MITSUBISHI ELECTRIC RESEARCH LABORATORIES
http://www.merl.com

Beamforming Networks Using Spatial Covariance Features for Far-field Speech Recognition

Xiao, Xiong; Watanabe, Shinji; Chng, Eng Siong; Li, Haizhou

TR2016-162 December 13, 2016

Abstract

Recently, a deep beamforming (BF) network was proposed to predict BF weights from phase-carrying features, such as generalized cross correlation (GCC). The BF network is trained jointly with the acoustic model to minimize automatic speech recognition (ASR) cost function. In this paper, we propose to replace GCC with features derived from input signals' spatial covariance matrices (SCM), which contain the phase information of individual frequency bands. Experimental results on the AMI meeting transcription task shows that the BF network using SCM features significantly reduces the word error rate to 44.1% from 47.9% obtained with the conventional ASR pipeline using delay-and-sum BF. Also compared with GCC features, we have observed small but steady gain by 0.6% absolutely. The use of SCM features also facilitate the implementation of more advanced BF methods within a deep learning framework, such as minimum variance distortionless response BF that requires the speech and noise SCM.

Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Beamforming Networks Using Spatial Covariance Features for Far-field Speech Recognition

Xiong Xiao*, Shinji Watanabe†, Eng Siong Chng‡ and Haizhou Li‡§¶
* Temasek Laboratories, Nanyang Technological University (NTU), Singapore.
† Mitsubishi Electric Research Laboratories (MERL), USA.
‡ School of Computer Science and Engineering, NTU, Singapore.
§ Human Language Technology Department, Institute of Infocomm Research, Singapore.
¶ Department of Electrical and Computer Engineering, National University of Singapore.
E-mail: xiaoxiong@ntu.edu.sg, watanabe@merl.com, aseschng@ntu.edu.sg, hli@i2r.a-star.edu.sg

***Abstract*—Recently, a deep beamforming (BF) network was proposed to predict BF weights from phase-carrying features, such as generalized cross correlation (GCC). The BF network is trained jointly with the acoustic model to minimize automatic speech recognition (ASR) cost function. In this paper, we propose to replace GCC with features derived from input signals' spatial covariance matrices (SCM), which contain the phase information of individual frequency bands. Experimental results on the AMI meeting transcription task shows that the BF network using SCM features significantly reduces the word error rate to 44.1% from 47.9% obtained with the conventional ASR pipeline using delay-and-sum BF. Also compared with GCC features, we have observed small but steady gain by 0.6% absolutely. The use of SCM features also facilitate the implementation of more advanced BF methods within a deep learning framework, such as minimum variance distortionless response BF that requires the speech and noise SCM.**

I. Introduction

Far-field ASR remains a challenging research topic, even with neural network based acoustic model and a large amount of training data. It attracts considerable attentions from the speech processing community, and several benchmarking tasks have been devoted to this topic, such as the REVERB Challenge [1], CHiME-3 speech source separation challenge [2], and the 2015 Jelinek Summer Workshop on Speech and Language Technology [3]. Although considerable progress has been made in the past few years, the performance of state-of-the-art ASR systems is still poor on far-field recordings, e.g. the WER on the AMI meeting transcription task [4] is still around 50% even when 8-channel beamforming and neural network based acoustic model have been used [5, 6]. Beamforming is an important technique to improve far-field ASR performance as it allows us to utilize spatial information to separate target signal and interferences, in addition to spectral and temporal information.

Traditional beamforming methods [7] are usually optimized with signal level criteria that are not directly related to the ASR performance. For example, the delay-and-sum (DS) beamforming aligns the input channels to cancel their phase difference w.r.t. the target signal and sum the channels. The processing reinforces the target signal more than the interference from other directions and hence improve signal-to-noise ratio (SNR). The MVDR beamforming [8] also takes into consideration the spatial characteristics of the interference and is able to achieve higher SNR improvement. Compared to MVDR beamforming, the multi-channel Wiener filter [9] sacrifices some distortions on the target signal to achieve higher noise attenuation. Both MVDR and multi-channel Wiener filter are maximizing output SNR, which is a meaningful objective function but not directly related to the ASR's evaluation metrics, such as WER.

