INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Analysis of Contextual Voice Changes in Remote Meetings

Hector A. Cordourier Maruri[1], *Sinem Aslan*[2], *Georg Stemmer*[3], *Nese Alyuz*[2], *Lama Nachman*[2]

[1]Intel Labs, Intel Corporation, Mexico
[2]Intel Labs, Intel Corporation, USA
[3]Intel Labs, Intel Corporation, Germany

hector.a.cordourier.maruri@intel.com, sinem.aslan@intel.com, georg.stemmer@intel.com, nese.alyuz.civitci@intel.com, lama.nachman@intel.com

Abstract

People participating in remote meetings in open spaces might choose to speak with a restrained voice due to concerns around privacy or disturbing others. These contextual voice changes might impact the quality of communications. To investigate how people adjust their voices in certain situations, we performed an exploratory data collection study with 41 participants in 18 simulated remote meetings. A scenario was provided to the participants to naturally trigger contextual voice changes. We collected multi-modal data from the participants including in-situ labels for the voice quality. We implemented content analysis, t-test, and linear regression to analyze the multi-modal data. Results showed that the participants primarily preferred to use soft voice over whispered voice to avoid being overheard during the meetings. Speaking softly was often sufficient to successfully conceal private conversations, while using whispered voice had only a negative impact on the intelligibility. Overall, we found that participants perceived soft voice as less pleasant to listen to than normal voice during meetings and discovered factors related to speaker demographics and meeting context that impacted the concealing behavior (soft or whispered). For our future research, we will expand to different scenarios and consider the impact of audio feedback on voice concealing.
Index Terms: Paralinguistics, voice intelligibility, voice pleasantness, voice quality, remote communications, whispering

1. Introduction

During remote meetings or calls in shared spaces (e.g., cafe, open office, etc.), lack of privacy and concern of disturbing others become important issues to address [1], which could negatively impact the experience of (1) the people sitting close by, who cannot avoid overhearing a conversation; (2) the speakers who need to adjust their voices to preserve privacy or not to disturb others; and (3) other meeting participants who could struggle to understand the adjusted voice on the other side of the line. The goal of this study is to further understand these issues during remote meetings, which will enable us to evaluate different solutions, ranging from designing new environments or devices to voice processing/transformation algorithms to enhance voice signals, in favor of comfortable and effective communications.

Most related research focuses on office-like settings and investigates issues from the perspective of involuntary eavesdroppers (i.e., the people sitting close by, who cannot avoid overhearing a conversation). It has been reported that lack of privacy in conversations is a main cause of acoustic dissatisfaction in office environments, in a level comparable with the displeasure generated by interfering noises [1, 2]. It is well known that this type of annoyance is more related with speech intelligibility than sound level, and that understanding speech is the main reason of dissatisfaction for casual people in the environment [3, 4]. Therefore, it is best to implement multiple strategies, like installing absorbing materials to reduce voice sound fields, producing broadband masking sound, and enabling large enough spaces for the expected number of workers to create a comfortable environment [5]. Research focusing on speakers' perspectives is more limited. People have some common ways to try to conceal their voice when they do not want to be overheard. For instance, the phonation pipeline in the vocal tract can be changed with a range that goes from dampening phonation (i.e., speaking softly), to canceling phonation (i.e., whispering) [6]. In that sense, speech production based on vocal effort can generally be categorized in five ascending modes: whispered, soft, neutral, loud, and shouted [7, 8, 9]. For this study, we focus on the first three as they are relevant to our considered scenarios: Office workers that perform verbal communications in remote meetings with regular laptops and headsets.

This study aims to address four major research questions: (1) How did the participants adjust their voices (from speaking softly to whispering) in remote meetings when given a scenario for not to be understood by another person sitting close to them?, (2) to what extent did the participants successfully conceal their voices from the person sitting close to them?, (3) how did the concealing behaviors impact the perceived voice quality by the other meeting participants?, (4) was there an effect of demographics and meeting context on how the participants concealed their voices?

