Yelling In the Hall: Using Sidetone to Address a Problem with Mobile Remote Presence Systems

Andreas Paepcke, Bianca Soto, Leila Takayama
Willow Garage
68 Willow Road
Menlo Park, CA 94025
paepcke,sotob,takayama@willowgarage.com

Frank Koenig
1944 Tasso St.
Palo Alto, CA
94301
frank@dunmovin.com

Blaise Gassend
Willow Garage
68 Willow Road
Menlo Park, CA, 94025
blaise@willowgarage.com

ABSTRACT
In our field deployments of mobile remote presence (MRP) systems in offices, we observed that remote operators of MRPs often unintentionally spoke too loudly. This disrupted their local co-workers, who happened to be within earshot of the MRP system. To address this issue, we prototyped and empirically evaluated the effect of sidetone to help operators self regulate their speaking loudness. Sidetone is the intentional, attenuated feedback of speakers' voices to their ears while they are using a telecommunication device. In a 3-level (no sidetone vs. low sidetone vs. high sidetone) within-participants pair of experiments, people interacted with a confederate through an MRP system. The first experiment involved MRP operators using headsets with boom microphones (N=20). The second experiment involved MRP operators using loudspeakers and desktop microphones (N=14). While we detected the effects of the sidetone manipulation in our audio-visual context, the effect was attenuated in comparison to earlier audio-only studies. We hypothesize that the strong visual component of our MRP system interferes with the sidetone effect. We also found that engaging in more social tasks (e.g., a getting-to-know-you activity) and more intellectually demanding tasks (e.g., a creativity exercise) influenced how loudly people spoke. This suggests that testing such sidetone effects in the typical read-aloud setting is insufficient for generalizing to more interactive, communication tasks. We conclude that MRP application support must reach beyond the time honored audio-only technologies to solve the problem of excessive speaker loudness.

ACM Classification: H.5.5 [Information Interfaces and Presentation]: Sound and Music Computing. H.5.2 [Information Interfaces and Presentation]: User Interfaces - Auditory (non-speech) feedback.

General terms: Design, Human Factors

Keywords: Telepresence, sidetone, computer mediated communication, loudness regulation, experiment

UIST'11, October 16-19, 2011, Santa Barbara, CA, USA.

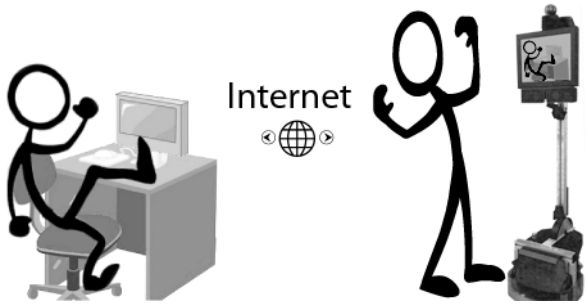

Figure 1: Remote *pilot* (left) interacts with a *local* (right) via an Internet-connected mobile remote presence (MRP) system, which is the pilot's physical 'avatar.' Problem: The pilots tend to speak too loudly, which threatens acceptance of the technology.

INTRODUCTION

Mobile remote presence systems (MRPs) have reached the cusp of commercial viability. An MRP system represents a remote operator, who uses the system to interact with others at a distance. Remote operators can drive the system around a physical environment, conversing with people at the MRP's location. A typical environment for MRP systems is a company's headquarters, in which a few workers are located elsewhere around the world (aka: hub-and-satellite teams [24]). MRPs are available at the central office for roaming around the office space. Remote employees 'inhabit' one of these MRPs via Internet connection (Figure 1). They remotely drive the machine around the premises, visiting co-workers in their offices, conversing in the hallway (e.g., Figure 2), or attending meetings.

Figure 1 depicts one example, an MRP system constructed and used at several companies for field testing. It is used for supporting communication between geographically distributed co-workers. This MRP is 1.57 meters tall (5' 2") and has a rolling base that holds a vertical rod. At the top, a pan-tilt web camera, 48.26 cm (19 inch) LCD display, microphone and loudspeakers provide live duplex video and audio. In our discussion, we call an MRP's controller the *pilot*; we call people at the MRP's location *locals*. The locals who engage in conversations with the pilots are considered to be *participants* and *side participants*, whereas locals who are near the MRP systems, but do not interact with them, are considered to be *bystanders* [4].

