Impact of Microphone position Measurement Error on Multi Channel Distant Speech Recognition & Intelligibility

Karan Nathwani
IIT Jammu,
India
karan.nathwani@iitjammu.ac.in

Sunil Kumar Kopparapu
TCS Research,
Tata Consultancy Services Ltd., India.
sunilkumar.kopparapu@tcs.com

Abstract

It was shown in (Raikar et al., 2020) that the measurement error in the microphone position affected the room impulse response (RIR) which in turn affected the single channel speech recognition. In this paper, we extend this to study the more complex and realistic scenario of multi channel distant speech recognition. Specifically we simulate m speakers in a given room with n microphones speaking without overlap. The n channel audio is beamformed and passed through a speech to text (`s2t`) engine. We compare the `s2t` accuracy when the microphone locations are known exactly (ground truth) with the `s2t` accuracy when there is a measurement error in the location of the microphone. We report the performance of an end-to-end `s2t` on beamformed input in terms of character error rate (CER) and and also speech intelligibility and quality in terms of STOI and PESQ respectively.

1 Introduction

The multi-path reflections (attributed by RT60) in an open enclosure is caused during hands free speech communication. Such multi-path reflections along side noise impinging on multiple microphones result in noisy reverberated speech which has a deteriorating impact on the distant speech recognition performance (Naylor and Gaubitch, 2010). Further, the quality and intelligibility of the speech in hands free communication would also deteriorate (Nathwani et al., 2017, 2016; Biswas et al., 2021). There is an urgent need for noise reduction, dereverberation and conjunction of both during communications. In this context, multi-microphone-based approaches exploiting spatial acoustic cues such as spatial diversity, inter-intensity differences and inter-time differences, receives particular interest

Accurate estimation of RT60 plays an important role in several applications like (a) sound reproduction of geometry aware room (Betlehem and Abhayapala, 2005; Tang et al., 2020; Kim et al., 2019), (b) reconstruction of the room geometry (Crocco et al., 2014; Moore et al., 2013; Yu and Kleijn, 2019), (c) robust automatic speech recognition (ASR) (Yoshioka et al., 2012; Krueger and Haeb-Umbach, 2010; Heymann et al., 2019) and (d) speech enhancement (Zhang et al., 2017; Li and Koishida, 2020; Gannot et al., 2017). In a single channel (microphone) scenario, several techniques exist to estimate the room impulse response (RIR) (Szöke et al., 2019) when the microphone position is not erroneous. Though, given the room geometry, RIR computing is non-trivial; RIR estimates require the exact location of both the source and the microphone. A comparative study for blind reverberation time estimation in single microphone scenario is explored in (Löllmann et al., 2019). A slight displacement (due to human interventions or due to routine maintenance etc. (Raikar et al., 2020; Muthukumarasamy and Donohue, 2009; Sachar et al., 2002)) in the microphone position could severely hamper the RIR estimates (Muthukumarasamy and Donohue, 2009; Sachar et al., 2002). In particular, the impact of measurement error in microphone position on speech intelligibility and quality is explored in (Raikar et al., 2020).

In practical applications, a microphone array in comparison to single microphone, grants more benefits (Nathwani et al., 2013; Stoica et al., 2002) especially due to the spatial information and associated applications like directional or arrival (DOA), location of sound source and room information (Pavlidi et al., 2013; Chen et al., 2015). However, as in single microphone case, calibration error because of displaced microphone array is not trivial to model in a real time scenarios (Sachar et al., 2004, 2002). This is, primarily because it is computationally expensive and prone to error. In (Muthuku-

Proceedings of the 18th International Conference on Natural Language Processing, pages 186–194
Silchar, India. December 16 - 19, 2021.

marasamy and Donohue, 2009), a delay and sum beamforming (DSB) technique is used to model location errors analytically, they show that the measurement error in microphone position affects the intelligibility and quality due to change in the overall RIR. DSB approach has two drawbacks, namely, it requires (a) a large number of sensors to improve the SNR and (b) it cannot adapt to varying noisy conditions.

To overcome the limitations of DSB, adaptive beamformers like capon (Stoica et al., 2002) and minimum-variance distortionless response (MVDR) beamformers have been introduced to perform joint noise and reverberation cancellation. In (Schwartz et al., 2014), a joint noise cancellation and dereverberation is illustrated in generalized side lobe canceller (GSC) framework, while in (Schwartz et al., 2015), a nested structure in the GSC framework is proposed. As opposed to beamforming, there are adaptive filtering based approaches that do not require spatial information of the speech source. In (Dietzen et al., 2017), a multi-channel linear prediction (MCLP) in Kalman Filtering domain is proposed for blind dereverberation. However, they fail to perform in the presence of noise as they focus only on reverberation.

