Learning to Rank Microphones for Distant Speech Recognition

Samuele Cornell[1]*, Alessio Brutti*[2]*, Marco Matassoni*[2]*, Stefano Squartini*[1]

[1]Università Politecnica delle Marche, Italy
[2]Fondazione Bruno Kessler, Italy

s.cornell@pm.univpm.it, (brutti, matasso)@fbk.eu, s.squartini@univpm.it

Abstract

Fully exploiting ad-hoc microphone networks for distant speech recognition is still an open issue. Empirical evidence shows that being able to select the best microphone leads to significant improvements in recognition without any additional effort on front-end processing. Current channel selection techniques either rely on signal, decoder or posterior-based features. Signal-based features are inexpensive to compute but do not always correlate with recognition performance. Instead decoder and posterior-based features exhibit better correlation but require substantial computational resources.

In this work, we tackle the channel selection problem by proposing MicRank, a learning to rank framework where a neural network is trained to rank the available channels using directly the recognition performance on the training set. The proposed approach is agnostic with respect to the array geometry and type of recognition back-end. We investigate different learning to rank strategies using a synthetic dataset developed on purpose and the CHiME-6 data. Results show that the proposed approach considerably improves over previous selection techniques, reaching comparable and in some instances better performance than oracle signal-based measures.

Index Terms:speech recognition, channel selection, learning to rank, array signal processing

1. Introduction

Nowadays, many application scenarios envision the presence of multiple heterogeneous recording devices. Examples are meeting scenarios [1] or multi-party conversations such as in the CHiME-6 Challenge [2]. However, distant automatic speech recognition (ASR) in the presence of ad-hoc microphone networks is still an open issue and the potential of fusing information from multiple devices towards the common goal of reducing the word error rate (WER) is still not fully exploited.

Audio signals captured by different microphones can be suitably combined at the front-end level by using beamforming techniques [3–9]. However, most of these approaches [3–6] are designed for microphone array applications and do not perform well in ad-hoc microphone scenarios where sensors can be far from each other. Few exceptions are [1, 8, 9] in which ad-hoc microphone networks are explicitly considered in the design of the method.

Another intriguing approach, is to pick up, for each utterance, the best channel without any further processing, or, in alternative, sorting the channels and choosing a promising subset before applying signal-based combination methods or Recognizer Output Voting Error Reduction (ROVER) [10]. This channel selection problem has been addressed in the past either using signal-based hand-crafted features [11–13], decoder-based [14, 15] and posterior-based features [16]. Among the most representative past studies on automatic channel selection, [12] (and previous works from the same authors) investigated both signal-based and decoder-based measures, as well as different strategies for their combination. It was found that envelope variance (EV), despite being signal-based, represents one the most effective channel selection strategy thanks to its ability to detect the reduced dynamic ranges introduced by reverberation. More recently, in [13], another signal-based method relying on Cepstral Distance (CD) was proposed. The main advantage of these signal-based methods is that they are inexpensive with respect to decoder-based measures which require full decoding of all channels. Another option is the posterior-based channel selection method proposed in [16] in which microphones are selected using an entropy measure of posterior probabilities produced by an Acoustic Model (AM) trained on clean speech. While less expensive than decoder-based methods, as it requires only an AM forward pass for each channel, it assumes that a matched AM trained on clean speech is available.

In this paper, we propose MicRank, an alternative, fully neural approach for channel selection. MicRank is agnostic with respect to the properties of the acoustic environment, recording set up and type of ASR back-end. Borrowing from information retrieval [17], we formulate the channel selection task as a learning to rank (LTR) problem where a DNN is trained to rank microphones based on the errors obtained with the ASR back-end on a training set. Within this framework, we explore different loss functions and training strategies by performing experiments on a purposely developed synthetic dataset and CHiME-6. We show that MicRank considerably outperforms several previously proposed channel selection methods and even, in some instances, signal-based oracle measures. Importantly, this is achieved with remarkably lower computational requirements compared to decoder and posterior-based approaches. Our source code is made open source at `github.com/popcornell/MicRank`.