Recently, several learning-based beamforming methods are proposed for the ASR task [5, 10–14]. These methods usually require a training data set from which prior knowledge of speech signal can be extracted and then used in determining the beamforming parameters. An early study is the LIMABEAM [14] that applies filter-and-sum (FS) beamforming on multi-channel waveforms and estimates the beamforming weights by maximizing the likelihood of the features extracted from the enhanced signal. Gaussian mixture model (GMM) based acoustic model is used to evaluate the likelihood of the enhanced features and guide the beamforming weight estimation. In [10, 11, 15, 16], neural network based acoustic models are trained directly on multi-channel waveforms. The first layer of the networks is temporal convolution layer that takes in multi-channel time domain signals. These temporal convolution filters usually have both frequency and spatial location selectivity. In other words, the network learns a set of filterbanks, each looking at a specific direction-of-arrival (DOA) and frequency band. These filters are trained directly on the ASR's cost function and shown to outperform traditional beamforming methods. In two other studies [12, 17], neural network is used to predict whether a time frequency bin is dominated by target speech or not, and this information is used to estimate spatial covariance matrices of speech and noise which in turn are used to determining beamforming parameters. This method could also be combined with the acoustic model to achieve the joint training of acoustic model and mask estimation.

While the work in [10, 11, 15, 16] uses very few domain knowledge of beamforming and relies mainly on the data-driven principle to perform spatial filtering, the deep beamforming networks proposed in [5] aims more straightforward applications of traditional beamforming. The multi-channel waveforms are first used to generate phase-carrying features derived from generalized cross correlation (GCC) [18]. Then a feedforward DNN maps the GCC features to complex filter-and-sum beamforming weights directly in frequency domain. Spatial filtering is carried out by applying the predicted weights to the input signals in the frequency domain in the same way as traditional beamforming. Features are extracted from the enhanced power spectrum and used for DNN based acoustic model. Thus, the whole pipeline of signal processing from waveforms to acoustic modeling are built into a computational graph, and the beamforming weight predicting DNN and the acoustic model DNN are trained jointly on the ASR cost function. Experimental results on the AMI meeting transcription task shows significant ASR performance improvement was obtained by the beamforming network over the DS beamforming.

In this paper, we improve the deep beamforming network [5] by using features derived from spatial covariance matrices to replace the GCC features. This is motivated by the fact that many beamforming methods, such as MVDR beamforming, use spatial covariance matrix to determine beamforming weights. GCC is time domain features and captures the overall time difference of arrival (TDOA) between microphones. Spatial covariance features, on the other hand, captures the phase differences at individual frequency bins. To make the spatial covariance features suitable for neural network inputs, we propose to normalize the covariance matrices by the average power of received signal at each time frequency bin. We also propose several ways of reducing the number of feature dimensions. The main contribution of this paper is to enable the use of spatial covariance features in a beamforming network, which can also achieve better ASR performance than using GCC features on the AMI meeting transcription task.

The rest of this paper is organized as follows. In section II, we briefly review the beamforming network approach and the extraction of GCC features. In section III, we describe the extraction of spatial covariance matrix features and provide theoretical connection to the GCC features. In section IV, we present the experimental settings and results on the AMI task. Finally, we conclude the study in section V and discuss about future research directions.

II. Review of Deep Beamforming Networks Using GCC Features

In this section, we briefly review the beamforming network proposed in [5]. The system diagram is shown in Fig. II. From the multi-channel input speech signals, we generate both the GCC features and short-time Fourier transform coefficients. The GCC features are used to predict the complex-valued beamforming weights in the frequency domain (filter-and-sum) by using a feedforward DNN. The beamforming weights are then used to filter the input Fourier coefficients in the beamforming module to produce the enhanced single-channel Fourier coefficients, which are used to generate log Mel filterbanks for acoustic modeling. The main advantage of the beamforming network over traditional beamforming techniques is that both the beamforming and acoustic modeling networks can be optimized together using the ASR cost function. Therefore, the beamforming network has the potential to perform better than traditional methods for the ASR task.

