



Effect of Loudspeaker Emitted Speech on ASR performance

Vikram C M¹, Sanjoy Pal¹, Nidhi Mantri¹, Gopal Kumar Agrawal¹

¹Samsung R&D Institute Bangalore, India

{vikram.cm, sanjoy.pal, nidhi.matri, gopal.agrawal}@samsung.com

Abstract

Speech signal played out from the loudspeaker is referred as loudspeaker emitted speech or loud speaker speech. Most of the automatic speech recognition (ASR) systems are trained on the natural speech signals, recorded directly from the human speakers and gives higher word error rate (WER) for the loudspeaker speech. In this paper, first, we analyzed the whisper-medium ASR performance on the loudspeaker emitted speech. Five different equalizer modes, i.e., normal, pop, rock, jazz, and classic along with the distances 0m, 3m, and 5m are considered for the study. Further, based on the spectral differences between natural and loudspeaker speech, an algorithm is proposed to generate the loudspeaker quality speech from natural speech recordings. This algorithm is used to augment the Librispeech data and used to fine-tune the whisper-medium. The fine-tuned ASR on simulated loudspeaker quality speech showed significant improvement when compared to baseline system.

Index Terms: Distant speech, equalizer, loudspeaker emitted speech

1. Introduction

Loudspeaker emitted speech refers to the speech signal played out from the loudspeaker [1]. Loudspeaker emitted speech may be a replay of recorded speech or audio from loudspeaker devices such as radio or television. Most of the research works consider loudspeaker emitted speech as the replay spoofing attacks [2]. Several efforts have done to detect the replay attacks [1] and reduce the false wake-ups in voice interface devices [2]. Throughout this paper, the loudspeaker emitted speech and loudspeaker speech are used interchangeably.

The research related to spoofing attacks focused on loudspeaker emitted speech. However, accurate recognition of loudspeaker emitted speech was found to have several real-life applications. Modern smartphones have on-device automatic speech recognition (ASR) systems with streaming transcribing and translation capabilities. The on-device transcriber and translator applications in smartphones can be used in lecture/seminar halls, where the speech signal comes from loudspeaker devices. Also, the on-device ASR can be used to translate the speech in virtual meetings—most meeting applications like Microsoft Teams and Zoom support streaming translation. However, in situations like if multiple attendees belonging to different languages want to translate the meeting contents into their native language, then they can use their smartphones to translate meetings into their desired languages. In such situations, accurate recognition of loudspeaker emitted speech is necessary. However, a minimal focus is given in the literature on loudspeaker speech recognition-based applications.

1.1. Motivation

Most of the ASR systems are trained on natural speech, i.e., speech signals directly recorded from human speakers [1]. When we play the recorded speech from the loudspeaker device, the device induces several distortions in the output. In addition to the effects of loudspeaker devices, ASR needs to be robust for room reverberation, as the ASR is kept a bit far away from the loudspeaker device. Fig. 1 shows the spectra of speech signal directly recorded from the microphone and the playback audios with different equalizer effects, i.e., classic, jazz, normal, pop, and rock [3]. As shown in Fig. 1, the loudspeaker speech spectrum exhibits different spectral shapes than the natural speech. In the case of classic and pop, the low-frequency components ($< 1000Hz$) have relatively lower energy than the high-frequency components. The deviant spectral characteristics of the loudspeaker device may affect the ASR performance for loudspeaker-emitted speech.

While processing loudspeaker speech, the speech recognition system is generally kept far away from the device. Hence, the ASR performance will further degrade due to room reverberation in addition to loudspeaker device effect. To the best of our knowledge, no studies in the literature reported the impact of loudspeaker-emitted speech on ASR performance. In this paper, a detailed analysis of ASR's performance in loudspeaker-emitted speech is carried out. Based on the analysis, a spectral-modification algorithm is proposed to simulate the loudspeaker speech from the natural speech. Further, the simulated data is used to fine-tune the ASR system to improve its performance for loudspeaker speech.