This paper is organized as follows. Section 2 presents the learning to rank paradigm and how it can be adapted to address the channel selection problem. Section 3 describes the experimental set-up, including datasets, baseline methods and neural architectures. Following, in Section 4 we discuss our experimental results and, in Section 5, we draw conclusions.

2. Learning to Rank for Channel Selection

The problem of selecting the best channel among a set of available ones can be best formulated as a ranking problem. In fact, predicting an absolute quality metric (e.g. WER, signal-to-noise ratio (SNR) etc.) for each channel is not necessary as what matters most is relative performance: for a given utterance we want to find the best channel within the available ones, whatever its absolute quality metric is. This requires the model to learn, either implicitly or explicitly, to rank the channels.

Learning to rank is an established framework in the field of information retrieval. Therefore its formulation has to be re-

vised and adapted to channel selection for ASR, in particular for what concerns the relevance of observed samples. In principle the ranking approaches we propose can be used to rank the channels with respect to any metric. Since in this work the ultimate goal is ASR, training labels are derived directly from WER or word accuracy (WA) obtained by the ASR back-end on the training material.

2.1. Ranking Strategies and Losses

Let us assume that U utterances are recorded by M microphones. For each utterance u ($u = 0, \dots, U-1$), given the observation feature vector $\boldsymbol{x}_{u,i}$ for the i-th microphone ($i = 0, \dots, M-1$) and a ranking order (or relevance in information retrieval) $w_{u,i}$, our goal is to define a function $f(\boldsymbol{x}_{u,i})$ that generates the same ranking order: if $w_{u,i} > w_{u,j}$ then $f(\boldsymbol{x}_{u,i}) > f(\boldsymbol{x}_{u,j})$. In the following we describe different training strategies to achieve this goal, graphically depicted in Fig. 1.

2.1.1. Point-wise training

The most straightforward approach to rank the channels is to employ a model trained on each single channel individually to predict its relevance. In this method, given a set of training pairs $(\boldsymbol{x}_{u,i}, w_{u,i})$ for each utterance and microphone, the network is trained to minimize a cross-entropy loss:

$$\mathcal{L}_{\text{XCE}}^{point} = \sum_{u=0}^{U-1} \sum_{i=0}^{M-1} w_{u,i} \log \left[\sigma(f(\boldsymbol{x}_{u,i}))\right], \tag{1}$$

where $\sigma(\cdot)$ is the sigmoid operator. In this case, the relevance label $0 \leq w_j \leq 1$ is a soft label, representing the quality of the speech signal in an absolute term. WA for example, and any other bounded metric can be used straightforwardly. A clipping or normalization strategy instead can be adopted for metrics like WER which are unbounded. Alternatively, the cross-entropy training objective can be replaced by a Mean Squared Error (MSE) objective which does not require any bounded relevance assumption:

$$\mathcal{L}_{\text{MSE}}^{point} = \sum_{u=0}^{U-1} \sum_{i=0}^{M-1} \|w_{u,i} - f(\boldsymbol{x}_{u,i})\|^2. \tag{2}$$

2.1.2. Pair-wise training

With point-wise training the model implicitly learns to rank the channels by learning to predict their absolute quality. However, it does not consider relative performance of the other channels. One way to account for the other microphones is to train the network in a Siamese fashion, as it has been proposed in RankNet [18]. In this case, labels are not required to represent an absolute measure and thus even unbounded metrics can be used directly. For a given utterance u, let us consider feature vectors from two channels $x_{u,i}$ and $x_{u,j}$ with related relevance scores $w_{u,i}$ and $w_{u,j}$. We can define a binary pairwise label as:

$$y_{u,i,j} \begin{cases} 1 & \text{if } w_{u,i} > w_{u,j}, \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