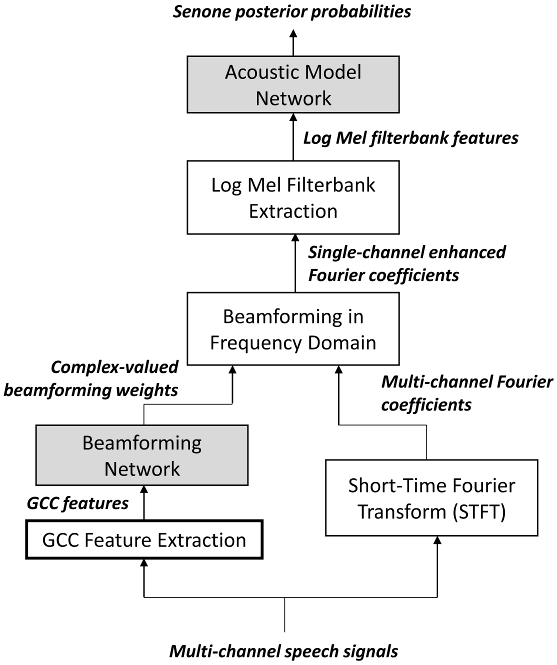
Senone posterior probabilities
Acoustic Model Network
Log Mel filterbank features
Log Mel Filterbank Extraction
Single-channel enhanced Fourier coefficients
Beamforming in Frequency Domain
Complex-valued beamforming weights
Multi-channel Fourier coefficients
Beamforming Network
GCC features
GCC Feature Extraction
Short-Time Fourier Transform (STFT)
Multi-channel speech signals

Fig. 1. System diagram of the joint training of beamforming network and acoustic model network. The two shaded boxes denote neural networks that are trained together, while other blocks denote deterministic processing.

A. Extraction of GCC features

Similar to traditional methods, the beamforming network needs time delay information between microphones to predict the beamforming weights. In [5], GCC is adopted as features for the beamforming network. We briefly review the extraction of GCC features in this section.

For signals recorded by two microphone channels $y_i[n]$ and $y_j[n]$, the cross correlation in the frequency domain can be computed using the GCC-PHAT method [18] by

$$G_{i,j}(f) = \frac{Y_i(f)Y_j^*(f)}{|Y_i(f)Y_j^*(f)|} \tag{1}$$

where $Y_i(f)$ and $Y_j(f)$ are the Fourier transform of $y_i[n]$ and $y_j[n]$ at frequency bin f, respectively. $Y_j^*(f)$ is the complex conjugate of $Y_j(f)$. The cross correlation in time domain can be obtained by

$$R_{i,j}(\tau) = \text{IFT}(G_{i,j}(f)) \tag{2}$$

where IFT() denotes the inverse Fourier transform. In classic methods, we can estimate the TDOA between the microphones i and j by finding the peak of the cross correlation function and use it to determining the beamforming weights. For beamforming network, we can directly use the correlation

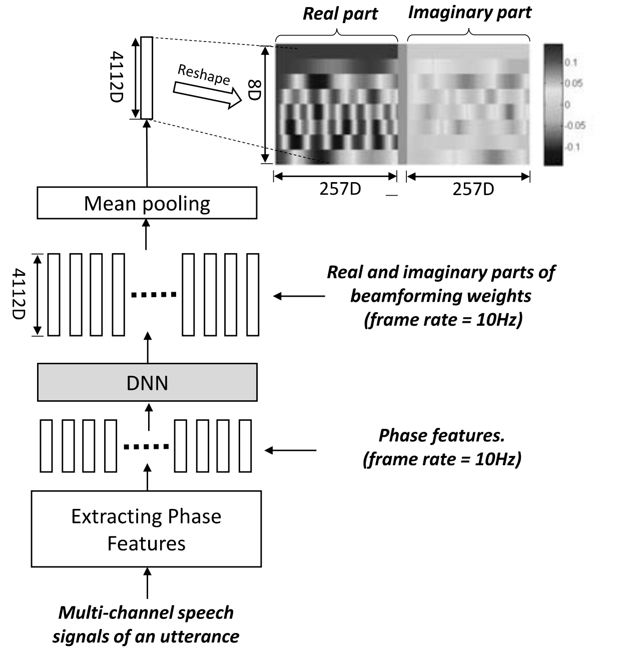
Real part
Imaginary part
4112D
Reshape
8D
257D
257D
0.1
0.05
0
-0.05
-0.1
Mean pooling
4112D
Real and imaginary parts of beamforming weights (frame rate = 10Hz)
DNN
Phase features. (frame rate = 10Hz)
Extracting Phase Features
Multi-channel speech signals of an utterance