Multi-channel beamformers are prone to measurement error due to change in microphone array position, which affects the RIR. This brings to focus, the question, *does microphone measurement error affect beamforming performance?* In particular, the impact of such displacement error, for single microphone channel, on speech intelligibility and quality has been explored in (Raikar et al., 2020). As an extension, it is of interest to explore and investigate the performance of DSB and adaptive beamformer (MVDR) for multi channel distant automatic speech recognition (ASR), intelligibility and quality. Towards this study, we simulate m speakers in a given room with n microphones speaking without overlap. The output of n channel audio is beamformed and passed through a speech to text (`s2t`) engine.

We compare the `s2t` accuracy when the microphone locations are known exactly (i.e. ground truth) with the `s2t` accuracy, when there is a measurement error in the location of the microphone location. The experimental results illustrate that the measurement error in microphone position has a significant effect on `s2t` performance. Consequently, the main contribution of this paper is the formulation of the problem to enable analysis of the effect of microphone position measurement error on distant speech recognition as well as speech intelligibility and quality. Note that in this paper, we make no attempt to introduce a new technique or algorithm to improve the distant speech recognition and intelligibility scores; rather the experimental studies reported in this paper should allow for development of new techniques in the future.

2 Problem Formulation

Let us assume a room $\mathcal{R}(L, W, H)$ of dimension $L \times W \times H$. Let there be N, $s_1, s_2, \cdots s_N$ speakers located at $\{(x_i^s, y_i^s, z_i^s)\}_{i=1}^N$ respectively and M microphones, $r_1, r_2, \cdots r_M$, located at $\{(x_j^r, y_j^r, z_j^r)\}_{j=1}^M$ respectively in the room $\mathcal{R}$. Let $u_i(t)$ be the utterance spoken by the speaker s_i at location (x_i^s, y_i^s, z_i^s) and let h_{kl} be the RIR computed for the speaker s_k and microphone r_l pair. Let o_l be the speech recorded at the microphone r_l. We can now write the output at the M microphones as $[o_1(t), o_2(t), \cdots, o_M(t)]^T =$

$$\begin{bmatrix} h_{11} & h_{21} & \cdots & h_{N1} \\ h_{12} & h_{22} & \cdots & h_{N2} \\ \vdots & \vdots & \ddots & \vdots \\ h_{M1} & h_{M2} & \cdots & h_{NM} \end{bmatrix} * \begin{bmatrix} u_1(t) \\ u_2(t) \\ \vdots \\ u_N(t) \end{bmatrix}$$

where $*$ is the convolution operator such that

$$o_l(t) = \sum_{i=1}^{N} h_{il} * u_i(t). \tag{1}$$

Note that h_{kl} is the RIR and is a function of c the speed of sound, f_s the sampling frequency of utterance, $L \times W \times H$ volume of the room, β the reverberation time, (x^s, y^s, z^s) the location of the speaker, and (x^r, y^r, z^r) the rectangular coordinates of the microphone. RIR $h_{kl} =$

$$\texttt{rir\_gen}(c, f_s, (x_l^r, y_l^r, z_l^r), (x_k^s, y_k^s, z_k^s), L, \beta) \tag{2}$$

Standard utilities to simulate h are readily available (Habets, 2006) and as mentioned in (Raikar et al., 2020) h_{kl} is prone to measurement errors in the position of the microphone, namely (x^r, y^r, z^r) as seen in (2).

Let an error $\epsilon = [\epsilon_x, \epsilon_y, \epsilon_z]$ be made in measuring the position of the l^{th} microphone r_l, so the measured location of the r_l is $r_{l\epsilon} = [x_l^r + \epsilon_x, y_l^r + \epsilon_y, z_l^r + \epsilon_z]$. Subsequently there is an error introduced in the RIR, namely, $h_{*l\epsilon} =$

$\texttt{rir\_gen}(c, f_s, r_{l\epsilon}, s_*, L, \beta, n)$. Clearly an error in measurement of the microphone $r_{l\epsilon}$ results in an error in the output speech, namely,

$$o_{l\epsilon}(t) = \sum_{i=1}^{N} h_{il\epsilon} * u_i(t). \tag{3}$$

The actual output of the l^{th} microphone, if there were no measurement errors, is (1). Also in all our experiments we assume ϵ to be Gaussian $\mathcal{N}(0, \sigma^2)$ with 0 mean and $\sigma^2 = 0.1, 0.5, 1$ as was done in (Raikar et al., 2020).