The major contributions of the present work are as follows:

- The impact of loudspeaker speech on ASR performance is analyzed for
 - Five different equalizers, i.e., rock, classic, jazz, pop, and normal
 - Distances: near (0m), 3m and 5m
- Based on the analysis of loudspeaker speech, a spectral modification algorithm is proposed to simulate the loudspeaker quality speech from natural speech.
- The simulated speech signals are used to fine-tune the whisper-medium ASR to improve the performance for loudspeaker speech.

2. Database of loud speaker speech

The recording protocol used for loudspeaker speech database creation is described in Fig. 2. First, we recorded the natural speech samples directly from human speakers. These recordings are played back through loudspeakers and the out-

Table 1: *Band-wise estimated gain for different equalizer profiles*

Equalizer	Band-wise Gain (Mean±std)					
	100-150 Hz	200-400 Hz	500-1000 Hz	1100-2500 Hz	3000-6000 Hz	6000-7900 Hz
Rock	0.538±3.016	-9.282±1.949	-5.306±3.441	0.522±1.874	6.471±1.07	4.151±3.409
Classic	-5.123±2.13	-11.22±2.873	-4.017±3.326	10.521±5.743	11.498±5.304	8.249±5.303
Jazz	-3.183±2.516	-9.696±1.941	-5.093±3.548	-0.816±1.769	2.703±1.552	0.078±1.037
Pop	-2.274±4.355	-8.171±1.719	-0.018±1.785	12.698±2.191	12.05±2.543	6.535±3.181
Normal	-2.9±3.11	-4.542±2.728	2.544±3.144	14.162±3.799	13.547±4.168	10.391±4.215

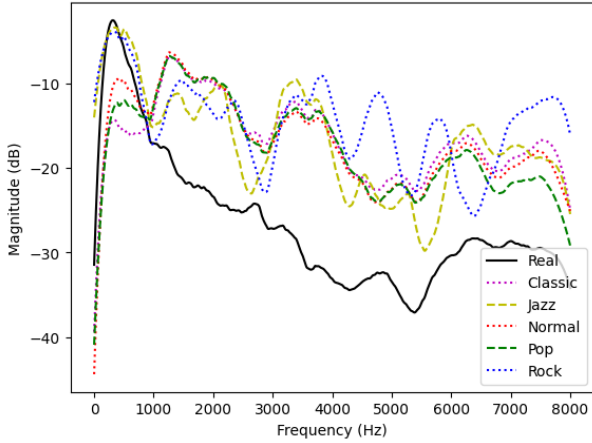


Figure 1: *Long-term averaged spectra of real (natural) and loudspeaker speech signals. The loudspeaker speech spectra are plotted for five equalizer models: classic, jazz, normal, pop, and rock. The loudspeaker speech spectrum is relatively flat when compared to real (natural) speech.*

put speech is recorded. The natural speech samples are played back with the following settings:

- Five different equalizer settings, i.e., classic, jazz, normal, pop, and rock
- Three distance conditions: 0m (near), 3m, and 5m.

The natural speech samples are recorded from professional English-speaking male and female speakers. Each speaker is instructed to speak 100 sentences in English language. Hence, in total, 200 utterances are recorded to form a natural speech database. All 200 recordings are played under classic, jazz, normal, pop, and rock equalizer conditions. For each equalizer, the playback speech are recorded by placing the recording device close (0m), 3m, and 5m from the loudspeaker. For each equalizer and distance case, we recorded 200 sentences. Finally, for five equalizer and three distance conditions, $5 \times 3 \times 200 = 3000$ utterances were recorded to create a loudspeaker speech database. We used a dedicated smartphone under voice recorder mode to record natural and loudspeaker speech samples. All the samples are recorded at a 16KHz sampling rate and saved in .wav format. In addition to these, we also recorded 3m and 5m distant speech samples directly from human speakers.

3. Analysis of spectral characteristics of loud speaker speech

As shown in Fig. 1, natural speech signals exhibit a negative spectral slope, i.e., higher energy in low-frequency bands

than high frequency. Meanwhile, loudspeaker speech exhibits a relatively flat spectral slope compared to natural speech signals. Most high-energy phonemes like vowels, exhibit low-frequency spectral dominance, whereas weak phonemes like plosive and fricatives, exhibit high-frequency spectral dominance. Equalizer boosts the high-frequency components of speech, and hence high-frequency dominant phonemes get enhanced.