Note that $y_{u,i,j}$ is an hard label (i.e. either 1 or 0) whose value depends on which relevance $w_{u,i}, w_{u,j}$ is higher than the other, and thus on the relative ranking of the two channels. For each training sample $(\boldsymbol{x}_{u,i}, \boldsymbol{x}_{u,j}, y_{u,i,j})$ we can then define a binary cross-entropy loss as:

$$\begin{aligned} \mathcal{L}_{u,i,j} = y_{u,i,j} \log[P(w_{u,i} > w_{u,j})] \\ + (1 - y_{u,i,j}) \log[1 - P(w_{u,i} > w_{u,j})], \end{aligned} \tag{4}$$

where $P(w_{u,i} > w_{u,j})$ is the probability estimated by the network $f(\cdot)$ that $\boldsymbol{x}_{u,i}$ is more relevant than $\boldsymbol{x}_{u,j}$, which can be computed as:

$$P(w_{u,i} > w_{u,j}) = \sigma\left(f(\boldsymbol{x}_{u,i}) - f(\boldsymbol{x}_{u,j})\right). \tag{5}$$

The overall training loss is obtained by summing over all unique microphone pairs and utterances:

$$\mathcal{L}_{\text{BCE}}^{pair} = \sum_{u=0}^{U-1} \sum_{(i,j) \in \mathcal{I}_u} \mathcal{L}_{u,i,j}. \tag{6}$$

where $\mathcal{I}_u = \{(i,j) : |w_{u,i} - w_{u,j}| > \delta, i \neq j\}$ is the set of microphone pairs whose relevance difference in utterance u is larger than δ with $\delta \geq 0$. Thus the size of the training set is upper bounded to $(U-1)(M-1)(M-2)/2$.

2.1.3. List-wise training

In RankNet, the network learns to order the items by comparing them only in a pairwise fashion. However, due to the use of hard labels, the learning process does not take into account the actual difference between two samples as it cares only for relative pairwise ordering. Nonetheless, swapping the ranks of two samples with very similar relevance should be less critical than swapping two samples with a very different relevance.

These problems can be addressed by employing ListNet [19]. Contrary to the pair-wise approach, for each utterance u all available microphones M are used to compute a cross-entropy loss:

$$\mathcal{L}_{\text{XCE}}^{list} = \sum_{u=0}^{U-1} \sum_{i=0}^{M} \mathcal{S}(w_{u,i}) \log[\mathcal{S}(f(\boldsymbol{x}_{u,i}))]. \tag{7}$$

$\mathcal{S}(\cdot)$ is the softmax operator which ensures that both labels and network outputs can be treated as probability distributions. It also enforces that ranking, for each utterance, is determined only by relative performance of each microphone.

3. Experimental Analysis

3.1. Datasets

In order to evaluate our method we experimented with two datasets: a synthetic dataset generated on purpose and the data used in the CHiME-6 challenge. We describe them thereafter.

3.1.1. Synthetic Dataset

We generated a multi-channel synthetic dataset featuring an ad-hoc network with 8 cardioid microphones. Clean speech utterances are uniformly sampled from LibriSpeech [20] using `train-clean-100` for training, `dev-clean` for validation and `test-clean` for test. We used a total of 20k utterances for train and 2k for validation and test splits. Point-source noise from the dataset in [21] is also employed to make the data more realistic. A different acoustic scenario is sampled for each utterance. Using gpuRIR [22] we simulate a rectangular room whose size and reverberation time (T60) are sampled uniformly

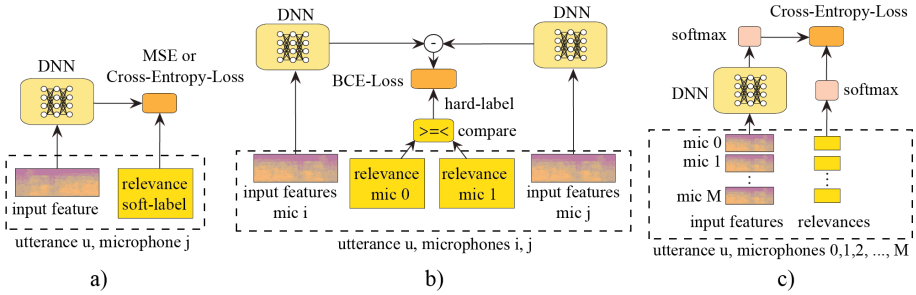

Figure 1: *Training strategies: a) point-wise training; b) pair-wise training with RankNet; c) list-wise training with ListNet.*

between 10 and 60 m^2 and between 0.2 and 0.6 s respectively. The positions and orientations of the speaker, noise and of the 8 microphones are chosen randomly inside the room but with the constraints that the speaker cannot be closer than 0.5 m from any microphone or wall and each microphone should be at least 0.5 m apart from any other. Relevance labels are obtained by training an ASR on the training set using a modified Kaldi [23] LibriSpeech recipe and computing the errors on such set.