Multi channel distant speech recognition involves beamforming ($\mathcal{B}$) the multi channel speech, namely, $(o_1, o_2, \cdots, o_M)$ from M microphones, to form an equivalent of a close microphone (single channel) speech followed by ASR ($\texttt{s2t}$). For convenience let us represent this process as

$$\mathcal{T} = \texttt{s2t}(\mathcal{B}(o_1, o_2, \cdots o_l \cdots o_M)) \tag{4}$$

As can be seen an error $(\epsilon_x, \epsilon_y, \epsilon_z)$ in measuring the position of a microphone results in an error at the output of the microphone, namely, $o_{l\epsilon}$ (3), this results in an error in speech recognition output T_ϵ, namely

$$T_\epsilon = \texttt{s2t}(\mathcal{B}(o_{1\epsilon}, o_{2\epsilon}, \cdots o_{l\epsilon} \cdots o_{M\epsilon})) \tag{5}$$

In this paper, we analyze the error in the recognition (5) of speech because of an error in measurement of the location of the microphone.

3 Experimental Results and Discussion

3.1 Experimental Setup

We assume[1] a room of dimension $5 \times 5 \times 5\ m^3$ and $M = 4$ microphones and $N = 2$ speakers. Unlike in a microphone array setup we assume that the microphones can be located anywhere in the room, preferably closer to the walls and the speakers are inside the room. As an example, the room and the location of microphone and the speakers is show in Figure 1 (a) corresponding to the location of microphones and speakers shown in Table 1.

Let $u_i(t)$ be the utterance spoken speaker s_i and define $\lambda(t)$ (Figure 1 (b)) to be an arbitrary multi-valued function (the number of values depend on the number of speakers, in our case 2 corresponding to the two speakers). As seen in Figure 1 (b) s_1 (s_2) is active during the time interval when $\lambda(t) = 1 (\lambda(t) = 2)$ where $\lambda_i(t)$ is

$$\begin{aligned} &= 1 \quad \text{for} \ \ \lambda(t) = i \\ &= 0 \quad \text{for} \ \ \lambda(t) \neq i \end{aligned} \tag{6}$$

Let $\bar{u}_i(t) = u_i(t)\lambda_i(t)$ represents the utterance of speaker s_i (the duration for which speaker s_i was active). In all our experiments we construct the speech utterance as

$$U(t) = \sum_i \bar{u}_i(t) \tag{7}$$

Subsequently, we construct the multi-channel output (4 channels) as

$$o_j = \sum_i \bar{u}_i(t) * h_{ij} \ \ \text{for} \ \ \forall j = 1, 2, 3, 4 \tag{8}$$

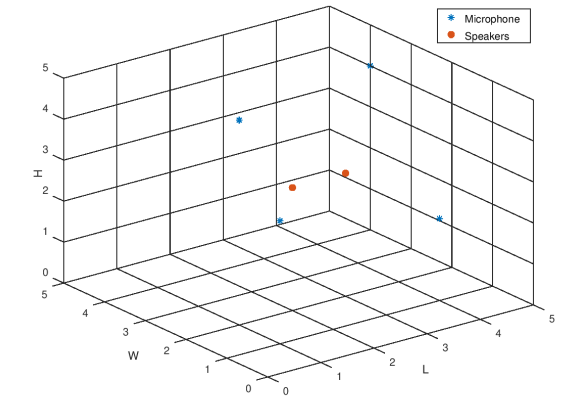
Let $\mathcal{T}_g = \texttt{s2t}(U(t))$ be the transcription of the utterance $U(t)$ which we consider as the ground truth. Now we get $\mathcal{T} = \texttt{s2t}(\mathcal{B}(o_1, o_2, o_3, o_4))$ when there is no error in the measurements of the location of the microphones. And $\mathcal{T}_\epsilon = \texttt{s2t}(\mathcal{B}(o_{1\epsilon}, o_{2\epsilon}, o_{3\epsilon}, o_{4\epsilon}))$ when there is an error (ϵ) in measurements of the location of the microphones as mentioned in Section 2. We experiment with two different beamformers, namely, (a) delay and sum (DSB) and (b) minimum variance distortionless response (MVDR) (Kumatani et al., 2015; Wei et al., 2021) (namely, $\mathcal{B} \in \{\text{DSB}, \text{MVDR}\}$). The MVDR is an adaptive beamformer which optimizes the desired speech in a given direction by filtering out interfering signal (Wei et al., 2021). This is achieved by selecting the weights of beamformer with the idea of minimizing the output power under the constraint that the target speech is unaffected. On the other hand, DSB is a fixed beamformer which is quite effective when the environment only contains uncorrelated noise between microphones (Wei et al., 2021) which is the case in our study.