2. Data Collection and Analysis

2.1. Participants

We designed a multi-modal data collection study with 41 participants (recruited from the US, Canada and the UK) in 18 simulated remote meetings (see Fig. 1). There were three different roles assigned to the participants: (1) voice-concealing participant, (2) eavesdropper participant, and (3) other meeting participant. Voice-concealing participant and other meeting participant(s) attended a remote meeting while the eavesdropper participant sat six feet apart from the voice-concealing participant and was involved in a different task to simulate a shared working environment. In total, 24 participants were assigned to the role of voice-concealing participant, having sometimes two concealing participants (with their respective eavesdroppers) in the same session. Additionally, we collected demographics information, i.e., gender, age, and ethnicity, from all of the participants. Although there was a balanced gender (46% male, 54% female), majority of the participants were Caucasian (67%), 54% in Gen Z (18-25 years old) and 38% Gen Y (26-44 years old) age groups.

http://dx.doi.org/10.21437/Interspeech.2021-1932

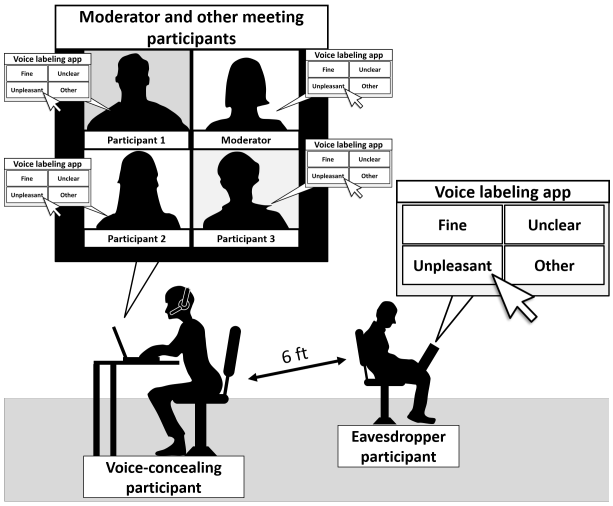

Figure 1: *Schematic diagram of the meeting setting.*

2.2. Simulated Remote Meetings

We designed remote meetings based on three distinct criteria to diversify meeting context: (1) Meeting location (café, i.e. public place, vs. home), (2) number of meeting participants (1:1 vs. group of five) and (3) participant familiarity (meeting participants know each other personally vs. not). Based on these criteria, nine of the meetings took place in a home setting whereas the other nine took place in a café. Similarly, half of the meetings were designed to be 1:1 while the other half were group meetings. The participants also knew each other at a personal level in half of these meetings. Each meeting was facilitated by a trained moderator. We implemented a set of pre- and post-meeting procedures during each meeting. As a part of the pre-meeting procedure, the moderator conducted a 15 minute training on the voice data collector app and labeling app. During the next 25 minutes, the moderator facilitated the simulated remote meeting. Each meeting had three distinct phases: (1) Informal conversations among the participants to collect baseline (normal) speech data, (2) scenarios customized to individual roles were delivered to the participants in breakout rooms to trigger adjustment of voice levels in the meeting, (3) guided conversations after receiving the scenario to collect concealed speech data. For informal conversations, the moderator asked the participants to introduce themselves as an ice-breaking activity and share the impact of COVID-19 on their lives. After the informal conversations, the moderator used the breakout rooms to explain the specific scenario to each participant (see Tab. 1). Upon providing the scenarios, the moderator initiated the guided conversations phase. For this phase, we defined three themes, i.e., travel, food, and technology, and came up with a set of questions for each theme (e.g., what is your favorite weekend trip?). These questions were used by the moderator to enable seamless conversations among the participants. During guided conversations phase, we also created a feedback loop between voice-concealing participant and eavesdropper participant: After providing the scenario, the moderator checked with the eavesdropper participant whether they could still clearly understand the conversations of the voice-concealing participant. If so, the moderator explained through the chat to the voice-concealing participant that the eavesdropper could still hear the conversation and he/she should adjust the voice level accordingly. After a minute, the moderator checked back with the eavesdropper participant and if the participant still understood the conversations, the moderator informed the voice-concealing participant, but suggested whispering instead. After this feedback, the moderator asked the eavesdropper participant one more time to record the answer but did not intervene any more. After each meeting, the moderator conducted a short debriefing interview (5-10 minutes) with each participant to further understand their experiences, attitudes, and perceptions regarding voice level adjustment. Upon completion of the interviews, the moderator guided the participants to upload their study data.