3.1.2. CHiME-6

The CHiME-6 Challenge [2] dataset features real dinner parties attended by 4 participants, recorded by 6 Kinect arrays, each with 4 microphones. Devices are distributed in space in order to cover the whole apartment, which may include multiple rooms. The dataset also features oracle speech segmentation and a manually selected reference device for each speech segment. In our experiments we employed the ASR back-end provided by the challenge, using the official Kaldi recipe with the acoustic model and the two-pass decoding in [24].

3.2. Neural Network Architecture

We studied the proposed MicRank LTR framework with the Temporal-Convolutional-Network (TCN) used in [25] based on ConvTasNet separator [26]. We employ as input features 40 logmel filterbanks extracted from 25 ms windows with 10 ms stride. These are fed to a layer normalization and a 40×64 fully connected layer. This latter is followed by 3 blocks each comprised of 5 residual blocks with 1-D dilated convolutions. Each residual block has the same structure as described in [25], with the dilation factor increasing for each successive residual block as $2^0, 2^1, \ldots, 2^4$. As in [25] we use 64 channels for bottleneck convolutions, 128 channels and a kernel size of 3 for depth-wise separable convolutional layers.

The network is fed a fixed-length input corresponding to 200 frames. Speech segments longer than 2 s, are split in chunks which are processed individually. Zero-padding is used for segments shorter than 200 frames. During training, the same relevance is used for all chunks derived from the same speech segment. The network is applied to each microphone channel independently and relevance is obtained via a final 40×1 fully connected layer followed by mean pooling over each 200 frames chunk. In inference, the output score is averaged over all chunks, which are extracted with an overlap factor of 4. This architecture ($\sim 266k$ parameters) requires $\sim 26M$ Floating Point Operations per second (FLOPs) in inference. For comparison, posterior-based selection [16] requires $\sim 359M$ FLOPs for the AM used in the synthetic dataset ($\sim 3.8M$ parameters). Thus the proposed approach is significantly more computationally efficient than decoder and posterior-based techniques.

Models are trained using Stochastic Gradient Descent (SGD). To improve generalization, we employ SpecAugment [27] based Mel-band masking. Learning rate, batch size, weight decay and SpecAugment parameters are tuned on each dataset validation set. As relevance score we use WA, computed by scoring each utterance and each channel of the training set using the ASR back-end. We experimented also with WER using normalization strategies, but without noticeable differences.

3.3. Oracle and Baseline Methods

We evaluate our proposed method against a set of baselines and oracle approaches. We consider as the upper-bound for this task the oracle channel selection obtained by taking, for each utterance, the channel with lowest WER among all the available ones. Moreover, we consider a set of selection strategies that relies on the distance between the speaker and microphones and on oracle signal-based quality metrics. Regarding the latter, we consider Short-Time Objective Intelligibility STOI [28], Signal-to-Distortion Ratio (SDR) [29] and Perceptual Evaluation of Speech Quality (PESQ) [30]. These are computed with respect to oracle non-reverberated clean speech for the synthetic dataset and with respect to close-talk per-speaker microphones for CHiME-6.

As oracle distance from the speaker is not available in CHiME-6, we instead consider the baseline system provided by the challenge organizers which employs weighted prediction error (WPE) [31] followed by BeamformIt [4] applied on a "pseudo-oracle" manually selected array for each utterance. The manual selection is based on the positions and orientations of the speakers obtained via video recordings and is provided by challenge organizers. The alternative, more performing, baseline system based on Guided Source Separation [7] is not considered here as it also exploit oracle diarization.