Of course, a work environment is just one example scenario

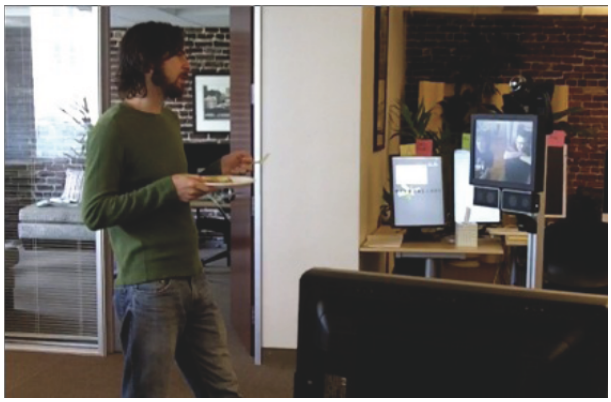

Figure 2: *Local participant* (left) in California chats with remote *pilot* (right) in Colorado after a project meeting at a field site, standing in a shared workspace near an open office door.

for telepresence. Others include remote family visits, remote medical consultations [13], or supporting independence for older adults [3].

Most mobile remote presence systems today share the basic design that is exemplified in Figure 1, including the personal roving presence (PRoP) [15], the Geminoid (android telepresence robot) [17], and BiReality [9], as well as commercially available products such as the Giraffe, Anybots QB, TiLR, and vGo. Interactions between people through these devices are quite lively, and can quickly become second nature. Considering findings in the computers-as-social-actors paradigm, [16] such ease is not surprising. While this comfortable interaction is desirable, MRP system designers encounter pitfalls that stem precisely from the resulting semi-transparency of the mediation. The root of these pitfalls is that locals transfer social norms they internalized in their lifetime of unmediated interaction onto the MRP. These norms include deeply seated and sometimes subconscious elements, like social distance [7, 2, 20, 11], politeness in yielding the way when meeting in a narrow hallway [26, 27], and turn-taking in conversation [22]. Consistent with [23], we have found that when people using MRP systems break these social norms surrounding appropriate speaking volumes in the workplace, locals react with consternation and anger. This has been observed across four of our field sites. We do not yet know the full list of norms that this technology must respect to ensure wide acceptance although we have concrete evidence for parts of this list (e.g., [11]).

Problem Statement

In this work we focus on a deal-breaking item on the list of behavioral norms that locals insist on seeing respected: Pilots must modulate the loudness of their voices through the MRP in accordance with convention. We hush our voices when in the halls of a workplace, a hospital, or in an elevator. In meetings, we use the loudness of our voice with consideration to our organizational status, the personalities of other attendees, and the intention behind our statements.

From field deployments of our MRP prototypes, we know that pilots very frequently speak inappropriately loudly, even when this behavior is pointed out to them (e.g., Figure 3).

Figure 3: *Local participant* (right) in California shushes a remote *pilot* (left) in New York, who was disrupting the nearby standing meeting by speaking too loudly at a field site.

These findings are consistent with field studies conducted by Tsui, et al. [23], who used the QB and vGo systems. At an office, where these MRP systems are used every day, we have observed co-workers slamming their office doors shut when a hallway conversations with MRP pilots become too loud. Even when local participants, who are engaged in conversation through an MRP system, often hesitate to tell pilots that they are being too loud. Possibly because of politeness norms, they seem to find it to be easier to tell pilots when they are being too quiet than when they are being too loud. In fact, what often happens is that local participants end up aligning to the loudness levels of the remote pilots so both the pilots and the locals end up speaking too loudly, exacerbating the problem of disrupting other local bystanders.

A physical volume control is available on the MRP. Locals could thus moderate the MRP system loudness themselves. Yet they rarely avail themselves of this power, even when they are annoyed by how loudly the MRP system is projecting the pilot's voice. Operating the volume control not only requires penetration into the pilot's (virtual) personal space, but even manipulation of the pilot's 'body.' Thus, locals suffer the uncouth behavior (disruptively loud conversations), which is, of course, purely a breakdown in mediation, and not an indicator of pilot intent or rudeness.