Table 1: *Scenarios customized for individual participant types*

Participant type	Summary of the scenario
Voice-concealing	Adjust your voice level so that the person sitting next to you does not understand what you're saying. Your voice level will be adjusted, so other meeting participants will still hear you fine regardless of how low you speak.
Eavesdropper	Try to understand what the person next to you is talking about. Based on what you hear, keep labeling the voice as instructed. Do not provide any direct feedback about the voice.
Other	The other participant will need to adjust voice level not to be understood by the person sitting next to them. Based on what you hear, keep labeling the voice as instructed. Do not provide any direct feedback about the voice.

Table 2: *In-situ labeling for voice quality*

Labels	Summary of the definitions
Fine	Can clearly understand what the other person says. Voice is not disturbing
Unpleasant	Can clearly understand what the other person says, but it sounds disturbing (e.g., too high/low).
Unclear	Cannot clearly understand what the other person says.
Cannot decide	Are not sure which label to assign.

2.3. In-Situ Data Labeling for Voice Quality

In the meetings, both the eavesdropper and other meeting participant(s) were asked to assess the quality of the voice in terms of intelligibility and pleasantness and label it using the app they were provided (see Tab. 2 for a description of the labels). This app produced a time-coded data file that allowed the labels to be synchronized later with the audio recordings.

2.4. Data Processing and Analysis

A total of 40 hours of audio and 21 hours of video data were recorded. The voice signals obtained from the voice concealing participants were 5 hours and 34 minutes: 52 minutes correspond to normal voice (i.e., regular amplitude speech), and 4 hours and 42 minutes to concealed voice (i.e., speech emit-

ted while the participant was giving an active effort to conceal the voice). These recordings were obtained using both the recording capability of the meeting software and our own voice recording app, which worked in the background to capture the audio from all the participants. Once the collection of all the audio, video, and labeling data were completed, the audio signals of the voice-concealing participants were manually segmented into normal or concealed voice. Using these segments, Root Mean Square (RMS) amplitude and pitch analysis were performed to find the main characteristics of concealed voice, in comparison to normal voice. The voice-concealing participants' audio signals were also manually aligned with the labels from all the participants of the same meeting. This alignment enabled us to correlate the labels with the corresponding speech signals from the voice-concealing participants. Averaging was used to integrate all the labels from the different participants. Using fine/unclear/unpleasant labels, we were able to identify how the voice-concealing participant's voice was perceived by the other participants. The segments labeled as *unclear* by the eavesdropper were considered as concealed, whereas the labels of *unclear/unpleasant* given by the other meeting participants were used as is to evaluate the negative impact on these participants. Considering the segments of normal and concealed voice, perception percentages (concealed% for eavesdropper, unclear/unpleasant% for the other meeting participants) for each voice-concealing participant were obtained. These percentages were then used in statistical analysis to understand how the perception changed by (1) considering different voice types (normal vs. soft vs. whispered), and (2) different factors related to either the speaker demographics or the meeting context. Video data was only used to help the behavioral assessment of each session.

3. Results

3.1. Discrete classes of voice emerged: Normal, soft, whispered.

We noticed all recordings from the voice concealing participants could be easily grouped into three voice types related to vocal effort, in agreement with previous research [7, 8, 9]:

- **Normal voice:** Captured while the participant spoke freely without concern of being overheard. This is the voice type with the largest amplitude and pitch, and shows clear harmonic tones in its spectrum.
- **Soft voice:** Captured when the participant is trying to avoid being overheard, this signal has lower amplitude and pitch, but still shows harmonic tones in its spectrum.
- **Whispered voice:** Also obtained when the participant is trying to conceal his/her voice, this voice type has the lowest amplitude, and no pitch, since this voicing does not produce harmonics at all.

