VoiceAssist: Guiding Users to High-Quality Voice Recordings

Prem Seetharaman
Northwestern University
Evanston, IL, USA
prem@u.northwestern.edu

Gautham Mysore
Adobe Research
San Francisco, CA, USA
gmysore@adobe.com

Bryan Pardo
Northwestern University
Evanston, IL, USA
pardo@cs.northwestern.edu

Paris Smaragdis
University of Illinois at Urbana Champaign
Urbana, IL, USA
paris@illinois.edu

Celso Gomes
Adobe Research
Seattle, WA, USA
cegomes@adobe.com

ABSTRACT

Voice recording is a challenging task with many pitfalls due to sub-par recording environments, mistakes in recording setup, microphone quality, etc. Newcomers to voice recording often have difficulty recording their voice, leading to recordings with low sound quality. Many amateur recordings of poor quality have two key problems: too much reverberation (echo), and too much background noise (e.g. fans, electronics, street noise). We present VoiceAssist, a system that helps inexperienced users produce high quality recordings by providing real-time visual feedback on audio quality. We integrate modern audio quality measures into an interactive human-machine feedback loop, so that the audio quality can be maximized at capture-time. We demonstrate the utility of this feedback for improving the recording quality with a user study. When presented with visual feedback about recording quality, users produced recordings that were strongly preferred by third-party listeners, when compared to recordings made without this feedback.

CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**; **Graphical user interfaces**; • **Applied computing** → **Sound and music computing**.

CHI 2019, May 4–9, 2019, Glasgow, Scotland Uk

ACM ISBN 978-1-4503-5970-2/19/05…$15.00
https://doi.org/10.1145/3290605.3300539

KEYWORDS

Audio quality; active capture; feedback; interfaces; speech; voice recording; narration; creativity support tools

ACM Reference Format:
Prem Seetharaman, Gautham Mysore, Bryan Pardo, Paris Smaragdis, and Celso Gomes. 2019. VoiceAssist: Guiding Users to High-Quality Voice Recordings. In *CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019), May 4–9, 2019, Glasgow, Scotland Uk.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3290605.3300539

1 INTRODUCTION

Voice recording is central to production of audio and audio-visual media, such as podcasts, educational content, film, advertisements, video essays, and radio. Newcomers to voice recording often make mistakes when recording their voice, leading to a poor recording. High recording quality is a hallmark of successful voice-based media (e.g. radio broadcast such as NPR or popular podcasts and YouTube channels). Two key problems in many amateur recordings of poor quality are bad room acoustics (reverberation), and too much background noise (e.g. fans, electronics, street noise). LibriVox - a site where volunteers record audio books - has sections in their contributor guide on avoiding "room echo" and "background noise", illustrating how common these problems are. Denoising and dereverberation of recorded speech are entire fields of study because they are such common problems in speech recordings [11, 12].

A common workflow in voice recording is to record a “take” and then apply audio enhancement tools to the recording to improve its quality (post-processing of the recording). Denoising tools are used to reduce unwanted background noise. Dereverberation tools are used to reduce the effect of a room and echos within the room. However, the output of these tools is imperfect, with noticeable distortions and artifacts on the resultant audio [19]. Therefore, to get high-quality

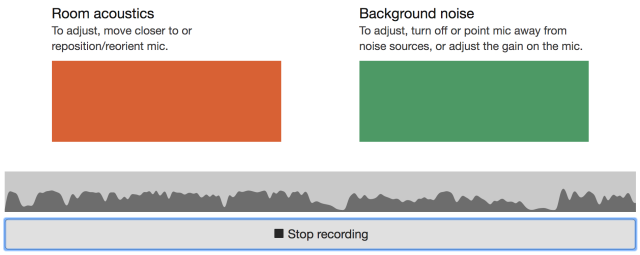

Figure 1: VoiceAssist, a system that displays visual feedback on audio quality in order to help users create high-quality voice recordings. We display two quality measures, which are indicated by a color gradient ranging from red (poor quality) to green, (high quality). In this recording, there is little background noise but the room acoustics are poor. The user can make adjustments and respond to the visual feedback to improve audio quality.

voice recordings, we think it is beneficial to simply record the audio with less noise and reverberation in the first place. When a professional recording engineer and recording studio are available, the engineer provides feedback and guidance on microphone placement and recording technique, resulting in a high-quality recording with little need for denoising or dereverberation. For many applications, however, a recording engineer and studio may not be practical or readily available. People may wish to record late at night, in their home, or without prior scheduling. The nature of the project may not allow for the expense of a recording engineer and studio.