Fig. 2. Predicting beamforming weights in the frequency domain by using a DNN. An example weight map predicted by the network is shown.

function as the input features to avoid the estimation of TDOA which may be erroneous.

In practice, only the central part of the correlation function corresponding to the $\pm\tau_{max}$ samples are used as features, where τ_{max} is the maximum possible delay between microphones in the array in terms of samples. For the array used in this paper, $\tau_{max} = 10$, hence, a 21 dimensional feature vector is extracted from each microphone pair. To improve the robustness of the features, we can include the feature vectors of all microphone pairs. For example, if we have 8 microphones, we will have C(8,2)=28 pairs. As a result, the final feature dimension for the beamforming weight prediction is 21x28=588. For more details of GCC feature extraction, please refer to [5, 19].

B. Predicting Beamforming Weights in Frequency Domain

The details of the beamforming network are illustrated in Fig. II-A. The phase features (i.e. the GCC features in this section) are extracted from the input waveforms using a 0.2s window length and 0.1s window shift, resulting in 10Hz frame rate. A feedforward DNN is used to map the phase features to beamforming weights in the frequency domain. For example, if there are 8 channels, the sampling rate is 16 kHz, and the FFT length is 512, the weight vector will have 257x8x2=4112 real values in every frame. Mean pooling is used to take the average weight vectors from an utterance as we assume that the speaker does not move within one utterance. Readers are referred to [5] for more details of beamforming network description.

III. Spatial Covariance Features

GCC captures the TDOA information in time domain. In this section, we describe frequency domain phase features extracted from spatial covariance matrix and discuss about the connection between these two types of features.

A. Computing Spatial Covariance Matrix

The spatial covariance matrix of the observed multi-channel signals at a time frequency bin is defined as

$$\mathbf{\Sigma}(t,f) = E[\mathbf{y}(t,f)\mathbf{y}^H(t,f)] \tag{3}$$

where t is the frame index, $E[x]$ denotes the expectation of random variable x, and ^{H} denotes Hermitian transpose. $\mathbf{y}(t,f) = [Y_1(t,f), ..., Y_J(t,f)]^T$ is the vector of observed signals of all channels in frequency domain. The frame length and shift used in (3) are 0.025s and 0.01s respectively, which are shorter than those used for GCC computation. In practice, we assume that speech statistics are slowly varying and estimate the spatial covariance matrix as a moving average:

$$\hat{\mathbf{\Sigma}}(t,f) = \frac{1}{2L+1} \sum_{c=-L}^{c=L} [\mathbf{y}(t+c,f)\mathbf{y}^H(t+c,f)] \tag{4}$$

where L is the context size. In this study, we use $L = 10$ to cover 21 contextual frames. Hence, the spatial covariance matrix is estimated from slightly more than 0.21s of input speech if the frame shift is 0.01s. This context span is similar to the 0.2s window used to compute the GCC features.

The diagonal elements of the spatial covariance matrix are the power spectrum of the signals at different channels and hence carry the spectral information of the inputs. The off-diagonal elements are the cross spectrum between channels, so they also contain the phase difference between channels and the spatial information of the target signal and interference. In the MVDR beamforming formulation, the beamforming weights can be solely determined by two types of spatial covariance matrix, i.e. the noise covariance estimated from noise dominant time frequency bins and the noisy speech covariance estimated from speech dominant bins [9]. Therefore, the spatial covariance matrix of the input signals contains the necessary information for neural networks to predict beamforming weights. In the following sections, we will describe how to extract suitable features for the beamforming network.