We use an end-to-end transformer based state-of-the-art speech to alphabet engine for $\texttt{s2t}$ (Hugging Face Team). The $\texttt{s2t}$ inference is based on the greedy Connectionist temporal classification

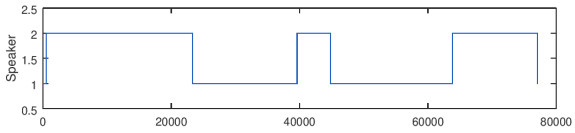
Table 1: Microphone and Speaker location used in our Experiments (Figure 1(a)).

Microphone/Speaker	Location
r_1, r_2	(1, 2, 5), (5, 4, 4)
r_3, r_4	(4, 1, 2), (1, 1, 3)
s_1, s_2	(2, 2, 3), (3, 2, 3)

[1] though not realistic, it is common, in literature to assume a cuboid room dimension



(a)



(b)

Figure 1: Experimental Setup. (a) Four microphones and two speakers (Table 1) in a room of dimension $5 \times 5 \times 5$ (we assume a cuboid so that we are consistent with (Raikar et al., 2020)), (b) A sample $\lambda(t)$ where the x-axis represents the time (expressed in samples) and y-axis can take a value of 1 or 2 depending on who is speaking. Note that at any given time only one of the speaker is speaking (no overlap).

(CTC) output without the use of a language model. We compute the character error rate $\text{CER}(\mathcal{T}_g, \mathcal{T})$ (yi Wang et al., 2003) using the state of the art speech to alphabet engine (Hugging Face Team) when (a) there is no microphone location measurement error and $\text{CER}(\mathcal{T}_g, \mathcal{T}_\epsilon)$ and (b) when there is a measurement error in the location of the microphone. We hypothesize that the CER degrades with increased ($\epsilon \equiv \sigma^2$) measurement error in microphone location. We conducted a number of experiments using the above mentioned experimental setup. We first assumed the measurement error in microphone position has a Gaussian distribution with different variances (σ^2). We study the degradation of CER, the speech intelligibility, and speech quality as a function of the microphone location measurement error σ^2.

3.2 Data

We used the popular LibreSpeech database (OpenSLR, 2021) to generate real utterances as mentioned in the experimental setup. We randomly selected two audio files u_1, u_2 (when the duration is less than 5 s we append zeros to make them of duration 5 s) from the LibreSpeech clean dataset and constructed $U(t)$ of duration 5 s (see Algorithm 1). All experiments were conducted on 100 audio samples generated in this way and all results reported (Table 2 and 3) are averaged over these samples.

Algorithm 1: Constructing $U(t)$ (7).

input : $L = 80000$ (5 seconds);
$u_1(t), u_2(t)$
output : $U(t)$ of length 5 s
$ind = round\left(\frac{L}{6} * [1:6]\right)$;
$t_1 = randi([0, ind(1)])$;
for $ind \leftarrow 1$ **to** $length(ind)$ **do**
  $t_{i+1} = randi([ind(i), ind(i+1)])$;
end
$U(t) = [u_1([1:t_1]); u_2([t_1:t_2])]; u_1([t_2:t_3]); u_2([t_3:t_4]); u_1([t_4:t_5]); u_2([t_5:t_6]); u_1([t_6; L])]$;

3.3 Experimental Validation

We evaluate the distant speech recognition performance for the experimental setup (see Figure 1). The distant speech recognition performance are presented in the form of CER averaged over 100 audio samples (Table 2). Further, the validation of CER is achieved by computing the impact of microphone position measurement error on intelligibility and

quality (Table 3).

To measure speech intelligibility, the well known (a) short time objective intelligibility (STOI) (Taghia and Martin, 2014) and (b) mutual information (MI) (Taghia et al., 2012) are used. STOI can take a value between 0 (completely unintelligible) and 1 (perfect intelligibility) and depends on the average amount of speech information available to a listener (Taal et al., 2010). The MI scores are estimated by first transforming the input signals into 15 sub-bands by using a $1/3$ octave band filter bank. Thereafter, the MI between the amplitude envelopes of the reference signal (7) and the beamformed signal with no microphone position error (bs, namely, $\mathcal{B}(o_1, o_2, o_3, o_4)$) and beamformed signal with microphone position error (bs_ϵ, namely, $\mathcal{B}(o_{1\epsilon}, o_{2\epsilon}, o_{3\epsilon}, o_{4\epsilon})$) are computed. MI is estimated per sub-band to evaluate the auditory perception (Kumatani et al., 2008). For speech quality assessment, we have used perceptual evaluation of speech quality (PESQ), signal to distortion ratio (SDR) and log-likelihood ratio (LLR). In PESQ, the speech signal is analyzed sample-by-sample after temporal alignment of corresponding excerpts of the original signal *w.r.t* to bs_ϵ and bs. In principal, PESQ models a mean opinion score (MOS) that ranges from 1 (bad) to 5 (excellent). Thereafter, we have used LLR objective measure which forms the distance measures. The LLR computes the spectral envelope difference between the original signal *w.r.t* bs_ϵ and bs (Gannot et al., 2001).