We analyzed the band-wise energy difference between natural and loudspeaker speech. The corresponding loudspeaker speech is considered for each utterance in real speech, and the band-wise energy difference is computed. Generally, audio equalizers operate up to 20 kHz. Our goal is to study the impact of loudspeaker speech on ASR, where most ASRs are trained by using speech samples of 16kHz sampling rate. Hence, we analyzed spectral energy difference up to 8kHz by dividing the spectrum into six bands: (100-150 Hz), (200-400 Hz), (500-1000 Hz), (1100-2500 Hz), (3000-6000 Hz), (6000-7900 Hz).

The distribution of band-wise energy difference is mentioned in Table 1. The spectral energy difference values emphasize that the high-frequency components in loudspeaker speech and suppression in low-frequency energy.

4. Proposed augmentation method

Based on the spectral analysis, we proposed an algorithm that simulates the equalizer effects on natural speech recordings. Let $s(t)$ be the natural speech, which is passed through the 6-band audio equalizer to give $s'(t)$. That is given by

$$s'(t) = G_{b_1} s_{b_1}(t) + G_{b_2} s_{b_2}(t) + \dots + G_{b_n} s_{b_n}(t) \quad (1)$$

where $n=6$, G_{b_i} refers to the gain of bandpass filter and $s_{b_1}(t)$ corresponds to output of i^{th} bandpass filter. The gain for each band (G_{b_i}) is estimated by computing the band-wise spectral difference between the natural and playback signals. Let $S(f)$ and $S_{pb}(f)$ be the spectra of original and playback speech signals, respectively. In the i^{th} band, the gain parameter is estimated as the spectral energy difference between the original and playback speech signals. That is given by

$$G'_{b_i} = 10 * \log_{10} \sum_{f1_{b_i}}^{f2_{b_i}} (|S_{pb}(f)|^2 - |S(f)|^2) \quad (2)$$

For each equalizer case, i.e., classic, jazz, normal, pop, and rock, the band-wise gain is chosen from Table 1. For a given equalizer profile and band, the gain parameter is randomly generated using the distribution mentioned in Table. 1. Further, we added reverberation effects on the equalizer output by convolving it with the room impulse responses taken from [4].

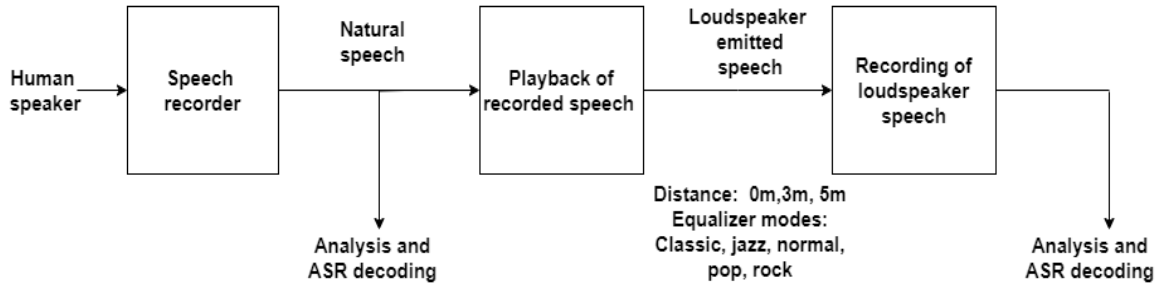


Figure 2: Overview of loudspeaker speech recording procedure.