B. Normalization by Average Power Spectrum of Channels

Although the spectral information contained by the spatial covariance matrix (e.g. the power spectrum at the diagonal elements) may be useful for beamforming weight prediction, it causes different scales in different time frequency bins as illustrated in Fig. III-B. As our objective is to predict beamforming weights which are mainly determined by the spatial information, we need to reduce the variation due to spectral information and make spatial information more salient. In this paper, we propose to normalize the elements of the covariance matrix by the average power of the signal at

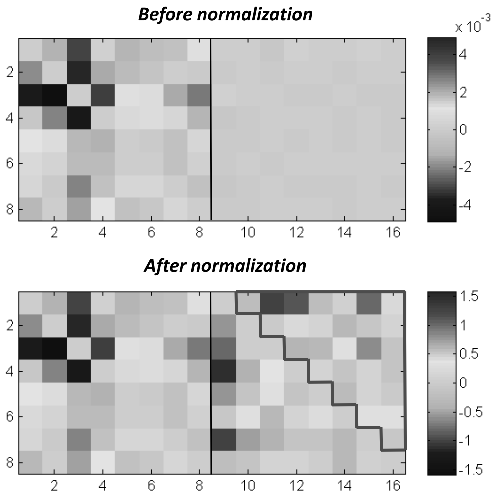

Fig. 3. Effect of normalizing spatial covariance by average power spectrum. The top figure shows the spatial covariance matrices (imaginary part only) of two randomly selected time frequency bins before normalization. Each spatial covariance is an 8x8 matrix as we used 8 microphones to generate this figure. Due to the very different speech power in these two bins, the values of the spatial covariance matrices have very different scales. The spatial covariance after normalization is shown in the bottom figure, and similar scale is observed in the two bins. The elements in the upper triangular matrix within the red lines as shown in the bottom right figure are used as the features for beamforming network.

a time frequency bin. Specifically, the normalized covariance matrix is obtained by

$$\tilde{\mathbf{\Sigma}}(t,f) = \hat{\mathbf{\Sigma}}(t,f)/\hat{P}(t,f) \tag{5}$$

where $\hat{P}(t,f)$ is the estimated average power of the channels and estimated as

$$\hat{P}(t,f) = \frac{1}{J}\sum_{j=1}^{J}\hat{\sigma}_j^2(t,f) \tag{6}$$

where $\hat{\sigma}_j^2(t,f)$ is the j^{th} diagonal element of the covariance matrix and represents the power of the received signal at channel j. The effect of the normalization is shown in Fig. III-B. After the normalization, the values of the two time frequency bins are similar, making it easier for the DNN to learn.

C. Extracting Features from Spatial Covariance

After the normalization, the upper triangular matrix elements (excluding the diagonal elements) as illustrated in Fig.III-B are used as input features for beamforming network. These elements mainly capture the phase differences of channel pairs. For each spatial covariance matrix, we construct a feature vector by stacking the real and imaginary values of the selected elements. The dimension of the feature vector is 2*C(J,2), where C(J,2) is the number of unique channel pairs. For example, if J=8, a feature vector of 56 dimensions will be extracted from each spatial covariance matrix.

For each frame, there are totally K frequency bins, and hence K spatial covariance matrices. The value of K is typically 129 and 257 for 8 kHz and 16 kHz sampling rates, respectively. If we simply concatenate the feature vectors of all frequency bins, the final feature vector for each frame will be huge. For example, if K=257 and J=8, we will have a 56x257=14,392 dimensional feature vector. Such a high dimensional feature vector will cause the network to overfit to the training data easily and also introduces high computational cost.