3.3.1 Speech Recognition Performance

The performance of distant speech recognition is computed in the form of CER for 100 random runs. It may be noted that lower values of CER suggest better performance. From Table 2, it can be seen that with higher measurement error (higher σ^2), the CER scores increase for both DSB and MVDR beamformers. We observe a maximum of 3% and 9% change in CER for MVDR and DSB respectively at $\sigma^2 = 1$ compared to when there is no measurement error in the microphone ($\sigma^2 = 0$). Comparing MVDR and DSB beamformers, it is observed that MVDR (adaptive beamformer) is not able to achieve the performance displayed by DSB. This lower performance can be attributed to the fact that MVDR is highly susceptible to singularity of the inverse matrix being used to calculate the weight matrix. This may result in musical noise or artifacts in the reconstruction. Figure 2 illustrates the box plot across 100 runs for DSB only (note that DSB outperforms MVDR in terms of CER). It can also be noticed that with increased σ^2 (0.5 or more), the variance and outliers in CER increase. Moreover at $\sigma^2 = 1$, the variations in CER is significantly high.

Table 2: Mean CER(%) for DSB and MVDR with varying microphone position measurement errors ($\sigma^2 = 0 \rightarrow$ no error).

$\sigma^2 \rightarrow$	0	0.01	0.5	1
DSB	26.23	27.12	27.58	28.92
MVDR	32.77	33.95	33.42	33.82

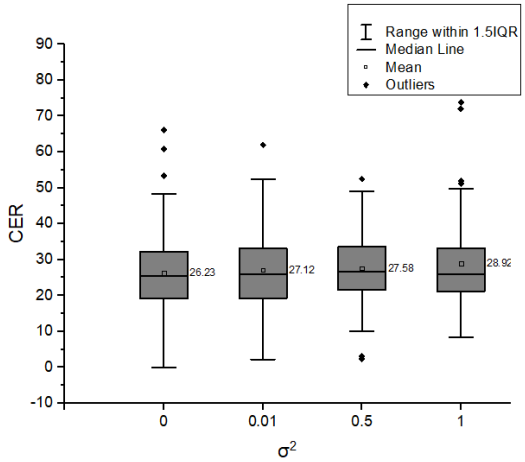


Figure 2: Variations of CER with the varying σ^2 for DSB.

Further to study how the microphone position measurement error compares with ambient or environmental noise effecting $U(t)$, we computed CER for noisy $U(t)$. We constructed $U_\epsilon(t) = U(t) + n(t)$, where $n(t)$ is an additive white Gaussian noise with different noise levels. We computed the CER on $\texttt{s2t}(U_\epsilon(t))$ with different noise levels (Figure 3). As expected, with an increase in the SNR levels (better signal strength), the CER scores decrease (better recognition) significantly. It can be also observed that with better signal (high SNRs), the outliers and variance in CER (computed over 100 runs; Figure 3) also decrease significantly, suggesting consistency in CER performance with increased signal strength. Also comparing the CER values in Table 2 and Figure 3, one can hypothesise that the measurement error in the microphone position is equivalent to an ambient noise of between 5 and 10 dB effecting the original signal. To further verify the impact of number of random runs (previously 100), we increased the random runs to 1000. It is observed that there is no difference in

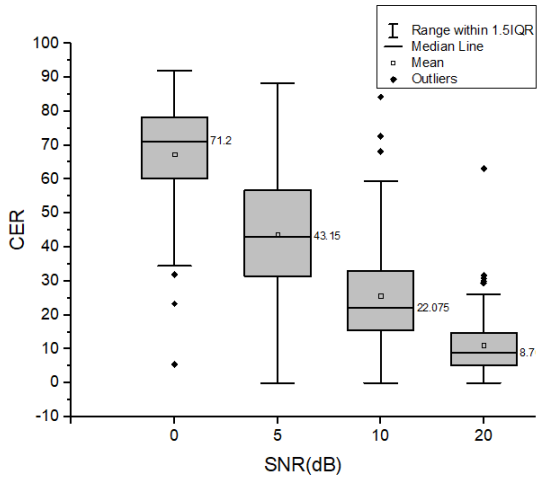


Figure 3: Variations of CER with the different noise levels.

