# CHANNEL SELECTION BASED ON MULTICHANNEL CROSS-CORRELATION COEFFICIENTS FOR DISTANT SPEECH RECOGNITION

*Kenichi Kumatani*[1], *John McDonough*[2], *Jill Fain Lehman*[1,2], *and Bhiksha Raj*[2]

[1]Disney Research, Pittsburgh
Pittsburgh, PA 15213, USA

[2]Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh, PA 15213, USA

**ABSTRACT**

In theory, beamforming performance can be improved by using as many microphones as possible, but in practice it has been shown that using all the possible channels does not always improve recognition performance [1, 2, 3]. In this work, we present a new channel selection method in order to increase the computational efficiency of beamforming for distant speech recognition (DSR).

To achieve better performance, we treat a channel that is uncorrelated with others as unreliable and choose a subset of microphones whose signals are more highly correlated with each other. We use the multichannel cross-correlation coefficient (MCCC) [4] as a measure for selecting the reliable channels. The selected channels are then used for beamforming.

We evaluate our channel selection technique with DSR experiments on real children's speech data captured using a linear array with 64 microphones. A single distant microphone provided a word error rate (WER) of 15.4%, which was reduced to 8.5% by super-directive beamforming with all the sensors. The experimental results suggest that almost the same recognition performance can be obtained with half the number of sensors in case of super-directive beamforming. Maximum kurtosis beamforming [5] with 48 channels out of 64 sensors achieved a WER of 5.7%, comparable to the close talking microphone's WER of 5.2%.

***Index Terms***— channel selection, microphone arrays, beamforming, speech recognition

## 1. INTRODUCTION

There has been great interest in distant speech recognition (DSR) [6]. Such techniques can relieve users from the necessity of donning close taking microphones (CTMs) before interacting with automatic speech recognition (ASR) systems. DSR can be especially useful for young children who may find CTMs too cumbersome and intrusive to use in interactive attractions.

The presence of a noise source and reverberation effects in real environments severely degrade the performance of a DSR system. Depending on the distance between each microphone and the noise source signal, some channels will have lower signal-to-noise ratios (SNR) than others, especially when a large microphone array is used. The reverberation effects also differ among the sensors. Therefore, the performance of speech enhancement might not always be improved by using as many microphones as possible in a real environment. Moreover, it is generally assumed in microphone array processing that all the microphones have the same properties. This assumption may not hold due to variations in the system response from the microphone to the analog-to-digital converter (ADC) [7, 8].

Various methods have been proposed for selecting a suitable channel or using a cluster of microphones. These methods can be categorized into the following approaches:

- selecting a channel with a high SNR [2];
- choosing a channel to which a speech recognizer gives the maximum likelihood [9];
- measuring how much the system's outputs have changed by a noise adaptation technique based on the comparison of word hypotheses of uncompensated and compensated features, and choosing the one with the smallest change [1];
- calculating the class separability measure of feature vectors and selecting the channel which maximizes the separation measure [3]; and
- clustering microphones based on the distance between two microphones and choosing the cluster of microphones according to the proximity measure to a speaker that considers the distance between the reference microphone and speaker as well as the size of the cluster [10, 11].

The SNR-based method is simple and fast to calculate but requires voice activity detection that often fails in noisy environments. Moreover, the SNR measure does not consider any information about ASR.

In terms of ASR, it might be straightforward to use outputs from the speech recognizer for channel selection. As Wölfel noted in [3], however, the disadvantage of this approach is that at least one decoding process is required for each channel in order to avoid mismatch between different channels. Such additional calculation leads to a drastic increase in computational time.

In contrast to the SNR measure, the class separability criterion can take into account speech features for ASR and requires less computation than the decoder-based methods. Wölfel demonstrated in [3] that the channel selection method based on the class separability criterion provided better recognition performance than the SNR-based approach. However, Wölfel selected a single channel and thus did not consider using beamforming, which could improve

First of all, the authors would like to thank Prof. Jessica Hodgins for giving us the opportunity to study this work. The authors would also like to thank Cedrick Rochet for his support in developing the Mark IV microphone array. Also due thanks are Wei Chu, Spencer Diaz, Jerry Feng, Ishita Kapur, and Moshe Mahler for their assistance in collecting the audio-visual material used for the experiments described in this work.

recognition performance. Moreover, the computation required by his method is still significant in the case of multi-channel processing. In contrast, we develop a technique which selects a set of channels for microphone array processing.