To deal with the large number of potential features, we proposed three methods of dimension reduction listed as follows:

1) **Sampling method** Use only spatial covariance matrices of frequency bins sampled uniformly from the frequency axis. For example, if we use only 1 frequency bin from every 10 consecutive frequency bins (denoted as 10:10:257 following the Matlab notation), we can reduce the number of features by 10 times.
2) **Row method** Take only the first row of the spatial covariance matrix which contains the phase delay between the first channel (used as the reference channel) and all other channels. The number of features from each spatial covariance matrix is reduced from 2*C(J,2)=J*(J-1) to 2*(J-1).
3) **Filterbank method** First sum neighboring frequency bins into filterbanks, then extract features as described above. For example, if we merge every 5 frequency bins into one filterbank, we can reduce the number of features by 5 times.

It is not easy to see which of the three methods is the best way of building feature vectors from spatial covariance matrices. They will be compared experimentally.

D. Comparison of GCC and Spatial Covariance Features

The features extracted from the spatial covariance matrices and the GCC features are closely related. The GCC in the frequency domain shown in (1) is a cross-spectrum between channel i and j and normalized by the product of magnitude in the two channels. The (i,j)th element of spatial covariance matrix $\tilde{\mathbf{\Sigma}}(t,f)$ is also a normalized cross-spectrum between channel i and j, while the normalization term is the average power of all channels. Hence, the GCC in frequency domain carries similar information as the spatial covariances. The major difference between the GCC features and the spatial covariance features is that the GCC features are converted back to time domain and measures the time delay between microphone pairs in terms of number of samples. The phase information in all frequencies is summed together. On the other hand, the spatial covariance features are extracted directly from the frequency domain and measure the phase delay in individual frequency bins or filterbanks. Hence, we can loosely say that the GCC features focus on the time resolution while the spatial covariance features focus on the frequency resolution. Another difference is that GCC is usually extracted using long windows, e.g. 0.2s, while spatial covariance is estimated by taking the average over many shorter windows, e.g. 0.025s.

IV. Experiments

A. Settings

The beamforming network with spatial covariance features is evaluated on the AMI meeting transcription task. We used both simulated and real array signals for the training of the beamforming and acoustic model networks. The real and simulated data share the same array geometry, i.e. a circular array with 8 microphones and 20cm diameter. The sampling rate is 16 kHz. The simulated data are generated by convolving single-channel clean speech utterances with artificial room impulse responses (RIRs). The clean utterances come from the training set of the WSJCAM0 training set [20] which contains 7,861 sentences. The RIRs are generated by using the image method [21] with various room sizes and T60 reverberation times. Three room sizes are used, including small, medium, and large rooms. The T60 reverberation time is randomly sampled from 0.1s to 1.0s. After the reverberant array signal is simulated, additive noise samples from the REVERB Challenge corpus [1] are added at SNR levels randomly chosen from 0dB to 30dB. In total, 90 hours of simulated array data are generated.

The real array signals are from the multiple distant microphone (MDM) scenario of the AMI meeting corpus [4]. The training set contains about 75 hours of data, while the eval set contains about 8 hours of data. Besides the array signals, the AMI corpus also contains close-talk microphone data that were recorded in parallel with the array signals. The close-talk microphone data is used to train and test another acoustic model to show the upper bound of beamforming and other speech enhancement techniques.

There are two steps in the training of the beamforming networks. In the first step, we teach the network to perform DS beamforming. This is achieved by training the beamforming network to minimize the mean square error between the predicted beamforming weights and the ideal DS weights in the frequency domain. The first step is also called initialization step and carried out on the simulated data where we know the true DOA and hence the ideal weights. The weights of the beamforming network are initialized as Gaussian random noise. In the second training step, the beamforming network and the acoustic model network are trained together on the training set of the AMI task to minimize the cross entropy (CE) cost function of ASR. Therefore, the second step optimizes the beamforming network specifically for ASR. According to our previous study [5], the initialization step is necessary to achieve successful training of beamforming network. Both training steps are implemented in Matlab.

After the beamforming network is trained, it is used to generate enhanced filterbank features. The features are then used to train the DNN acoustic model from scratch using the Kaldi speech recognition toolkit [22], first using the CE cost function, then using the sequential cost function. For ASR decoding, a trigram language model trained from the word label of the 75 hours training data is used.