CER between $\texttt{s2t}(U_\epsilon(t))$ for increase number of runs.

It may also be emphasized from Table 2 that although MVDR being less sensitive to position/directional errors, their performance is still not satisfactory in comparison to DSB. One of the plausible argument is that the locations of microphones are fixed once the displacement or no displacement is made in microphone position. Hence it might be possible that the MVDR beamformer would not change the weights significantly once the locations of microphones are fixed. A more extensive analysis is deferred to future work to understand the underlying reasons for such a directional behaviour of MVDR in comparison to DSB.

3.3.2 Speech Intelligibility Performance

With an aim to answer the following question, namely, (a) Does the change in the microphone position impact the speech quality and speech intelligibility? (b) Is there any relationship between CER scores and speech quality (intelligibility) scores? and (c) How does the intelligibility and quality vary with respect to the two beamformers? To address these questions, we measured the speech intelligibility and speech quality with varying microphone measurement errors. Table 3 captures the mean (variance) scores of speech intelligibility and speech quality for varying microphone position measurement errors and for the two different beamformers (namely, DSB and MVDR). It may be noted that higher the STOI and MI scores, better is the intelligibility. On the other hand, higher the SDR, PESQ and lower the LLR scores, better is the quality of the speech signal.

From Table 3, it can be observed that DSB holds better speech quality while on the other hand MVDR claims better speech intelligibility. However with increasing microphone position measurement error (increasing σ^2), both MVDR and DSB performances for intelligibility and quality degrades significantly. In particular for STOI, the maximum degradation in DSB and MVDR performances is observed to 25% and 49% respectively, when we move from no microphone position error ($\sigma^2 = 0$) to $\sigma^2 = 1$. Similarly, this degradation in quality (SDR scores) for DSB and MVDR goes to 20% and 3% respectively.

Interestingly, it is observed that the quality, intelligibility and CER scores of both the beamformers do not change significantly, while error in microphone position varies from $\sigma^2 = 0.5$ to $\sigma^2 = 1$. These results indicate that the convergence in the error in the microphone position is achieved after $\sigma^2 = 0.5$. Further, we also able to verify the claim made in (Loizou and Kim, 2010) that non-correlation between improvement in quality and improvement in intelligibility. It is clearly visible from the MVDR and DSB performances in Table 3.

Similar to Figure 3, we also address how variation in microphone position error compares with the effect of environmental noise on intelligibility and quality. To achieve this, intelligibility and quality measures are computed between $U(t)$ and $U_\epsilon(t)$. It can be seen from Table 4 that as SNR increases, the mean intelligibility and quality scores increase as expected. Although, the mean scores for STOI PESQ and SDR vary relatively slower with change in SNR, than MI and LLR scores. Further, the variance decreases for STOI and LLR but on contrary it increases for MI SDR and PESQ scores. Similar to CER scores, the effective change is observed at 10 dB SNR. This is an indication of an equivalence between measurement error in microphone position and original signal effected by ambient noise at 10 dB SNR.

4 Conclusions

In this paper, we addressed the impact of error in measuring the position of microphone position on (a) the performance of a multi-channel distant speech recognition in terms of CER and (b) quality and intelligibility of beamformed speech with two well know beamformers, namely, DSB and MVDR. Experimental analysis showed, that with increased

Table 3: Mean (variance) of speech intelligibility and quality for different microphone position measurement errors.

BF	σ^2	Speech Intelligibility		Speech Quality		
		STOI	MI	SDR	PESQ	LLR
DSB	0	0.23 (0.07)	4.72 (2.07)	-18.08 (0.89)	1.48 (0.19)	1.34 (0.29)
	0.01	0.22 (0.07)	4.66 (2.08)	-18.03 (2.18)	1.50 (0.21)	1.24 (0.27)
	0.5	0.17 (0.07)	2.55 (0.81)	-21.70 (2.95)	1.18 (0.22)	3.76 (1.75)
	1	0.17 (0.07)	2.51 (0.83)	-21.70 (3.80)	1.18 (0.22)	3.74 (1.78)
MVDR	0	0.37 (0.34)	22.14 (36.10)	-21.1 (3.75)	1.16 (0.19)	3.82 (1.81)
	0.01	0.37 (0.34)	22.05 (37.03)	-21.3 (3.88)	1.20 (0.22)	3.51 (1.68)
	0.5	0.20 (0.16)	10.55 (36.19)	-21.9 (2.86)	1.20 (0.14)	3.76 (1.74)
	1	0.19 (0.13)	8.42 (30.17)	-21.7 (3.72)	1.17 (0.21)	3.74 (1.81)