In addition, we evaluate MicRank against three aforementioned state-of-the-art channel selection methods. In detail, we consider the posterior-based approach proposed in [16] (AM-Entropy in the following) for the synthetic dataset only and two signal-based approaches for both datasets EV [12] and CD [13]. For the former we used the pre-trained LibriSpeech AM available in Kaldi. Regarding [13] we evaluate both the blind version (CD-blind) as well as the oracle version (CD-informed) computed in the same way as the aforementioned signal-based oracle measures. Sub-band weights in EV are tuned on each dataset training set using SGD and a cross-entropy objective for selecting the best channel.

4. Results

Results on the synthetic dataset are reported in Table 1. The upper part of the Table reports results obtained by randomly selecting one of the microphones as well as using oracle measures. Note that, as expected, picking the closest microphone

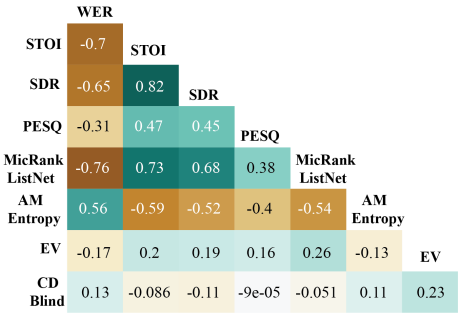

Figure 2: *Pearson correlation plot for different channel selection techniques on synthetic data.*

leads to better WER with respect to a random choice. Nevertheless, this is not the best strategy as signal-based oracles further improves the performance with STOI providing the best results. Among blind channel selection techniques, EV and AM-Entropy considerably improve over random selection and perform slightly worse than the oracles. All MicRank-based techniques are able to bring substantial gains over such previous blind selection methods. In particular, we can observe that, as expected, pair-wise and list-wise methods outperform point-wise ones which cannot account for relative performance. Notably, the best WER for RankNet and ListNet is lower than the Top-3 averaged WER of oracle WER selection, indicating that these methods are able to pick up always the best or second-best channel among the top 3. Amidst previously proposed selection methods, EV and AM-Entropy have comparable performance despite the former is remarkably less computational expensive.

Table 1: *WER on the synthetic dataset. We report both the best WER as well as the average WER on the Top-3 selected microphones.*

Ranking Method		Dev		Test	
		Best	Top-3	Best	Top-3
	Random Selection	51.7	51.5	40.9	41.1
oracle	CD-Informed [13]	45.1	47.7	36.9	38.3
	PESQ	41.9	45.8	33.1	36.4
	closest	37.0	45.1	29.9	36.1
	SDR	37.4	43.8	29.6	34.9
	STOI	36.3	44.2	29.2	35.2
	WER	32.0	39.6	24.8	30.6
baseline	CD-blind [13]	46.1	48.1	36.2	39.4
	EV [12]	39.0	44.9	31.8	35.8
	AM-Entropy [16]	41.2	45.8	31.1	35.5
MicRank	Point-wise XCE	37.3	44.1	30.4	34.6
	Point-wise MSE	36.9	43.7	30.0	34.3
	RankNet	36.5	43.4	28.8	34.1
	ListNet	**36.0**	**43.2**	**28.5**	**33.9**

In Figure 2 we report a Pearson correlation plot for a subset of selection metrics obtained on synthetic dataset test set. Interestingly, EV has rather low correlation with WER despite properly selecting favorable channels as shown in Table 1. We observed that EV fails to rank the channels with high WER. CD-Blind has the same behaviour while AM-Entropy, which is posterior based, shows much better correlation even for unfavourable channels. Again, we can notice that the proposed method is the one with highest absolute correlation value and surpasses even some oracle measures.