VoiceAssist (Figure 1) helps inexperienced users produce high quality recordings by providing real-time visual feedback on the quality of the recording, like an expert recording engineer would. This is opposed to traditional recording software, which does not provide visual feedback on sound quality. Traditional recording software usually only provides volume or frequency information (e.g. Adobe Audition or Audacity). Our system helps the user find the "sweet spot" of the microphone[1] - the optimal area within the microphone's pickup pattern for recording. The feedback from our system simulates part of the expertise a recording engineer would bring to the recording session. Our work integrates audio quality measures directly into an interactive human-machine loop to maximize audio quality at capture-time. We demonstrate the utility of this feedback for creating high-quality voice recordings. We compare VoiceAssist (Figure 1) to a traditional recording interface (Figure 3). We show that users, when presented with visual feedback about audio quality, produce higher-quality voice recordings.

[1]https://www.voices.com/blog/microphone_sweet_spot

2 RELATED WORK

Active Capture [6] is a paradigm for media production that combines capture, interaction, and processing. Active capture systems use an iteration loop between the human and the machine to improve the quality of produced media. They aim to reduce the amount of effort required to produce high-quality media. These systems have been used to help people create better videos and photos by guiding them using automated quality feedback towards better framing [5, 7] or better vantage points [15]. For example, *NudgeCam* [3] helps users record interviews that follow good practices, such as ensuring the interviewee is framed correctly. VoiceAssist is an analogous system for improving sound quality.

Presentation Sensei [10] uses speech and image processing techniques to provide capture-time feedback on the way the person presents themselves: amount of eye contact with the camera, speech speed and pitch. *Narration Coach* [17] also provides feedback on a number of measures that affect speech performance quality. The feedback is focused on speech performance characteristics, such as emphasis, variety, flow, and diction. In their system, the user first records speech and then edits their recording using the feedback. The user than records the speech again, using the edited recording as a guide. The iterative process leads to a better speech performance. These systems focus on the performance quality of the text rather than the sound quality.

Speech enhancement. Users often follow a post-processing paradigm where they record the audio and then edit the recording using audio enhancement tools such as denoisers and dereverberators. However, speech enhancement tools [1, 13, 14] either often leave behind audible artifacts or only work in a limited set of cases. We aim to help users make recordings with higher sound quality that do not need post-processing.

Audio quality measures. There are several automated speech audio quality measures such as PESQ [16], PEAQ [22], and STOI [21]. We use two measures in this work - the speech transmission index (STI) [8] and the signal to noise ratio (SNR) [20]. A few methods have been developed to estimate audio quality directly from speech audio without comparing it to a reference "clean" recording [18, 23, 23, 24]. None of these methods have been incorporated into a real-time recording interface. VoiceAssist integrates two of these algorithms into a system for improving recording quality at capture-time.

3 VOICEASSIST

Our system analyzes the audio quality of a user's recording in real-time. It then presents feedback about audio quality to the user for them to improve their recording setup. We analyze two aspects of audio quality - the effect of the room on the recording (room acoustics, or reverberation) and the

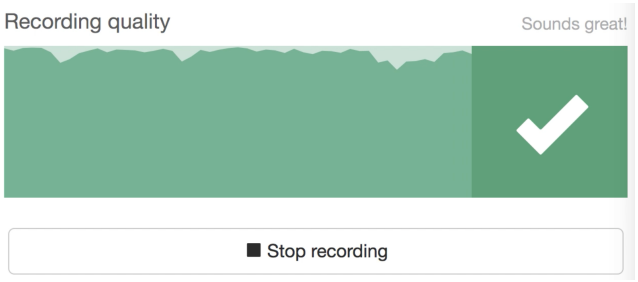

Figure 2: An early iteration of VoiceAssist, using a time-series representation. Users in a small pilot study stated in interviews that it was difficult to record a prepared text and monitor the feedback due to its active visual footprint. In response to these reports, we eliminated the time-series representation, replacing it with simple colored boxes.

amount of background noise present in the recording (signal to noise ratio). We integrate two algorithms into VoiceAssist and modify them to work in real-time. The first is a deep learning based approach which measures the effect of the room on the recording [18]. The second is a signal processing based approach that measures the amount of background noise [20].