Spectrum samples of these voice types can be seen at Figure 2, to emphasize the specific differences of these types. We used independent t-test to check for significant differences among amplitude reduction ratios of different concealing voices (soft vs. whispered). Since some participants had used both concealing types during a meeting, we considered the averages only for whispered voice to ensure sample independence and data balance for these participants. Figure 3 shows the amplitude reduction ratios for soft and whispered voice, relative to the normal speech section of each participant. Soft voice was on average 8.94 dB lower in amplitude and 21 Hz lower in pitch,

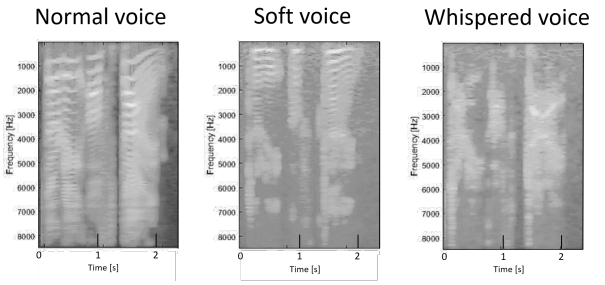

Figure 2: *Samples of the spectrums from normal, soft (lower amplitude, lower pitch with closer harmonics), and whispered voice (no harmonics), while uttering the same phrase.*

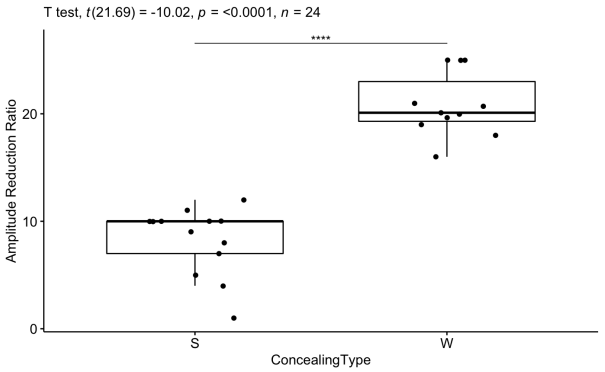

Figure 3: *Comparison of amplitude reduction ratio for different concealing behavior (S: Soft voice, W: Whispered voice).*

whereas whispered voice was 20.9 dB lower in amplitude. Little whispering fatigue over time could be observed in some participants, but not in a scale that could make voice to be re-classified.

3.2. Soft voice was good enough for concealing.

We investigated the perception differences between different voice types, comparing normal to soft and soft to whispered voice. For this analysis, we employed paired t-test by pairing perception percentages (concealed % for eavesdropper, unclear and unpleasant %s for the other meeting participants) for normal-soft or soft-whispered for each individual. The statistical analysis results are summarized in Table 3. The differences between normal and soft voice were significant for perception by eavesdropper and other participants: The perception as concealed was increased by 35% (from 0.14 for normal to 0.49 for soft voice), and the perception as unclear and unpleasant was increased by 2% and 20%, respectively. However, there was no significant difference for concealment between soft voice and whispered voice, while the unclear and unpleasant percentages were significantly increased when the speaker was whispering (by 24% and 7%, respectively).