In [11], Himawan et al. addressed the situation where microphones are placed on an ad hoc basis. Accordingly, clustering of microphones must be done without any knowledge of microphone positions. In contrast, we consider the situation where the microphones are located methodically and those positions are given as prior knowledge. This assumption allows us to simplify the problem significantly.

In essence, we consider the multichannel cross-correlation coefficient (MCCC) [4] as a measure for selecting the reliable channels. The MCCC represents correlation among more than two channels and the cross-correlation coefficient can be viewed as the special case where the MCCC is calculated with two channels. Although Benesty et al. originally proposed the MCCC for the speaker localization problem in [4], we use the maximum MCCC criterion for channel selection.

The basic idea behind the algorithm is that signals of unreliable channels are uncorrelated with many others. For the sake of computational efficiency, we first compensate for the delays of the signals based on the phase transform (PAHT) [6]. After the multichannel signal is aligned, we compute the maximum MCCC and then choose a set of channels with the maximum MCCC. Finally, beamforming and post-filtering are performed on the selected channels. We demonstrate the effectiveness of our channel selection technique through a series of DSR experiments on the *real* data captured with the *real* microphones. In these experiments we used both traditional super-directive beamforming [6, S13] and state-of-the-art maximum kurtosis beamforming, which adjusts its weights to maximize kurtosis of the beamformer's outputs subject to the distortionless constraint for a look direction [5].

We also investigated other microphone array design methods [12, 13] in order to reduce the number of microphones for beamforming. Logarithmically spaced and non-redundant linear array design methods were evaluated in terms of recognition performance.

The balance of this paper is organized as follows. Section 2 describes the formulation of the problem for microphone array processing and defines the notations used in this paper. Section 4 reviews the MCCC. Section 5 presents our channel selection method based on the maximum MCCC. Recognition experiments are described in section 6. Conclusions of this work and future plans are summarized in section 7.

## 2. PROBLEM FORMULATION

Consider the anechoic situation shown in Figure 1 where a single source signal is captured with the microphone array.

In the time domain, a vector of the $M$-channel signal captured with $M$ microphones at discrete time $n$ can be denoted as

$$\mathbf{x}_M[n] = [x_1[n], x_2[n], \cdots, x_M[n]]^T . \tag{1}$$

In that case, the observation vector of source signal $s[n]$ can be expressed as

$$\mathbf{x}_M[n] = \begin{bmatrix} a_1 s[n - T_p] \\ \vdots \\ a_M s[n - T_p - \tau_{1M}] \end{bmatrix} + \begin{bmatrix} v_1[n] \\ \vdots \\ v_M[n] \end{bmatrix} \tag{2}$$

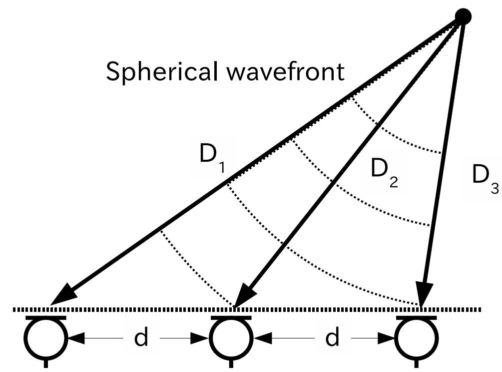


**Fig. 1**. Illustration of the single source signal model under the near-field assumption.

where $a_m$ is the attenuation factor from the source to microphone $m$, $T_p$ is the propagation time to the microphone 1, $\tau_{1m}$ is a time delay between microphone 1 and $m$, and $v_m[n]$ is an additive noise signal.

We denote the signal model of (2) in the (subband) frequency domain as

$$\mathbf{X}_M(e^{j\omega n}) = \begin{bmatrix} A_1 S_1(e^{j\omega(n-T_p)}) \\ \vdots \\ A_M S_M(e^{j\omega(n-T_p-\tau_{1M})}) \end{bmatrix} + \begin{bmatrix} V_1(e^{j\omega n}) \\ \vdots \\ V_M(e^{j\omega n}) \end{bmatrix}. \tag{3}$$

Estimating the time delay of arrival (TDOA) $\tau_{1m}$ is the first step for source localization. This problem is not an easy task in real acoustic environments where each microphone also captures multiple attenuated and delayed replicas of the source signal due to reflections from, for example, tables and walls.