One reason why pilots speak too loudly may be that they operate in a much more complex environment than when using a more commonplace remote presence tool, the telephone. Pilots are very engaged and busy attending to views from the MRP system cameras (e.g., pointing the pan-tilt head camera) and navigating (e.g., driving the MRP system around the office space) on top of talking and listening. Another issue is that reproduction of MRP environment sound levels at the pilot side are not necessarily faithful; therefore, pilots cannot properly compare their voice loudness levels to that of locals.

Previous Attempts at Solutions

As a first measure, we tried graphical Sound Pressure Level (SPL) meters in the pilot's graphical user interface. The me-

ters communicated sound levels measured at the MRP to the pilot, similar to a stereo sound system gain monitor. Unfortunately, pilots quickly forgot to observe the meter. Such meters present additional work for pilots, who have to learn how to interpret the meters and continuously check them.

Another solution we tried was audio compressors. These digital signal processing modules automatically reduce the gain of an incoming signal processing chain without clipping, once the input level exceeds a threshold. A related technology is ambient noise compensation, which similarly adjusts gain, but in this case to compensate for fluctuating noise levels in the surroundings at the receiving end. While a well known and readily available technology, these devices must be adjusted so that they react quickly to sudden increases in loudness, while also letting go of control within a reasonable time.

In an MRP's frequently changing environment it is difficult to keep these adjustments properly tuned. For example, predicting the duration of noises is easy for humans, but not for these loudness regulators. A sneeze, for instance, is known to a human as a brief burst that does not warrant a subsequent increase in voice amplification. In contrast, the onset of a drill or printer noise is expected to last for a while. These fluctuations in appropriate thresholds, attack, and decay times compromised our success with compressors.

While more sophisticated solutions can be attempted, we decided to try a telephony feature: the *sidetone* [12]. This technology consists of letting human speakers hear an attenuated stream of their own voices. When we speak into the mouthpiece microphone of a traditional landline telephone, our voice is in fact played into our own handset receiver's earpiece at an attenuated level. The level is just enough to induce in us a subconscious tendency towards appropriate adjustment of our voice loudness. Too little feedback, and the effect is lost. Too much, and we might be self conscious or distracted. Many cell phones today do not include sidetone; this lack of auditory feedback is one of the reasons why cell phone users often yell into their phones [14]. Determining the appropriate levels of sidetone for the current study was an involved process, which is discussed in the Experiment Manipulation section.

Sidetone cannot easily be applied in the context of speaker phones, because microphone and speakers are close together. That proximity easily causes screeching acoustic feedback. The question is whether the technology can be employed for MRP pilots, who sometimes use a headset/microphone combination, other times a loudspeaker to listen to their MRP's environment.

If usable, sidetone would be beautifully simple to implement. Two advantages over more complex solutions would be that no loudness level information would need to be transmitted across the Internet, and that the human sources of sound themselves would accomplish the loudness control through mostly subconscious self regulation.

We conducted an experiment to test whether sidetone can help pilots overcome this serious problem of inadvertent yelling when using MRP systems. After reporting on related work, we describe the experimental setup and results. These sections are followed by a discussion of our findings, and concluding remarks.

RELATED WORK

The loudness of speakers and singers is impacted by two components: the noise level that surrounds the speaker, and the level and clarity at which speakers can hear their own voices (sidetone) [12]. Lane et al. [10] have shown that when confronted with changes in either surrounding noise or sidetone level, speakers will produce their compensatory reaction at one half the size of those changes. For example, when sidetone level is quadrupled, speakers will cut their loudness in half. That is, they determined that the slope of the sidetone reaction on a decibel/decibel scale is -0.5.

Siegel and Pick [19] showed that the loudness moderating effect of sidetone interacts with surrounding noise levels. When surrounding noise levels are higher, the effect of varying sidetone levels is amplified. This finding did not bode well for our experiment. Our remote pilots were most annoying in quiet environments, where according to Siegel and Pick's study the effect of sidetone is least powerful (albeit significant).