Interface

Feedback is shown as two labeled boxes corresponding to room acoustics quality and background noise level (see Figure 1). These boxes change color on a gradient that goes from red (poor audio quality) to green (excellent audio quality). Below the two boxes is the traditional volume-based visual feedback.

Pilot studies revealed that a simple feedback mechanism (colored box) was important. An earlier version of our interface (shown in Figure 2) used a more traditional time-series representation to indicate audio quality. Users in the pilot study reported that monitoring the two objective metrics (room acoustics and background noise) when represented as a time-series while also reading a prepared text (a common recording scenario) was difficult. This was possibly because the time-series representation has a very active visual footprint, distracting from the task of recording a prepared text. When the time-series representation was substituted for simple colored boxes, users were able to keep track of the audio quality in their peripheral vision while still being able to focus on the text.

We considered other visualizations as well, such as a speed meter style visualization, reporting the raw quality numbers, and happy/sad emoji visualization. We settled on the final design because it provides intuitive situational awareness and is easily grasped with peripheral vision. Finally, we chose a red/green colormap which does not work for those with color blindness. This could be addressed by adjusting the colormap.

Room acoustics quality

When recording speech in a room, the sound waves reach the microphone directly and also indirectly via reflections off of the walls and other surfaces in the room. The effect that these reflections have on the recording have to do with the room acoustics. The reflections are called the indirect sound and the speech is called the direct sound. The quality of a recording is strongly influenced by the ratio between the direct and indirect sound. The size of and material of the surfaces in the room have an effect on this, as well as the relative positions of the speaker and the microphone. If the user is close to the microphone and is speaking inside the microphone's pick-up region (e.g. into the correct side of the microphone, rather than the side or rear of the mic), the direct sound will dominate the indirect sound, resulting in better recording quality.

The speech transmission index (STI) measures the effect a recording environment has on a recording [2]. Specifically, it measures how the recording environment warps the modulations of speech at frequencies that are important to speech perception. STI ranges between 0 and 1, where 0 indicates that the room has distorted the speech to noise and 1 indicates that the room has no effect on the speech. STIs above .75 are considered usable for public address systems, while STIs above .95 are found in professionally recorded speech. STI measurement typically requires specialized sound sources, equipment, and access to the recording environment. Our use case requires a low-effort and real-time approach to STI measurement so that the user can find the optimal recording setup quickly. For our purposes, we need to be able to measure the STI of a recording from the voice recording in real-time.

We use a method for estimating STI from speech via a convolutional neural network [18]. The network is trained with a synthetic dataset of reverberant speech with known STI values for each example in the dataset. The input to the network is 1 second of reverberant speech. The output of the network is the corresponding STI for the impulse response used to produce the reverberant speech. The trained network reliably predicts the speech transmission index from reverberant speech. It is a small network with 40000 parameters, making it suitable for real-time applications.

Background noise

Audio quality can also be affected by the amount of background noise in the recording. Not turning off background noise sources (e.g. air conditioners or fans or other appliances) or placing the mic too close or pointing towards a noise source are very common mistakes for amateurs. These mistakes result in a recording with a low signal to noise ratio

(SNR). The SNR is computed by dividing the power of the signal (speech) by the power of the noise. Professional voice recordings will generally have very high SNR.

To measure the signal to noise ratio, we first need to identify which parts of the recording are speech and which are noise. We use a state-of-the-art voice activity detector provided in the WebRTC [9] package to do this. We run the voice activity detector on the recording and segment the parts that are speech and those that are noise. We then compute the volume of the speech and the noise to estimate the SNR.