Table 4: Variation in speech intelligibility and quality with varying SNRs (in dB)

SNRs	Speech Intelligibility		Speech Quality		
	STOI	MI	SDR	PESQ	LLR
0	0.289 (0.04)	1.57 (0.39)	-21.47 (1.79)	1.25 (0.08)	5.21 (4.34)
5	0.285 (0.03)	1.72 (0.38)	-21.54 (4.17)	1.25 (0.10)	5.02 (4.18)
10	0.291 (0.01)	1.86 (0.41)	-21.40 (6.34)	1.27 (0.13)	4.83 (3.99)
20	0.299 (0.00)	2.11 (0.44)	-21.39 (8.01)	1.28 (0.16)	4.49 (3.65)

measurement error in the location of microphone the quality of mult-channel distant speech recognition deteriorates (higher CER) so does the speech intelligibility and quality, as expected. We further showed that the effect of microphone position measurement error on distant speech recognition in terms of CER is equivalent to a close microphone speech being effected by an additive environmental noise in the range of 5 to 10 dB.

References

Terence Betlehem and Thushara D Abhayapala. 2005. A modal approach to soundfield reproduction in reverberant rooms. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 289–292.

Ritujoy Biswas, Karan Nathwani, and Vinayak Abrol. 2021. Transfer learning for speech intelligibility improvement in noisy environments. *Proc. Interspeech 2021*, pages 176–180.

Xiaoyi Chen, Wenwu Wang, Yingmin Wang, Xionghu Zhong, and Atiyeh Alinaghi. 2015. Reverberant speech separation with probabilistic time–frequency masking for B-format recordings. *Speech Communication*, 68:41–54.

Marco Crocco, Andrea Trucco, Vittorio Murino, and Alessio Del Bue. 2014. Towards fully uncalibrated room reconstruction with sound. In *IEEE European Signal Processing Conference (EUSIPCO)*, pages 910–914.

T. Dietzen, S. Doclo, A. Spriet, W. Tirry, M. Moonen, and T. van Waterschoot. 2017. Low-complexity kalman filter for multi-channel linear-prediction-based blind speech dereverberation. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 284–288.

S. Gannot, D. Burshtein, and E. Weinstein. 2001. Signal enhancement using beamforming and nonstationarity with applications to speech. *IEEE Transactions on Signal Processing*, 49(8):1614–1626.

Sharon Gannot, Emmanuel Vincent, Shmulik Markovich-Golan, and Alexey Ozerov. 2017. A consolidated perspective on multimicrophone speech enhancement and source separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(4):692–730.

Emanuel AP Habets. 2006. Room impulse response generator. *Technische Universiteit Eindhoven, Tech. Rep*, 2(2.4):1.

Jahn Heymann, Lukas Drude, Reinhold Haeb-Umbach, Keisuke Kinoshita, and Tomohiro Nakatani. 2019. Joint optimization of neural network-based wpe dereverberation and acoustic model for robust online asr. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6655–6659. IEEE.

Hugging Face Team. wav2vec2. https://huggingface.co/transformers/model_doc/wav2vec2.html.

Hansung Kim, Luca Remaggi, Philip JB Jackson, and Adrian Hilton. 2019. Immersive spatial audio reproduction for vr/ar using room acoustic modelling from 360 images. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pages 120–126. IEEE.

Alexander Krueger and Reinhold Haeb-Umbach. 2010. Model-based feature enhancement for reverberant

speech recognition. *in IEEE Transactions on Audio, Speech, and Language Processing*, 18(7):1692–1707.

K. Kumatani, J. McDonough, S. Schacht, D. Klakow, P. N. Garner, and W. Li. 2008. Filter bank design based on minimization of individual aliasing terms for minimum mutual information subband adaptive beamforming. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1609–1612.

Kenichi Kumatani, Rita Singh, and Bhiksha Raj. 2015. Btk / Millennium ASR Manual. https://distantspeechrecognition.sourceforge.io/user_doc_btk10.html.

Li Li and Kazuhito Koishida. 2020. Geometrically constrained independent vector analysis for directional speech enhancement. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 846–850. IEEE.

Philipos C Loizou and Gibak Kim. 2010. Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions. *in IEEE Transactions on audio, speech, and language processing*, 19(1):47–56.

Heinrich Löllmann, Andreas Brendel, and Walter Kellermann. 2019. *Comparative study of single-channel algorithms for blind reverberation time estimation*. Universitätsbibliothek der RWTH Aachen.

Alastair H Moore, Mike Brookes, and Patrick A Naylor. 2013. Room geometry estimation from a single channel acoustic impulse response. In *IEEE European Signal Processing Conference (EUSIPCO)*, pages 1–5.