In our channel selection algorithm, the TDOA is first estimated in order to align the signal and calculate the correlation measure among the multiple microphones more accurately.

## 3. TIME DELAY ESTIMATION

In this work, we use the phase transform (PHAT) for time delay estimation. It is a variant of generalized cross-correlation (GCC) and, is perhaps, the most widely used method due to its computational efficiency and robustness against noise [14, 6]. The PHAT between two microphones $m$ and $n$ can be expressed as

$$\rho_{mn}(\tau) = \int_{-\pi}^{\pi} \frac{X_m(e^{j\omega\tau}) X_n^*(e^{j\omega\tau})}{|X_m(e^{j\omega\tau}) X_n^*(e^{j\omega\tau})|} e^{j\omega\tau} d\omega, \tag{4}$$

where $X_m(e^{j\omega\tau})$ denotes the spectrum of the signal captured with the $m$-th sensor. We use the Hamming analysis window for calculating the spectrum in (4). The normalization term in the denominator is intended to weight all frequencies equally. It has been shown that those weights lead to more robust time delay estimation [14]. Time delay of arrival (TDOA) is then estimated from

$$\hat{\tau}_{mn} = \max_{\tau} \rho_{mn}(\tau). \tag{5}$$

Thereafter, an interpolation is performed to overcome the granularity in the estimate corresponding to the sampling interval.

## 4. MULTICHANNEL CROSS-CORRELATION COEFFICIENT

Once the time delays of $M$ signals are estimated based on the PHAT, we can obtain the time-aligned signal as

$$\mathbf{x}_{d,M}[n] = [x_1[n], x_2[n+\hat{\tau}_{12}], \cdots, x_M[n+\hat{\tau}_{1M}]]^T . \quad (6)$$

In order to calculate the MCCC, we first need a spatial correlation (covariance) matrix of observations. The spatial correlation matrix can be expressed as

$$\mathbf{R}_M = E\left\{\mathbf{x}_{d,M}[n]\mathbf{x}_{d,M}^T[n]\right\}. \quad (7)$$

Then, given the TDOA estimates, the MCCC can be computed as

$$\varrho_M^2 = 1 - \frac{\det[\mathbf{R}_M]}{\Pi_{i=1}^M \sigma_i^2}, \quad (8)$$

where det[.] stands for the determinant and $\sigma_i^2$ is the $i$-th diagonal component of the spatial correlation matrix $\mathbf{R}_M$. It can be readily confirmed that the MCCC is equivalent to the cross-correlation coefficient normalized by the energy in the case of $M = 2$ [4].

Chen, Benesty and Huang originally used the MCCC for estimating the direction of arrival (DOA) in the far field assumption [15, 16]. In their work, the MCCC was viewed as a function of the time delays. In contrast to their work, we estimate the TDOA based on the PHAT which leads to a drastic computational reduction in the case of the near field assumption and calculate the MCCC with the fixed time delays for the channel selection.

In [15, 16], for source localization, Chen, Benesty and Huang showed

$$0 \leq \frac{\det[\mathbf{R}_M]}{\Pi_{i=1}^M \sigma_i^2} \leq 1, \quad (9)$$

and noted that the MCCC has the following properties:

- $0 \leq \varrho_M^2 \leq 1$;
- $\varrho_M^2 = 1$ if two or more signals are perfectly correlated;
- $\varrho_M^2 = 0$ if all the signals are completely uncorrelated with each other; and
- if one of the signals is completely uncorrelated with the $M-1$ other signals, the MCCC of all the signals will be equal to that of those $M-1$ remaining signals.

## 5. CHANNEL SELECTION

Here we describe our channel selection method. Let us assume that we select $M_s$ channels with the maximum MCCC out of $M$ microphones. We ideally want to find a set of channels $\mathcal{C}_{M_s}$ which provides the largest MCCC among all the possible combinations as follows:

$$\hat{\mathcal{C}}_{M_s} = \underset{\mathcal{C}_{M_s}}{\text{argmax}} \ \varrho_{M_s}^2. \quad (10)$$

An exhaustive search requires computing the MCCC ${}_M C_{M_s}$ times. If we have a large number of microphones, the computation becomes enormous.

We avoid this problem by iteratively reducing the number of the search candidates from $M$ to $M_s$. More specifically, we ignore the channel that provides the smallest MCCC and keep the remaining channels for the next step. This process is repeated until we obtain the desired number of channels, $M_s$. By doing so, the computation for the MCCC is reduced from ${}_M C_{M_s}$ to $\sum_{i=0}^{M-M_c} M - i$.

