on research into deep and recurrent neural networks for acoustic modelling in Googles speech recognition system. Before joining Google, he worked at IBM Research in the areas of handwriting, audio-visual speech, face and fingerprint recognition as well as video privacy protection and visual tracking. He has taught at Columbia University, written over 100 papers and holds 49 patents.

Kean Chin received the B.Sc. degree in computer engineering from University of Warwick, Coventry, U.K., in 1995 and the M.Phil. degree from the University of Cambridge, Cambridge, U.K., in 1999. He started his Ph.D. degree in the Speech, Vision, and Robotics Group, Engineering Department, University of Cambridge in 1999. After the B.Sc. degree, he began work as a Researcher in the Articial Intelligence Laboratory, Standards and Industrial Research Institute of Malaysia (SIRIM). He joined the Speech Technology Group (STG), Cambridge Research Laboratory, Toshiba Research Europe, Ltd., in 2002. He lead the ASR group in STG, Toshiba since 2008. He is currently a senior research scientist at Google Inc.

Ananya Misra received her A.B. in Computer Science and Mathematics from Bryn Mawr College, and her M.A. and Ph.D. in Computer Science from Princeton University. Her dissertation focused on interactive analysis and re-synthesis of real-world sounds for new music composition. She has since worked on automatic speech recognition at Google, with special interest in noise-robust acoustic modeling.

Chanwoo Kim has been a software engineer at Google, Inc. since 2013. He was a speech scientist and software development engineer at Microsoft from 2011 to 2013. Dr. Kim received a Ph.D. from the Language Technologies Institute of the Carnegie Mellon University School of Computer Science in 2010. He received his B.S and M.S. degrees in Electrical Engineering from Seoul National University in 1998 and 2001, respectively. Dr. Kim's doctoral research was focused on enhancing the robustness of automatic speech recognition systems in noisy environments. Toward this end he has developed a number of different algorithms for single-microphone applications, dual-microphone applications, and multiple-microphone applications which have been applied to various real-world applications. Between 2003 and 2005 Dr. Kim was a Senior Research Engineer at LG Electronics, where he worked primarily on embedded signal processing and protocol stacks for multimedia systems. Prior to his employment at LG, he worked for EdumediaTek and SK Teletech as a R&D engineer.