Finally, in Table 2 we report the performance achieved on CHiME-6 data for the most promising approaches as found on the synthetic set. Note that both EV and MicRank methods considerably improve with respect to the CHiME-6 Baseline, which benefits from "pseudo-oracle" knowledge of the speaker position and features dereverberation plus beamforming. Both RankNet and ListNet based systems improve over EV but, contrary to the synthetic dataset, are unable to outperform signal-based oracle-level performance especially on the development set. This is mainly due to the fact that CHiME-6 features a substantial amount of overlapped speech [25], while in the synthetic data only one speaker is present. This occurs in particular in the development set, which is where we observe the largest difference between signal-based oracles and the proposed method. Current selection methods, including MicRank, are unable to account for speaker identity when ranking the channels for a given utterance. This can lead to mistakenly rank the channels with respect to the interfering speaker. On the other hand, signal-based oracle measures are able to implicitly account for this because they are computed with respect to the correct speaker close-talk microphone. RankNet seems to generalize better than ListNet on CHiME-6 due to the fact that on CHiME-6 relevances are very close to each other in the training set but not in the dev and eval sets. Thus using hard labels, as in RankNet, can help boosting discriminability and generalization.

Table 2: *WER on CHiME-6 development and evaluation sets.*

Ranking Method		Dev	Eval
		Best	Best
	Random Selection	73.1	68.0
oracle	CD-Informed [13]	70.8	68.7
	PESQ	66.0	60.1
	SDR	65.2	58.9
	STOI	64.8	58.5
	WER	56.7	51.3
	CHiME-6 Baseline	69.2	60.5
baseline	CD-blind [13]	72.5	67.0
	EV [12]	68.6	59.9
MicRank	RankNet	67.4	**59.0**
	ListNet	**67.2**	59.5

5. Conclusions

In this paper we proposed MicRank, a fully neural channel selection framework for ad-hoc microphone arrays. In this framework the channel selection problem is formulated as a learning to rank (LTR) problem and a DNN is trained to rank the microphones using directly ASR errors on a training set. We explored three different LTR training strategies and validated our method on a synthetic dataset and CHiME-6. We showed that the proposed method is able to outperform previous state-of-the-art channel selection approaches which rely on signal-based or posterior-based features and is even able to surpass oracle signal-based selection on single-speaker synthetic data. Besides investigating other LTR training strategies in further work we could explore how to condition the channel selection on speaker identity in order to improve the performance on multi-party scenarios such as CHiME-6. Moreover, it would be interesting to study how much performance changes if labels are generated with a different back-end ASR than the one used in testing.

6. Acknowledgements

This research started at JSALT 2019, hosted at ETS (Montreal, Canada) and sponsored by JHU with gifts from Amazon, Facebook, Google, and Microsoft. This work has been supported by the AGEVOLA project (SIME code 2019.0227), funded by the Fondazione CARITRO. The authors would like to thank Maurizio Omologo for his valuable contribution.

7. References

[1] T. Yoshioka, Z. Chen, D. Dimitriadis, W. Hinthorn, X. Huang, A. Stolcke, and M. Zeng, "Meeting transcription using virtual microphone arrays," *ArXiv*, vol. abs/1905.02545, 2019.

[2] S. Watanabe *et al.*, "CHiME-6 Challenge: Tackling multispeaker speech recognition for unsegmented recordings," in *6th International Workshop on Speech Processing in Everyday Environments (CHiME)*, 2020.

[3] M. R. Bai, J. Ih, and J. Benesty, *Acoustic array systems: theory, implementation, and application*. John Wiley & Sons, 2013.

[4] X. Anguera, C. Wooters, and J. Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE/ACM TALSP*, vol. 15, no. 7, pp. 2011–2021, September 2007.

[5] H. Erdogan, J. R. Hershey, S. Watanabe, M. I. Mandel, and J. Le Roux, "Improved MVDR beamforming using single-channel mask prediction networks." in *Proc. of Interspeech*, 2016, pp. 1981–1985.

[6] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *Proc. of ICASSP*, 2016, pp. 196–200.

[7] C. Boeddeker, J. Heitkaemper, J. Schmalenstroeer, L. Drude, J. Heymann, and R. Haeb-Umbach, "Front-end processing for the CHiME-5 dinner party scenario," in *CHiME5 Workshop, Hyderabad, India*, 2018.

[8] X. L. Zhang, "Deep ad-hoc beamforming," *Computer Speech & Language*, vol. 68, p. 101201, 2021.