Our channel selection algorithm is summarized as follows:

1. estimate the time delays of the $M$-channel signal with (5) and align the signal,
2. push all the $M$ channels to a search stack,
3. denoting the number of the candidates in the search stack as $M_c$, find a set of the $M_c - 1$ channels with the largest MCCC,
4. remove the channel which does not belong to the set with the largest MCCC in Step 2. from the stack, and
5. go to Step 3. if $M_c > M_s$.

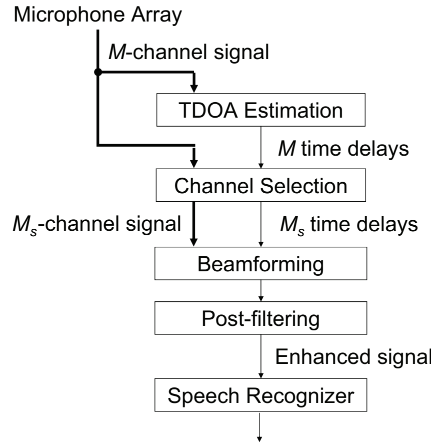


**Fig. 2**. A flow chart of our distant speech recognition system.

## 6. EXPERIMENTS

Figure 2 shows a flow chart of the distant speech recognition (DSR) system used in our experiments. Our DSR system involves the time delay estimation step described in Section 3, the channel selection method depicted in Section 5, beamforming, post-filtering and automatic speech recognition (ASR) components which we will describe in this section.

In our experiments, a beamforming algorithm is performed on the channel data selected by the algorithm proposed here. We consider both the widely used super-directive beamforming [14, 6] and one of the state-of-art techniques, maximum kurtosis beamforming [5]. By reducing the number of channels, the computation required for beamforming can be significantly decreased without degrading recognition performance (Section 6.1). Following beamforming, Zelinski post-filtering [17], a variant of Wiener filtering, is carried out in order to remove the noise uncorrelated among the sensors.

Our basic automatic speech recognition (ASR) system was trained on two publicly available corpora of children's speech:

1. the Carnegie Mellon University (CMU) Kids' Corpus, which contains 9.1 hours of speech from 76 speakers;

2. the Center for Speech and Language Understanding (CSLU) Kids' Corpus, which contains 4.9 hours of speech from 174 speakers.

The feature extraction used for the ASR experiments reported here was based on cepstral features estimated with a warped *minimum variance distortionless response* (MVDR) spectral envelope of model order 30 [6, §5.3]. Front-end analysis involved extracting 20 cepstral coefficients per frame of speech, and then performing *cepstral mean normalization* (CMN). The final features were obtained by concatenating 15 consecutive frames of cepstral coefficients together, then performing *linear discriminant analysis* (LDA), to obtain a feature of length 42. The LDA transformation was followed by a second CMN step, then a global semi-tied covariance transform estimated with a maximum likelihood criterion [18].

HMM training was conducted initializing a context independent model with three states per phone with the global mean and variance of the training data. Thereafter, five iterations of Viterbi training [6, §8.1.5] were conducted. This was followed by an additional five iterations whereby optional silences and optional breath phones were allowed between words. The next step was to treat all triphones in the training set as distinct and train three-state single-Gaussian models for each. Then state clustering was conducted as in [19]. In the final stage of conventional training, the context-dependent state-clustered model was initialized with a single Gaussian per codebook from the context-independent model; three iterations of Viterbi training followed by splitting the Gaussian with the model training steps. These steps were repeated until no more Gaussians had sufficient training counts to allow for splitting. The conventional model had 1,200 states and a total of 25,702 Gaussian components. Conventional training was followed by *speaker-adapted training* (SAT) as described in [6, §8.1.3].

In our experiments, the ASR system consisted of three passes:

1. Recognize with the unadapted conventionally trained model;
2. Estimate *vocal tract length normalization* (VTLN) [20], *maximum likelihood linear regression* (MLLR) [21] and *constrained maximum likelihood linear regression* (CMLLR) [22] parameters, then recognize once more with the adapted conventially trained model;
3. Estimate VTLN, MLLR and CMLLR parameters for the SAT model, then recognize with same.

For all but the first unadapted pass, unsupervised speaker adaptation was performed based on word lattices from the previous pass.