Putting it all together

The front-end records the audio and sends it to the back-end in real-time. The back-end keeps a buffer of 5 seconds of audio that is sent to it by the front-end. The back-end then analyzes the buffer whenever queried by the front-end and returns the speech transmission index, as measured using the deep learning model described in Section 3, and the signal to noise ratio, as measured by the method described in Section 3. If there is no vocal activity detected in the last second of the buffer (the most recent second recorded by the user), the back-end does not compute either quality measure, and instead reports that there is no vocal activity. The front-end then ceases visual feedback by greying out the boxes. The output of the back-end system is smoothed with median filtering to reduce the impact of outliers on user behavior. The user must maintain high sound quality over a considerable period of time.

4 USER STUDY DESIGN

The purpose of our study was to see if the visual feedback about room acoustics and background noise provided by VoiceAssist encouraged the user to make the kind of adjustments that result in higher audio quality. We had participants record themselves speaking aloud a written passage. Participants were first provided a traditional interface (comparable to Audacity, Adobe Audition) that provided visual feedback only for volume, (see Figure 3). They recorded the passage using the traditional recording interface. Participants were then either provided VoiceAssist (the test condition) or the traditional interface again (the control condition). We then had two sets of people evaluate the quality of the resulting recordings: those who made the recordings, and a set of third-party listeners.

Recruitment and setup

23 participants were recruited from nearby student and employee populations for our user study. Participants were fluent English speakers with no hearing loss or language function impairment. The control condition had 6 men and 5 women. The test condition had 6 men and 6 women. There were two locations for recording. Participants in location 1 recorded in a sound isolation booth. Participants in location 2 recorded in an empty office. All recordings were done on a Blue Yeti microphone (a popular choice for voice recording) with a cardioid pickup pattern. The participants monitored their recording over headphones. In both locations, if a participant sat down and recorded without making any adjustments to the environment or the microphone placement, they would produce a recording of low quality. Only by making adjustments to the environment could they achieve high recording quality. The adjustments suggested in Figure 1 for improving audio quality was communicated to the users in both conditions prior to the study. Users had the opportunity to improve audio quality with the same suggestions (e.g. adjust recording position, reorient microphone around, point away from noise) in both conditions.

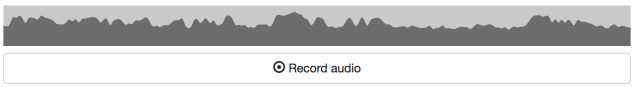

Figure 3: A traditional recording interface. The top graph indicates amplitude as measured by the microphone over time.

Procedure

Participants were provided headphones and placed before a computer, next to the microphone. Participants were asked to:

(1) Familiarize yourself with the text to be recorded.
(2) Record the text with the baseline interface.
(3) Listen to the entire recording and rate its quality.
(4) Experiment with the second interface.
(5) Record the same text again.
(6) Listen to the entire recording and rate its quality.

Participants in the control condition were given the baseline interface at both steps 2 and 4. Those in the test condition were given the baseline interface at step 2 and VoiceAssist at step 4. This allowed us to distinguish between the effect of the feedback provided from the system and the effect of simply learning to make a better recording when given time to experiment, listen to the first recording, and repeat the procedure. They were then asked to judge the recording on overall quality, room acoustics quality, and background noise quality with instructions:

"Rate the quality of the audio using the sliders below. Focus on characteristics such as the amount of reverberation (echos) or background noise in the recording, rather than the quality of the narration itself. For example, a recording with an error in the narration would still be good if the words are all easy to hear, with little reverberation or background noise."

5 USER STUDY RESULTS

Each participant made two recordings of the same text. The participants in the control condition make the recording using the same baseline interface twice, while the participants in the test condition make the recording first with the baseline interface and then with our interface. We analyze these pairs of recordings to see if there was improvement between the first and second recording. We do not compare recordings across participants to eliminate confounding variables, such as recording location. We measured the effectiveness of the feedback in improving audio quality of the resulting recording in three ways: audio quality measures (STI and SNR), third-party listener preference, and self-reported satisfaction.