[9] Y. Luo, Z. Chen, N. Mesgarani, and T. Yoshioka, "End-to-end microphone permutation and number invariant multi-channel speech separation," in *Proc. of ICASSP*, 2020, pp. 6394–6398.

[10] J. G. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER)," in *Proc. of ASRU*, 1997.

[11] K. Kumatani, J. McDonough, J. F. Lehman, and B. Raj, "Channel selection based on multichannel cross-correlation coefficients for distant speech recognition," in *2011 Joint Workshop on Hands-free Speech Communication and Microphone Arrays*. IEEE, 2011, pp. 1–6.

[12] M. Wolf and C. Nadeu, "Channel selection measures for multi-microphone speech recognition," *Speech Communication*, vol. 57, pp. 170 – 180, 2014.

[13] C. Guerrero Flores, G. Tryfou, and M. Omologo, "Cepstral distance based channel selection for distant speech recognition," *Computer Speech and Languages*, pp. 314–332, 2018.

[14] Y. Obuchi, "Noise robust speech recognition using delta-cepstrum normalization and channel selection," *Electronics and Communications in Japan (Part II: Electronics)*, vol. 89, no. 7, pp. 9–20, 2006.

[15] M. Wölfel, "Channel selection by class separability measures for automatic transcriptions on distant microphones," in *Proc. Interspeech*, 2007.

[16] F. Xiong, J. Zhang, B. Meyer, H. Christensen, and J. Barker, "Channel selection using neural network posterior probability for speech recognition with distributed microphone arrays in everyday environments," in *Proc. CHiME Workshop on Speech Processing in Everyday Environments*, 2018, pp. 19–24.

[17] J. Guo, Y. Fan, L. Pang, L. Yang, Q. Ai, H. Zamani, C. Wu, W. B. Croft, and X. Cheng, "A deep look into neural ranking models for information retrieval," *Information Processing & Management*, 2020.

[18] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender, "Learning to rank using gradient descent," in *International Conference on Machine Learning*, 2005, pp. 89–96.

[19] Z. Cap, T. Qin, T. Y. Liu, M. F. Tsai, and H. Li, "Learning to rank: From pairwise approach to listwise approach," in *International Conference on Machine Learning*, 2007, p. 129–136. [Online]. Available: https://doi.org/10.1145/1273496.1273513

[20] V. Panayotov *et al.*, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. of ICASSP*, 2015, pp. 5206–5210.

[21] N. Furnon, R. Serizel, I. Illina, and S. Essid, "DNN-based mask estimation for distributed speech enhancement in spatially unconstrained microphone arrays," in *submitted to TASLP*, 2020.

[22] D. Diaz-Guerra, A. Miguel, and J. R. Beltran, "gpuRIR: A python library for room impulse response simulation with gpu acceleration," *Multimedia Tools and Applications*, vol. 80, no. 4, pp. 5653–5671, 2021.

[23] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *Proc. of ASRU*, 2011.

[24] V. Manohar, S. Chen, Z. Wang, Y. Fujita, S. Watanabe, and S. Khudanpur, "Acoustic modeling for overlapping speech recognition: JHU CHiME-5 challenge system," in *Proc. of ICASSP*, 2019, pp. 6665–6669.

[25] S. Cornell, M. Omologo, S. Squartini, and E. Vincent, "Detecting and counting overlapping speakers in distant speech scenarios," *Proc. Interspeech*, pp. 3107–3111, 2020.

[26] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM TALSP*, vol. 27, no. 8, pp. 1256–1266, 2019.

[27] D. S. Park, W. Chan, Y. Zhang, C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *Proc. Interspeech*, pp. 2613–2617, 2019.

[28] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM TALSP*, vol. 24, no. 11, pp. 2009–2022, 2016.

[29] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE/ACM TALSP*, vol. 14, no. 4, pp. 1462–1469, 2006.

[30] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. of ICASSP*, vol. 2, 2001, pp. 749–752.

[31] T. Taniguchi, A. S. Subramanian, X. Wang, D. Tran, Y. Fujita, and S. Watanabe, "Generalized weighted-prediction-error dereverberation with varying source priors for reverberant speech recognition," 2019, pp. 293–297.