### 6.1. Recognition results

Test data for experiments were collected at the Carnegie Mellon University Children's School. The speech material in this corpus was captured with a 64-channel Mark IV microphone array; the elements of the Mark IV were arranged linearly with a 2 cm intersensor spacing. In order to provide a reference for the DSR experiments, the subjects of the study were also equipped with Shure lavelier microphones with a wireless connection to an RME Hammerfall Octamic II preamp and A/D converter. The Octamic II was connected via an ADAT optical cable to a RME Hammerfall HDSPe AIO sound card. A PNC coaxial connection between the Mark IV and the Octamic II ensured that *all* audio capture was sample synchronous. This was required to enable voice prompt suppression experiments. All the audio data were captured at 41.1 kHz with a 24-bit per sample resolution.

The test set consists of 354 utterances (1,297 words) spoken by nine children. The children were native-English speakers (aged four to six). They were asked to play *Copycat*, a listen-and-repeat paradigm in which an adult experimenter speaks a phrase and the child tries to copy both pronunciation and intonation. As is typical for children in this age group, pronunciation was quite variable and the words themselves sometimes indistinct.

| Algorithm | Pass (%WER) | | |
|---|---|---|---|
| | 1 | 2 | 3 |
| Single distant microphone | 38.1 | 19.8 | 15.4 |
| SD beamforming | 24.1 | 11.3 | 8.6 |
| MK beamforming | 21.8 | 8.3 | 5.7 |
| Lapel microphone | 19.3 | 5.9 | 5.2 |

**Table 1**. Word error rates (WERs) for each decoding pass.

The search graph for the recognition experiments was created by initially constructing a finite-state automaton by stringing *Copycat* utterances in parallel between a start and end state. This acceptor was convolved together with a finite-state transducer representing the phonetic transcriptions of the 147 words in the *Copycat* vocabulary. Thereafter this transducer was convolved with the $HC$ transducer representing the context-dependency decision tree estimated during state-clustering [6, §7.3.4].

Table 1 shows word error rates (WERs) of every decoding pass obtained with one of 64 microphones, super-directive (SD) beamforming, maximum kurtosis (MK) beamforming and lapel microphone. In Table 1, the numbers of channels for SD and MK beamforming are 32 and 48 respectively because those settings provided the best results.

Table 1 demonstrates that the improvement from the adaptation techniques is dramatic. The reduction in the WER from the first pass to the third is approximately four-fold in the case of MK beamforming. It is also clear that the performance of far-field speech recognition can be improved by beamforming techniques, and the MK beamforming algorithm achieves the best performance in the experiments. The MK beamforming technique provides almost the same recognition performance as the lapel microphone.

We also investigated the WERs as a function of the number of channels used for the beamforming algorithm. Figure 3 shows the WERs of the third pass for the number of channels when SD and MK beamforming algorithms were performed. The MK beamformer provides better recognition performance than SD beamforming when the same number of channels is used. Using all the microphones does not provide the best recognition performance because the several channels are distorted by the reverberation and noise. The results in Figure 3 suggest that we could improve recognition performance by automatically finding the optimum number of channels although the effect would be relatively small. In practice, the number of channels could be empirically decided based on the computer resources available for the application.

Another interesting result comes from a comparison of the channel selection method with logarithmically spaced and non-redundant microphone array designs [12, 13]. In our case, due to the fixed geometry of the Mark IV, our adoption is to select among the channels with the 2 cm inter-sensor spacing. In other words, the microphone array design method can be viewed as a channel selection method. However, we can combine the channel selection and array design method if we can configure the geometry of the microphone array arbitrarily. For example, we can select the channels from the non-redundant microphone array.