Audio quality measures

We want to determine whether users actually respond to the visual feedback in VoiceAssist, and whether the user would have made the appropriate adjustments even without visual feedback. We computed the speech transmission index and the signal to noise ratio on each pair of recordings using the same system that provided visual feedback to the user. We use the Wilcoxon signed-rank test to establish significance, with $N = 23$, for each pair of recordings. For STI, we found that participants in the control condition had no significant difference between the two recordings (p-value of .15). Participants in the test condition did have a significant difference between the two recordings (p-value of .03), with an average improvement of .03 in STI. For SNR, we found that participants in the control condition had no significant difference (p-value of .15) between the two recordings. However, participants in the test condition did have a significant difference between the two recordings (p-value of .02), with an average improvement of 2.3 dB in signal to noise ratio. We find that users improve quality only when there is visual feedback.

Third-party listener preference

STI and SNR improve significantly with visual feedback on audio quality. To see whether this improvement matters to real listeners, we used the Crowdsourced Audio Quality Evaluation toolkit [4] (CAQE) to run a pairwise audio comparison experiment. We recruited 50 listeners from Mechanical Turk. Listeners passed inclusion criteria - at least 1000 tasks completed with 97% approval rating and passing a hearing test.

Listeners were trained using three speech recordings of varying quality with instructions on how they should be rated. One recording was from a professional recording booth and was considered high quality. The other two were in a noisy office and a bedroom, and were considered low quality. Listeners were further instructed to focus on the recording quality (room acoustic quality and background noise amount) rather than the performance of the text. Each listener was given 5 pairs, drawn randomly from the 23 pairs produced by our user study participants. A pair consists of the two recordings made by the same individual in either the control or test condition. The order of stimuli was random for each comparison. We collected 249 pairwise comparisons from 50 listeners.

	1st recording	2nd recording
Control	61	62
Test	30	96

Table 1: The number of listeners preferring either the first recording an individual made or the second recording that person made. In the control condition the standard interface is used for both recordings. In the test condition VoiceAssist was used for the second recording.

Table 1 shows the listener preference for recordings in the control and test conditions. For recordings made in the control condition, where both recordings made with the baseline interface, there was no preference between the recordings. For recordings made in the test condition, where the second recording was made with VoiceAssist, over three times as many listeners prefer the second recording. This indicates VoiceAssist helps users create recordings that are preferred by third-party listeners.

Self-reported evaluation

Those who performed the recordings were asked to listen to their own recording and judge the audio quality of it on three quality scales immediately after the recording was made. These quality scales were “overall quality”, “room acoustics quality”, and “background noise quality”. We found no statistical differences between the self-reported quality evaluation in the first recording and the second recording across all three quality measures for all of our pairs of recordings. This result contrasts with both the objective evaluation and the third-party evaluation, which both show the visual feedback creates a noticeable effect on the recording quality. This may indicate that inexperienced people who record audio may not always perceive recording quality differences that are detectable both by objective measures and by third-party listeners.

6 CONCLUSION

Making high-quality voice recordings is difficult. Newcomers to voice recording often make mistakes in setup and environment, leading to a poor recording. We have presented VoiceAssist, a system that provides real-time visual feedback about audio quality at capture-time. With the feedback VoiceAssist provides, we found users were able to improve audio quality at capture-time, as measured by third-party listeners and audio quality measures. VoiceAssist lowers the barrier to entry to creating high quality voice recordings.

REFERENCES

[1] Michael Berouti, Richard Schwartz, and John Makhoul. 1979. Enhancement of speech corrupted by acoustic noise. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'79.*, Vol. 4. IEEE, 208–211.

[2] JS Bradley, R Reich, and SG Norcross. 1999. A just noticeable difference in C 50 for speech. *Applied Acoustics* 58, 2 (1999), 99–108.

[3] Scott Carter, John Adcock, John Doherty, and Stacy Branham. 2010. NudgeCam: Toward targeted, higher quality media capture. In *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 615–618.

[4] Mark Cartwright, Bryan Pardo, Gautham J Mysore, and Matt Hoffman. 2016. Fast and easy crowdsourced perceptual audio evaluation. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 619–623.

[5] Ana Ramírez Chang and Marc Davis. 2005. Designing systems that direct human action. In *CHI'05 Extended Abstracts on Human Factors in Computing Systems*. ACM, 1260–1263.