3.3. Demographics and meeting context impacted the perception of voice quality.

To understand the impact of speaker or meeting related details on how the speaker was perceived by the eavesdropper or other participants while concealing his/her voice, we used linear regression analysis. We investigated whether there were any differences among the participants using the percentages of be-

Table 3: *Paired t-test results for comparing different voice types (M: mean, SD: standard deviation).*

	Normal vs. Soft (n=21)						**Soft vs. Whispered (n=8)**					
	Normal		**Soft**		**t(n-1)**	**p-value**	**Soft**		**Whispered**		**t(n-1)**	**p-value**
	M	**SD**	**M**	**SD**			**M**	**SD**	**M**	**SD**		
Concealment	0.14	0.23	0.49	0.40	-4.27	**<0.001**	0.06	0.14	0.34	0.28	-2.13	0.07
Unclear	0.00	0.00	0.02	0.05	-2.38	**0.027**	0.01	0.01	0.25	0.15	-4.88	**0.002**
Unpleasant	0.03	0.09	0.23	0.24	-3.99	**<0.001**	0.05	0.07	0.12	0.05	-2.56	**0.038**

Table 4: *Standardized coefficients, confidence intervals, and significance values for the linear regression model predicting perception of concealing behavior (M: Male, A: Asian, C: Caucasian, H/L: Hispanic or Latino, H: Home, G: Group, Y: Familiar).*

	Concealment (Eavesdropper)			**Unclear** (Other Participants)			**Unpleasant** (Other Participants)		
Predictors	**Beta**	**95% CI**	**p-value**	**Beta**	**95% CI**	**p-value**	**Beta**	**95% CI**	**p-value**
(Intercept)	1.08	-0.21 – 2.36	**0.001**	-0.79	-1.74 – 0.17	0.129	-0.74	-2.11 – 0.63	0.253
Age	-0.10	-0.43 – 0.23	0.519	-0.24	-0.48 – 0.01	0.057	-0.18	-0.53 – 0.17	0.296
Gender [M]	-0.77	-1.43 – -0.11	**0.025**	-0.64	-1.13 – -0.15	**0.014**	0.50	-0.20 – 1.21	0.149
Ethnicity [A]	0.15	-1.67 – 1.96	0.866	-0.31	-1.66 – 1.04	0.634	-0.74	-2.68 – 1.19	0.426
Ethnicity [C]	0.17	-1.08 – 1.42	0.772	-0.09	-1.02 – 0.84	0.841	0.69	-0.65 – 2.02	0.291
Ethnicity [H/L]	0.37	-1.23 – 1.96	0.632	0.19	-1.00 – 1.37	0.740	1.51	-0.19 – 3.21	0.078
Location [H]	-1.18	-1.87 – -0.49	**0.002**	0.50	-0.01 – 1.01	0.056	0.85	0.11 – 1.58	**0.027**
Type [G]	-0.83	-1.55 – -0.11	**0.027**	1.10	0.56 – 1.64	**0.001**	-1.36	-2.13 – -0.59	**0.002**
Familiarity [Y]	0.28	-0.38 – 0.93	0.382	0.76	0.27 – 1.25	**0.005**	0.21	-0.49 – 0.91	0.530

ing concealed (by the eavesdropper) and unclear/unpleasant (by other participants). We considered two groups of information as independent variables: (1) Speaker information: age, gender, ethnicity; (2) meeting information: location (home vs. café), type (1:1 vs. group), familiarity (yes or no). First, we checked the difference among participants when they were using their normal voice and found no significant factor for any of the dependent variables. We then investigated the percentages during concealing behavior, and discovered factors creating significant differences among participants. In the case of the eavesdropper, Gender, location, and meeting type were significant factors affecting the perception of concealment. However, in the case of meeting participants, the significant factors were mostly related to the meeting scenario: gender, type, and familiarity for "unclear" perception; location and type for "unpleasant" perception (see Tab. 4 for a summary).

3.4. Voice-concealing felt subjective, unnatural, and challenging.

The content analysis of the debriefing interviews along with our meeting observations revealed critical qualitative insights. First, there was a subjective perception of "adjusting voice level" across the participants: A few participants started whispering immediately after getting the scenario (12.5%), while others used either normal voice (8.3%) or soft voice (79.2%) in the first place. Interestingly, in the café setting (relatively loud public space), 93% of participants did not whisper, some even after they were suggested to do so by the moderator. Second, there was a negative perception towards voice-concealing: Almost half of the participants indicated that voice-concealing during the meetings felt awkward, while eight of them pointed out that it was challenging for various reasons (e.g., worried of not being understood by the other participants, not speaking this way regularly in daily life, unprofessional to speak this way in a meeting, etc.).