First, we compare our channel method with the logarithmically

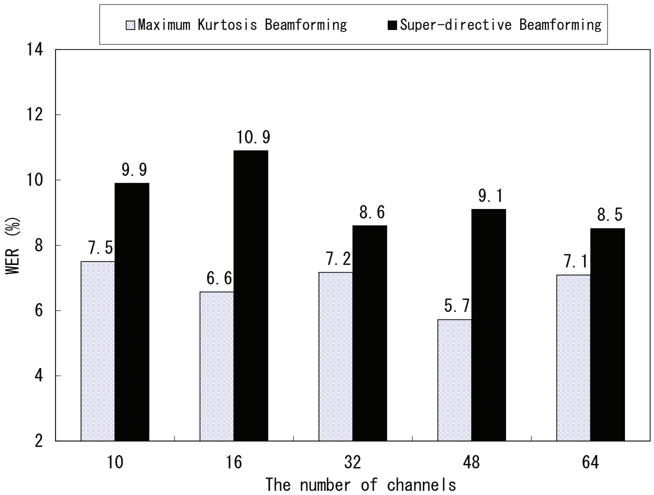


**Fig. 3**. WERs of the third pass as a function of the number of channels.

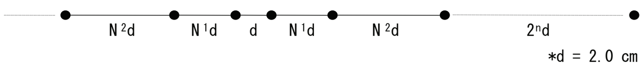


**Fig. 4**. Logarithmically spaced linear microphone array.

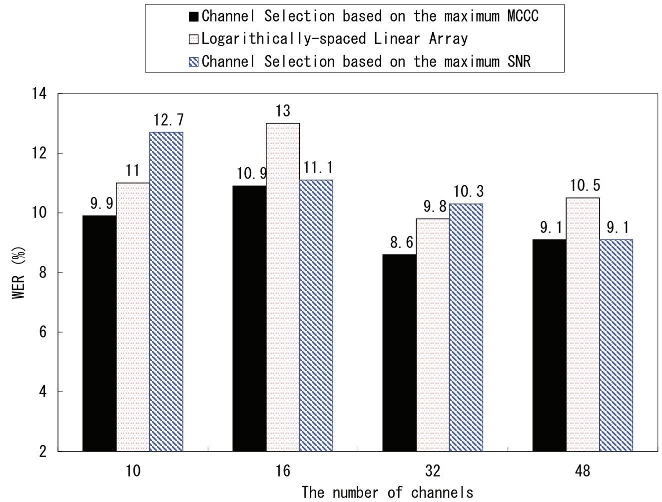


**Fig. 5**. WERs of the third pass as a function of the number of channels in the case of super-directive beamforming.

| Method | WER of the third pass |
|---|---|
| Logarithmically spaced array | 11.0 |
| Non-redundant array | 9.6 |
| Channel selection method | 9.9 |

**Table 2**. WERs for each array design method from the super-directive beamformer with 10 sensors.

spaced linear array shown in Figure 4. In the logarithmically spaced linear array, the sensor is symmetrically placed in the center of the linear array on a logarithmic scale. Figure 5 shows the WERs obtained by the channel selection and logarithmically spaced array for the number of channels. In Figure 5, SD beamforming is performed for the sake of efficiency. Again, due to the restriction of the Mark IV, we cannot change the spacing. Accordingly, the microphones are picked up so that the geometry gets close to the logarithmic design. It is clear from Figure 5 that our channel selection method provides lower WERs than the logarithmically spaced linear array. This improvement occurs because our channel selection method can adaptively choose the channel to some extent in contrast to the static array design.

Figure 5 also shows the WERs obtained by a channel selection algorithm based on the maximum SNR criterion as another baseline. The maximum SNR-based algorithm used here first measures the SNR of each channel from the noise and speech segments aligned by the speech recognizer and then select the channels with better SNRs. Figure 5 illustrates that the maximum SNR-based algorithm performs worse than the method based on the maximum MCCC criterion. The increases in the WERs occur mainly because it is not feasible to precisely measure the SNR in acoustically noisy environments due to the absence of perfect speech activity detection. The results might also suggest that the SNR is not related to the WER.

Finally, we tabulate the WERs of our channel selection method and two array design methods in Table 2 in the case of SD beamforming with 10 sensors. Again, because of the uniform spacing of the MarkIV, we cannot compare our channel selection method with the non-redundant linear array design [13, §3.9] in the case of more than 10 sensors. We can, however, observe from Table 2 that the non-redundant array and our channel selection method provide almost the same recognition performance in the experiment with 10 microphones. This result is promising because the techniques can be combined if we can choose the arbitrary geometry of the microphone array. For instance, we can select the channel of the non-redundant microphone array based on the maximum MCCC criterion.

## 7. CONCLUSIONS

In this work, we have presented a new channel selection algorithm for distant speech recognition (DSR). We have demonstrated through a series of DSR experiments that our algorithm can reduce the number of channels for beamforming effectively. Our channel selection method can also improve recognition performance.

We plan to combine our channel selection and array design methods as well as other conventional channel selection methods. We also plan to extend the algorithm proposed here to the situation where multiple sources are active. Future work also includes a plan to investigate the eigenvalues of the spatial covariance matrix and develop an automatic method to determine the number of channels.