TABLE I
Speech recognition performance (WER%) on AMI eval set. "Dim" is the size of feature vectors of the beamforming network. "Cost" is the cost function used to train the networks: "CE" is cross entropy cost of classifying frames into phone states of ASR, and "MSE" is the mean square error between the predicted and ideal DS beamforming weights. For the details of the different feature types, please see the description in Section IV-B.

Methods	BF Network Training Configurations			WER (%)
	Features	Dim	Cost	
IHM	-	-	-	25.5
SDM1	-	-	-	53.8
DS	-	-	-	47.9
BF networks	GCC	588	CE	44.7
	SpaCov 10:10:257 sample	1400	MSE	46.4
			CE	44.7
	SpaCov 10:10:257 sample + logSpec	1657	MSE	46.3
			CE	44.7
	SpaCov 5:5:257 sample	2856	MSE	46.1
			CE	44.4
	SpaCov 5:5:257 filterbank		MSE	45.9
			CE	**44.1**
	SpaCov 2:2:257 row	1792	MSE	46.3

B. Results

The word error rate (WER) on the eval set of the AMI corpus is shown in Table I. For comparison purpose, the performance of close-talk microphone (IHM), far-talk microphone (SDM1), and delay-and-sum beamforming are also shown. From the table, the BF network using GCC features and cross-entropy cost function obtains a WER of 44.7%, which is significantly lower than the DS beamforming's 47.9%. This result shows the advantage of the BF network which can be trained to optimize ASR results. The rest of the rows show the results obtained by using BF network with various kinds of spatial covariance features.

- **Sampling method** "SpaCov 10:10:257 sample" stands for we take 1 frequency bin's spatial covariance out of every 10 bins. The number of features in this case is 2C(8,2)x25=1400. It is observed that when using CE cost function, the results obtained is the same as that with the GCC features. If we double the number of selected bins in "SpaCov 5:5:257 sample", we obtain improvements in WER with both MSE and CE cost functions.
- **Sampling + log spectrum** In "SpaCov 10:10:257 sample + logSpec", we added 257 dimensions of average log spectrum of all frequency bins. The results obtained are similar to that without using log spectrum. This suggests that the spectral information, as represented by the log spectrum, is not very useful for determining the beamforming weights at the current network. Hence, we do not use log spectrum in the rest of the experiments.
- **Filterbank method** In "SpaCov 5:5:257 filterbank", we first average the spatial covariance matrices of every 5 frequency bins, then extract features. Although the number of features are the same as that in "SpaCov

5:5:257 sample", the WER obtained with the filterbank method is consistently lower than that obtained by the sampling method. This shows that the filterbank method is a more reliable way to extract spatial information for beamforming, perhaps due to the fact that all frequency bins information are used.

- **Row method** In "SpaCov 2:2:257 row", we take only the first row of each spatial covariance matrix from selected frequency bins. The WER obtained is not significantly better than the "SpaCov 2:2:257 row" although more frequency bins are used. This may suggest that it is important to take the phase information of all channel pairs for robustness.

From the results in Table I, we can conclude that the filterbank method of extracting spatial covariance features is the most effective. With the best spatial covariance features, the beamforming network achieved 44.1% WER, which is lower than the 44.7% obtained with the GCC features. The results show that spatial covariance features are an alternative choice for building beamforming network.

V. Conclusions

In this paper, we improved the deep beamforming network by using spatial covariance features to replace GCC features. Unlike the GCC features that capture the TDOA information of microphones in the time domain, the spatial covariance features capture the phase differences of microphones at different frequencies. Three methods are proposed to extract features from spatial covariance matrices and the filterbank-based feature extraction method is found to perform the best in experiments. Moderate and consistent improvements are observed by using spatial covariance features than using GCC features for beamforming network.

Currently, the beamforming network is initialized by the DS beamforming. In the future, we will investigate the use of beamforming network to implement more advanced beamforming. For example, the MVDR beamforming relies on both the noise and speech SCM to determine the beamforming parameters. A beamforming network with both speech and noise covariance information could perform better.