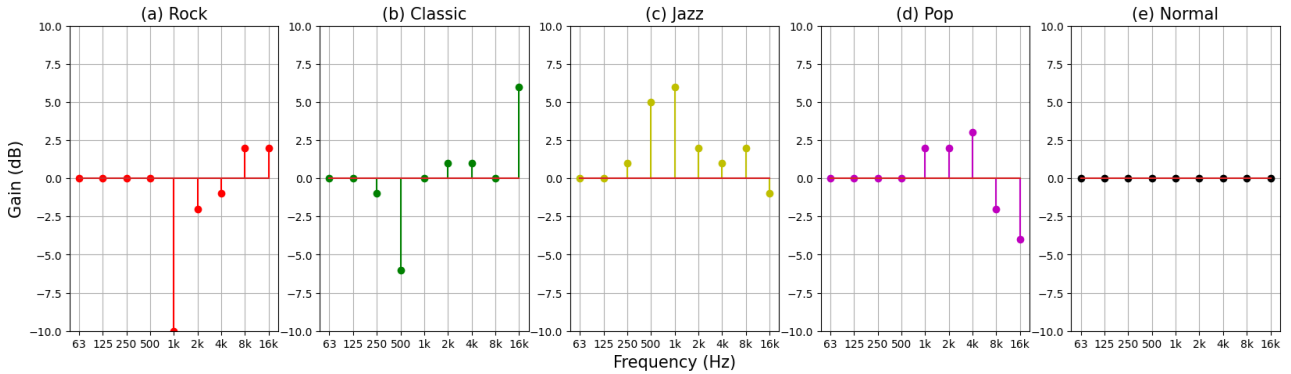


Figure 3: Different equalizer profiles.

5. Experimental Results

In loudspeaker device recordings, two effects are noticed, namely (1) equalizer in loudspeaker and (2) room reverberation effect due to the distance between a loudspeaker and recording device. We analyzed the effects of loudspeaker speech’s equalizer and room reverberation on ASR performance. In this work, whisper-medium ASR is considered for the study and the word error rate (WER) metric is used to report the ASR’s performance.

5.1. Effect of Equalizer

To analyze the equalizer effect on ASR, we considered the loudspeaker speech samples recorded by placing the recording device near the loudspeaker (0m distance samples). In the 0m case, the reverberation effect is considered to be minimal. The WER for natural and loudspeaker speech with different equalizer modes are presented in Table 2. Whisper-medium ASR is trained on natural speech recordings, which results in higher WER for loudspeaker speech than that of natural speech. This increased WER clearly indicates that the spectral changes caused by the equalizer degrade the ASR’s performance.

Table 2: Performance (WER (%)) of whisper-medium on real-clean and loudspeaker speech with different equalizer settings.

	Real-clean	Rock	Classic	Jazz	Pop	Normal
WER (%)	2.581	2.847	3.075	3.037	3.151	3.037

5.2. Effect of Distance on ASR performance

Generally, the loudspeaker device is kept far away from the ASR. Under these conditions, the loudspeaker speech is also affected by room reverberation. Fig. 4 shows the variation of WER concerning distance for different equalizers. Fig. 4 indicates that WER drastically increases with the distance for the loudspeaker speech case compared to natural or real speech. This shows that the degradation due to reverberation is more evident in loudspeaker speech cases than in real human recordings. Within the various equalizers, the classic showed higher degradation in ASR performance than others. In the classic case, the low-frequency components 250Hz-500 Hz get suppressed, which is in the range of fundamental frequency and first formant of the vowels. Since, most of the speech sounds have low-frequency spectral prominence, suppressing low-frequency components in loudspeaker speech degrades the speech quality with respect to distance. Next to classic, rock also shows highly increased WER with the distance due to the suppression of low-frequency components. Among the various equalizers, the normal mode shows minimum variation in WER concerning the distance. Because all the frequency components are equalized to 0 dB in normal, there is no relative emphasis on the high-frequency components compared to the low-frequency components.

5.3. Effect of data augmentation

The proposed loudspeaker speech simulation technique is applied on Librispeech train-clean-100 [5]. Using augmented data, we fine-tuned the whisper-medium [6] using low-rank adaptation (LoRA) technique [7, 8]. The augmented clean-dev set is used as a validation set. Augmented train-clean-100 sam-

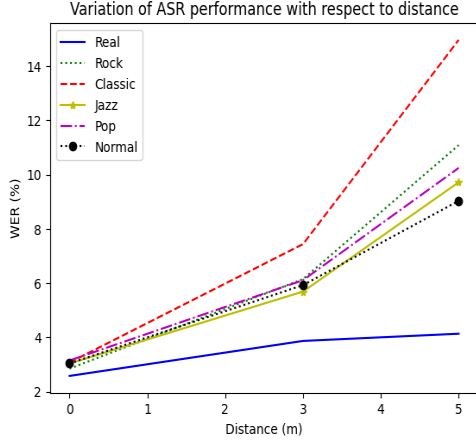


Figure 4: Effect of distance and equalizer on ASR performance

ples are used to fine-tune the whisper-medium by setting LoRA adaptation parameters: $r=32$, $\alpha=64$. The models are fine-tuned using 4 A10 GPUs with a batch size of 8 and learning rate of 0.0001. The data augmentation and LoRA adaptation is carried out for each equalizer profile which resulted in five different LoRA adapted whisper models. The performance of each adapted whisper model is evaluated by using the loudspeaker speech of corresponding mode.