## 8. REFERENCES


[1] Yasunari Obuchi, "Multiple-microphone robust speech recognition using decoder-based channel selection," in *Proc. ISCA Tutorial and Research Workshop on Statistical and Perceptual Audio Processing (SAPA)*, Jeju, Korea, 2004.

[2] Matthias Wölfel, Christian Fügen, Shajith Ikbal, and John W. Mcdonough, "Multi-source far-distance microphone selection and combination for automatic transcription of lectures," in *Proc. Interspeech*, Pittsburgh, Pennsylvania, 2006.

[3] Matthias Wölfel, "Channel selection by class separability mea-

sures for automatic transcriptions on distant microphones," in *Proc. Interspeech*, Antwerp, Belgium, 2007.

[4] Jacob Benesty, Jingdong Chen, and Yiteng Huang, *Microphone Array Signal Processing*, Springer, 2008.

[5] Kenichi Kumatani, John McDonough, Barbara Rauch, Philip N. Garner, Weifeng Li, and John Dines, "Maximum kurtosis beamforming with the generalized sidelobe canceller," in *Proc. Interspeech*, Brisbane, Australia, September 2008.

[6] Matthias Wölfel and John McDonough, *Distant Speech Recognition*, Wiley, New York, 2009.

[7] Ivan Jelev Tashev, *Sound Capture and Processing: Practical Approaches*, Wiley, 2009.

[8] Carsten Sydow, "Broadband beamforming for a microphone array," *Journal of the Acoustical Society of America*, vol. 96, pp. 845–849, 1994.

[9] Y. Shimizu, S. Kajita, K. Takeda, and F. Itakura, "Speech recognition based on space diversity using distributed multi-microphone," in *ICASSP*, Istanbul, Turkey, 2000.

[10] Ivan Himawan, Iain McCowan, and Sridha Sridharan, "Clustering of ad-hoc microphone arrays for robust blind beamforming," in *ICASSP*, Dallas, Texas, 2010.

[11] Ivan Himawan, Iain McCowan, and Sridha Sridharan, "Clustered blind beamforming from ad-hoc microphone arrays," *IEEE Trans. Speech Audio Processing*, vol. 18, pp. –, 2010.

[12] Saeed Gazor and Yves Grenier, "Criteria for positioning of sensors for a microphone array," *IEEE Trans. Speech Audio Processing*, vol. 3, pp. 294–303, 1995.

[13] H. L. Van Trees, *Optimum Array Processing*, Wiley-Interscience, New York, 2002.

[14] M. Brandstein and D. Ward, Eds., *Microphone Arrays*, Springer Verlag, Heidelberg, Germany, 2001.

[15] Jingdong Chen, Jacob Benesty, and Yiteng Huang, "Robust time delay estimation exploiting redundancy among multiple microphoens," *IEEE Trans. Speech Audio Processing*, vol. 11, pp. 549–557, 2003.

[16] Jacob Benesty, Jingdong Chen, and Yiteng Huang, "Time delay estimation via linear interpolation and cross-correlation," *IEEE Trans. Speech Audio Processing*, vol. 12, pp. 509–519, 2004.

[17] Claude Marro, Yannick Mahieux, and K. Uwe Simmer, "Analysis of noise reduction and dereverberation techniques based on microphone arrays with postfiltering," *IEEE Transactions on Speech and Audio Processing*, vol. 6, pp. 240–259, 1998.

[18] M. J. F. Gales, "Semi-tied covariance matrices for hidden Markov models," *IEEE Transactions Speech and Audio Processing*, vol. 7, pp. 272–281, 1999.

[19] S. J. Young, J. J. Odell, and P. C. Woodland, "Tree–based state tying for high accuracy acoustic modelling," in *Proc. of HLT*, Plainsboro, NJ, USA, 1994, pp. 307–312.

[20] Ellen Eide and Herbert Gish, "A parametric approach to vocal tract length normalization," in *Proc. of ICASSP*, 1996, vol. I, pp. 346–8.

[21] C.J. Leggetter and P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models," *Jour. on CSL*, vol. 9, no. 2, pp. 171–185, 1995.

[22] M. J. F. Gales, "Maximum likelihood linear transformations for hmm–based speech recognition," *Jour. on CSL*, vol. 12, pp. 75–98, 1998.