VI. Acknowledgment

Xiong Xiao, Eng Siong Chng, and Haizhou Li are supported by DSO funded project MAISON DSOCL14045. Shinji Watanabe is supported by MERL.

References

[1] Keisuke Kinoshita, Marc Delcroix, Takuya Yoshioka, Tomohiro Nakatani, Armin Sehr, Walter Kellermann, and Roland Maas, "The reverb challenge: A common evaluation framework for dereverberation and recognition of reverberant speech," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2013, pp. 1–4.

[2] Jon Barker, Ricard Marxer, Emmanuel Vincent, and Shinji Watanabe, "The thirdchime'speech separation and recognition challenge: Dataset, task and baselines," in *2015 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU 2015)*, 2015.

[3] Jelinek Summer Workshop on Speech and Language Technology, "https://www.ee.washington.edu/student/jsalt2015/index.html," 2015.

[4] Steve Renals, Thomas Hain, and Hervé Bourlard, "Recognition and understanding of meetings the ami and amida projects," in *IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU*, Kyoto, 2007.

[5] Xiong Xiao, Shinji Watanabe, Hakan Erdogan, Liang Lu, John Hershey, Michael L Seltzer, Guoguo Chen, Yu Zhang, Michael Mandel, and Dong Yu, "Deep beamforming networks for multi-channel speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016.

[6] Pawel Swietojanski, Arnab Ghoshal, and Steve Renals, "Hybrid acoustic models for distant and multichannel large vocabulary speech recognition," in *IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU*, 2013, pp. 285–290.

[7] Barry D Van Veen and Kevin M Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE assp magazine*, vol. 5, no. 2, pp. 4–24, 1988.

[8] Jack Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proceedings of the IEEE*, vol. 57, no. 8, pp. 1408–1418, 1969.

[9] Mehrez Souden, Jacob Benesty, and Sofiène Affes, "On optimal frequency-domain multichannel linear filtering for noise reduction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 260–276, 2010.

[10] Tara N. Sainath, Ron J. Weiss, Kevin W. Wilson, Arun Narayanan, and Michiel Bacchiani, "Factored spatial and spectral multichannel raw waveform cldnns," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2016.

[11] Tara N Sainath, Ron J Weiss, Kevin W Wilson, Arun Narayanan, Michiel Bacchiani, and Andrew Senior, "Speaker location and microphone spacing invariant acoustic modeling from raw multichannel waveforms," in *IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU*, 2015, pp. 30–36.

[12] Jahn Heymann, Lukas Drude, and Haeb-Umbach Reinhold, "Neural network based spectral mask estimation for acoustic beamforming," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016.

[13] Jahn Heymann, Lukas Drude, Aleksej Chinaev, and Reinhold Haeb-Umbach, "Blstm supported gev beamformer front-end for the 3rd chime challenge," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 444–451.

[14] Michael L. Seltzer, Bhiksha Raj, and Richard M. Stern, "Likelihood-maximizing beamforming for robust hands-free speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 5, pp. 489–498, sep 2004.

[15] Yedid Hoshen, Ron J. Weiss, and Kevin W. Wilson, "Speech acoustic modeling from raw multichannel waveforms," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. apr 2015, pp. 4624–4628, IEEE.

[16] Tara N Sainath, Ron J Weiss, Andrew Senior, Kevin W Wilson, and Oriol Vinyals, "Learning the speech front-end with raw waveform cldnns," in *Interspeech*, 2015.

[17] Hakan Erdogan, John Hershey, Shinji Watanabe, Michael Mandel, and Jonathan Le Roux, "Improved mvdr beamforming using single-channel mask prediction networks," in *INTERSPEECH*, 2016.

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

[19] Xiong Xiao, Shengkui Zhao, Xionghu Zhong, Douglas L Jones, Eng Siong Chng, and Haizhou Li, "A learning-based approach to direction of arrival estimation in noisy and reverberant environments," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 2814–2818.

[20] T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals, "WSJCAM0: a british english speech corpus for large vocabulary continuous speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1995, pp. 81–84.

[21] J.B. Allen and D.A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943?950, April 1979.

[22] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, Dec. 2011.