4. Conclusions

We conducted an exploratory study to identify how people adjusted their voices in a remote meeting scenario to conceal their speech from the eavesdroppers while maintaining the intelligibility of their communication. Amplitude and pitch analysis for different types of voices (normal vs. soft vs. whispered) indicated that these categories reside in discrete classes rather than in a continuous spectrum. Similarly, the analysis of amplitude reduction ratios showed a significant difference between soft and whispered voice. When perception percentages (fine/unclear/unpleasant) for different voice types were checked, it was clear that the speakers were able to successfully conceal (49%) their voice by using their soft voice, and switching to whispering only negatively impacted the perception of other participants (2-20% increase). Considering the concealing behavior (soft or whispered), there were factors creating significant differences within the participants, either related to demographics of the speakers or the meeting context. These factors should be considered for the development of solutions which can go from enabling appropriate environments or devices for comfortable and effective soft speaking, to voice processing algorithms to make speech more intelligible or pleasant. Future work includes using video data to assess the effect of multimodal perception in the participants. Moreover, extending the study to other voice types, depending on the environment, activity, person characteristics, or capturing device. Further experimenting can be done in systems with audio feedback, to help the speaker to conceal his/her voice more confidently.

5. Acknowledgements

We would like to thank Himanshu Bhalla, Wendy March, Caroline Foster, Srikanth Potluri, Kris D Fleming, Sai Prasad, Juan A. del Hoyo, and Marco Beltman for their support to undertake this study in the middle of a pandemic.

6. References

[1] K. Jensen and E. Arens, "Acoustical quality in office workstations, as assessed by occupant surveys," *Proceedings, Indoor Air 2005*, 01 2005.

[2] D. A. Ilter, E. Ergen, and I. Tekce, "Acoustical comfort in office buildings," in *7th Annual International Conference -ACE 2019 Architecture and Civil Engineering At Singapore*, 05 2019.

[3] W. J. Cavanaugh, W. R. Farrell, P. W. Hirtle, and B. G. Watters, "Speech privacy in buildings," *The Journal of the Acoustical Society of America*, vol. 34, no. 4, pp. 475–492, 1962. [Online]. Available: https://doi.org/10.1121/1.1918154

[4] H. Sato, M. Morimoto, S. Ohtani, Y. Hoshino, and H. Sato, "Subjective evaluation of speech privacy at consulting rooms in hospitals: Relationship between feeling evoked by overhearing speech and word intelligibility score," *Applied Acoustics*, vol. 124, pp. 38–47, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0003682X16304017

[5] P. Virjonen, J. Keränen, R. Helenius, J. Hakala, and V. Hongisto, "Speech privacy between neighboring workstations in an open office - a laboratory study," *Acta Acustica united with Acustica*, vol. 93, pp. 771–782, 09 2007.

[6] Z. Zhang, "Mechanics of human voice production and control," *The Journal of the Acoustical Society of America*, vol. 140, no. 4, pp. 2614–2635, 2016. [Online]. Available: https://doi.org/10.1121/1.4964509

[7] H. A. Patil, A. Neustein, and M. Kulshreshtha, *Signal and Acoustic Modeling for Speech and Communication Disorder*, ser. Speech Technology and Text Mining in Medicine and Health Care. De Gruyter, 2018. [Online]. Available: https://books.google.com/books?id=4-WTDwAAQBAJ

[8] H. Traunmüller and A. Eriksson, "Acoustic effects of variation in vocal effort by men, women, and children," *The Journal of the Acoustical Society of America*, vol. 107, no. 6, pp. 3438–3451, 2000. [Online]. Available: https://doi.org/10.1121/1.429414

[9] P. Zelinka, M. Sigmund, and J. Schimmel, "Impact of vocal effort variability on automatic speech recognition," *Speech Communication*, vol. 54, no. 6, pp. 732–742, 2012. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S016763931200009X