The performance of fine-tuned whisper-medium is evaluated using the loudspeaker speech samples, and the results are presented in Table 3. The fine-tuned ASR shows significant improvement when compared to the baseline model. The proposed approach showed good improvement for 0m distance case. However, the fine-tuned models showed relatively lesser improvement for 3m and 5m cases than the 0m case. The reverberation effect due to distance is observed to be more in the case of loudspeaker speech than natural speech (Fig.4). The augmented reverberant speech showed improvement in ASR performance for the far-field speech in natural speaking scenarios [4]. However, the conventional way of adding room reverberation effects may not be suitable for the loudspeaker speech case. Further analysis is required to improve the ASR performance on distant loudspeaker speech cases.

Table 3: ASR performance after fine-tuning

Dist (m)	Model	Rock	Classic	Jazz	Pop	Classic
0	Baseline	2.847	3.075	3.037	3.151	3.037
	LoRA	2.012	2.509	2.989	2.89	2.42
3	Baseline	6.15	7.441	5.657	6.112	5.923
	Lora	5.89	6.541	5.54	5.989	5.215
5	Baseline	11.086	14.958	9.719	10.251	9.036
	LoRA	10.12	14.011	9.087	10.012	8.596

6. Conclusion and Future Scope

This work presented a detailed analysis of on the effect of loudspeaker speech on ASR’s performance. The results revealed that the equalizer in loudspeaker modifies the speech spectral characteristics and degrades the ASR performance. Further, the

effect of distance on loudspeaker speech analyzed and the reverberation effect is found to be more prevalent in the case of loudspeaker speech than in natural speech. Further a loudspeaker speech simulation algorithm is proposed and the augmented speech samples are used to fine-tune the whisper-medium ASR. The proposed data augmentation approach showed significant improvement in the case of the loudspeaker speech recorded near the device; however, it did not show robustness against the reverberation effect. Future work includes developing loudspeaker simulation techniques using generative techniques that can model the loudspeaker-related reverberation effects in a better way than the proposed signal processing-based approach.

7. References

- [1] T.-H. Le, P. Gilberton, and N. Q. Duong, “Discriminate natural versus loudspeaker emitted speech,” in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 501–505, IEEE, 2019.
- [2] T. Kinnunen, M. Sahidullah, H. Delgado, M. Todisco, N. Evans, J. Yamagishi, and K. A. Lee, “The asvspoof 2017 challenge: Assessing the limits of replay spoofing attack detection,” in *Interspeech 2017*, pp. 2–6, International Speech Communication Association, 2017.
- [3] M. Karjalainen, E. Piirilä, A. Järvinen, and J. Huopaniemi, “Comparison of loudspeaker equalization methods based on dsp techniques,” *Journal of the Audio Engineering Society*, vol. 47, no. 1/2, pp. 14–31, 1999.
- [4] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, “A study on data augmentation of reverberant speech for robust speech recognition,” in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 5220–5224, IEEE, 2017.
- [5] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: an asr corpus based on public domain audio books,” in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 5206–5210, IEEE, 2015.
- [6] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, “Robust speech recognition via large-scale weak supervision,” in *International conference on machine learning*, pp. 28492–28518, PMLR, 2023.
- [7] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, “Lora: Low-rank adaptation of large language models,” *arXiv preprint arXiv:2106.09685*, 2021.
- [8] Z. Song, J. Zhuo, Y. Yang, Z. Ma, S. Zhang, and X. Chen, “Lora-whisper: Parameter-efficient and extensible multilingual asr,” *arXiv preprint arXiv:2406.06619*, 2024.