[6] Marc Davis. 2003. Active capture: integrating human-computer interaction and computer vision/audition to automate media capture. In *Multimedia and Expo, 2003. ICME'03. Proceedings. 2003 International Conference on*, Vol. 2. IEEE, II–185.

[7] Jeffrey Heer, Nathaniel S Good, Ana Ramirez, Marc Davis, and Jennifer Mankoff. 2004. Presiding over accidents: system direction of human action. In *Proceedings of the SIGCHI Conference on human factors in computing systems*. ACM, 463–470.

[8] T Houtgast and H JMi Steeneken. 1973. The modulation transfer function in room acoustics as a predictor of speech intelligibility. *Acta Acustica United With Acustica* 28, 1 (1973), 66–73.

[9] Alan B Johnston and Daniel C Burnett. 2012. *WebRTC: APIs and RTCWEB protocols of the HTML5 real-time web*. Digital Codex LLC.

[10] Kazutaka Kurihara, Masataka Goto, Jun Ogata, Yosuke Matsusaka, and Takeo Igarashi. 2007. Presentation sensei: a presentation training system using speech and image processing. In *Proceedings of the 9th international conference on Multimodal interfaces*. ACM, 358–365.

[11] Philipos C Loizou. 2007. *Speech enhancement: theory and practice*. CRC press.

[12] Patrick A Naylor and Nikolay D Gaubitch. 2010. *Speech dereverberation*. Springer Science & Business Media.

[13] Santiago Pascual, Antonio Bonafonte, and Joan Serra. 2017. SEGAN: Speech enhancement generative adversarial network. *arXiv preprint arXiv:1703.09452* (2017).

[14] Cyril Plapous, Claude Marro, and Pascal Scalart. 2006. Improved signal-to-noise ratio estimation for speech enhancement. *IEEE Transactions on Audio, Speech, and Language Processing* 14, 6 (2006), 2098–2108.

[15] Yogesh Singh Rawat and Mohan S Kankanhalli. 2017. Clicksmart: A context-aware viewpoint recommendation system for mobile photography. *IEEE Transactions on Circuits and Systems for Video Technology* 27, 1 (2017), 149–158.

[16] Antony W Rix, John G Beerends, Michael P Hollier, and Andries P Hekstra. 2001. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In *Acoustics, Speech, and Signal Processing, 2001. Proceedings. (ICASSP'01)*, Vol. 2. IEEE, 749–752.

[17] Steve Rubin, Floraine Berthouzoz, Gautham J Mysore, and Maneesh Agrawala. 2015. Capture-time feedback for recording scripted narration. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology*. ACM, 191–199.

[18] Prem Seetharaman, Gautham Mysore, Paris Smaragdis, and Bryan Pardo. 2018. Blind Estimation of the Speech Transmission Index for Speech Quality Prediction. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018*.

[19] Mike Senior. 2018. How can I remove background noise from a voice recording? (Oct 2018). https://www.soundonsound.com/sound-advice/q-how-can-i-remove-background-noise-voice-recording.

[20] Jongseo Sohn, Nam Soo Kim, and Wonyong Sung. 1999. A statistical model-based voice activity detection. *IEEE signal processing letters* 6, 1 (1999), 1–3.

[21] 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 *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*. IEEE, 4214–4217.

[22] Thilo Thiede, William C Treurniet, Roland Bitto, Christian Schmidmer, Thomas Sporer, John G Beerends, and Catherine Colomes. 2000. PEAQ-The ITU standard for objective measurement of perceived audio quality. *Journal of the Audio Engineering Society* 48, 1/2 (2000), 3–29.

[23] Masashi Unoki, Kyohei Sasaki, Ryota Miyauchi, Masato Akagi, and Nam Soo Kim. 2013. Blind method of estimating speech transmission index from reverberant speech signals. In *Signal Processing Conference (EUSIPCO), 2013 Proceedings of the 21st European*. IEEE, 1–5.

[24] Xiong Xiao, Shengkui Zhao, Xionghu Zhong, Douglas L Jones, Eng Siong Chng, and Haizhou Li. 2015. Learning to Estimate Reverberation Time in Noisy and Reverberant Rooms. In *Sixteenth Annual Conference of the International Speech Communication Association*.