Arulkumaran Muthukumarasamy and Kevin D Donohue. 2009. Impact of microphone placement errors on speech intelligibility. In *IEEE Southeastcon*, pages 323–328.

Karan Nathwani, Morgane Daniel, Gaël Richard, Bertrand David, and Vincent Roussarie. 2016. Formant shifting for speech intelligibility improvement in car noise environment. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5375–5379.

Karan Nathwani, Harish Padaki, and Rajesh M Hegde. 2013. Multi channel reverberant speech enhancement using LP residual cepstrum. In *Asilomar Conference on Signals, Systems and Computers*, pages 555–559.

Karan Nathwani, Gaël Richard, Bertrand David, Pierre Prablanc, and Vincent Roussarie. 2017. Speech intelligibility improvement in car noise environment by voice transformation. *Speech Communication*, 91:17–27.

Patrick A Naylor and Nikolay D Gaubitch. 2010. *Speech Dereverberation*. Springer Science & Business Media.

OpenSLR. 2021. Librispeech ASR corpus. https://www.openslr.org/resources/12/dev-clean.tar.gz.

Despoina Pavlidi, Anthony Griffin, Matthieu Puigt, and Athanasios Mouchtaris. 2013. Real-time multiple sound source localization and counting using a circular microphone array. *in IEEE Transactions on Audio, Speech, and Language Processing*, 21(10):2193–2206.

Aditya Raikar, Karan Nathwani, Ashish Panda, and Sunil Kumar Kopparapu. 2020. Effect of microphone position measurement error on RIR and its impact on speech intelligibility and quality. In *Interspeech 2020*, pages 5056–5060.

Joshua M Sachar, Harvey F Silverman, and William R Patterson. 2002. Position calibration of large-aperture microphone arrays. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 2, pages 1797–1800.

Joshua M Sachar, Harvey F Silverman, and William R Patterson. 2004. Microphone position and gain calibration for a large-aperture microphone array. *IEEE Transactions on Speech and Audio Processing*, 13(1):42–52.

Ofer Schwartz, Sharon Gannot, and Emanuël AP Habets. 2014. Multi-microphone speech dereverberation and noise reduction using relative early transfer functions. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(2):240–251.

Ofer Schwartz, Sharon Gannot, and Emanuël AP Habets. 2015. Nested generalized sidelobe canceller for joint dereverberation and noise reduction. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 106–110.

P. Stoica, Zhisong Wang, and Jian Li. 2002. Robust capon beamforming. In *IEEE Asilomar Conference on Signals, Systems and Computers*, pages 876–880.

I. Szöke, M. Skácel, L. Mošner, J. Paliesek, and J. Černocký. 2019. Building and evaluation of a real room impulse response dataset. *IEEE Journal of Selected Topics in Signal Processing*, 13(4):863–876.

Cees H Taal, Richard C Hendriks, Richard Heusdens, and Jesper Jensen. 2010. A short-time objective intelligibility measure for time-frequency weighted noisy speech. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4214–4217.

J. Taghia and R. Martin. 2014. Objective intelligibility measures based on mutual information for speech subjected to speech enhancement processing. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(1):6–16.

J. Taghia, R. Martin, and R. C. Hendriks. 2012. On mutual information as a measure of speech intelligibility. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 65–68.

Zhenyu Tang, Nicholas J Bryan, Dingzeyu Li, Timothy R Langlois, and Dinesh Manocha. 2020. Scene-aware audio rendering via deep acoustic analysis. *IEEE Transactions on Visualization and Computer Graphics*, 26(5):1991–2001.

Ye yi Wang, Alex Acero, and Ciprian Chelba. 2003. Is word error rate a good indicator for spoken language understanding accuracy. In *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 57–62.

Yangjie Wei, Ke Zhang, Dan Wu, and Zhongqi Hu. 2021. Exploring conventional enhancement and separation methods for multi-speech enhancement in indoor environments. *Cognitive Computation and Systems*.

Takuya Yoshioka, Armin Sehr, Marc Delcroix, Keisuke Kinoshita, Roland Maas, Tomohiro Nakatani, and Walter Kellermann. 2012. Making machines understand us in reverberant rooms:robustness against reverberation for automatic speech recognition. *IEEE Signal Processing Magazine*, 29(6):114–126.

Wangyang Yu and W Bastiaan Kleijn. 2019. Room geometry estimation from room impulse responses using convolutional neural networks. *arXiv preprint arXiv:1904.00869*.

X. Zhang, Z. Wang, and D. Wang. 2017. A speech enhancement algorithm by iterating single-and multi-microphone processing and its application to robust ASR. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 276–280.