Goodman and Johnston describe a variation of sidetone injection in which sidetone level is varied in response to the loudness of the speaker [5]. The authors did not report upon any experimental results.

In human-computer interaction, the use of sidetone has been explored in the domain of driver user interfaces. A controlled experiment found that the addition of sidetones increased conversational engagement (e.g., verbosity of spoken responses), but also increased the cognitive load that drivers experienced [21]. Sidetone can be a mixed blessing.

Sidetone is also used at concerts to help singers stay in tune. In that context, loudspeakers are sometimes used instead of headphones, and special care must be taken to avoid acoustic feedback [8]. To our knowledge, no formal studies have examined the efficacy of sidetone for speaker loudness self control when visual as well as aural communication is involved. In contrast to earlier studies, our exploration also included engaging and challenging creativity questions, which varied the tasks' intellectual demands and carried the potential of triggering raised speaking loudness in the participants.

Despite the widespread use of sidetone for telephones and live musical performances, it is rarely used at all in computer-mediated communication contexts. Norman raised this issue six years ago [14] and it has yet to hit the mainstream mobile phone or videoconferencing systems on the market. Providing auditory feedback in the form of sidetones improves the user experience for landline telephones [14] and holds the promise for being useful in these new computer-mediated communication contexts such as mobile remote presence.

The current work differs from previous uses of side tone in two major ways: (1) the addition of the video channel and (2) the addition of a mobile system that can wander around very diverse acoustic environments.

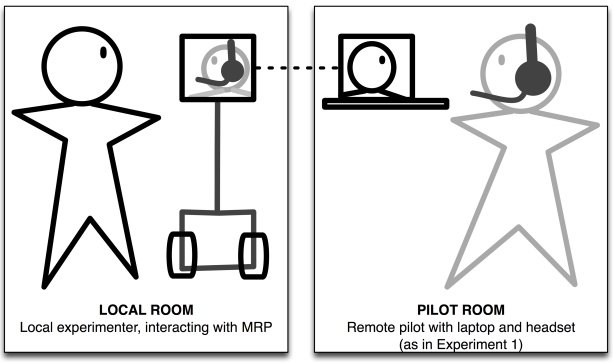

Figure 4: Local vs. pilot rooms.

	Example Participant 1		*Example Participant 2*	
Round1	*Low sidetone*	*wordlist*	*No sidetone*	*wordlist*
		warmup		*warmup*
		creativity		*creativity*
Round 2	*High sidetone*	*wordlist*	*Low sidetone*	*wordlist*
		warmup		*warmup*
		creativity		*creativity*
Round 3	*No sidetone*	*wordlist*	*High sidetone*	*wordlist*
		warmup		*warmup*
		creativity		*creativity*

Table 1: Summary of experimental procedure for two sample participants.

EXPERIMENT

We conducted two experiments in which participants interacted remotely with the experimenter. Sidetone level was a within-participants independent variable in both studies. In Experiment 1, participants used a headset with a boom microphone. In Experiment 2, participants used a loudspeaker and desktop microphone.

Experiment Manipulation

In both experiments we mixed each of three sidetone levels into the audio that participants heard when they were speaking. We call these levels S_{no}, S_{low}, and S_{high}. These values were measured in decibels comparing the pressure levels of the generated sidetone with those of the participant's voice. For example, a headset condition sidetone level of $-3dB$ resulted from attenuating a participant's voice by $3dB$ and mixing that signal into the participant's headset while they spoke.

S_{no} is zero; the absolute values of S_{low} and S_{high} differed between the loudspeaker and the headphone conditions (see below in the respective subsections). All participants experienced all three sidetone levels one after the other, albeit in random order.

Participants

The first experiment, which used headsets, included 20 adult volunteer participants (9 female, 11 male), ranging from 20 to 69 years of age, *M*=39.7, *SE*=3.7. The second experiment, which used loudspeakers, included 14 adult volunteer participants (8 female, 6 male), ranging from 24 to 68 years of age, *M*=32.1, *SE*=3.1.

Experiment Protocol

Study participants played the role of the MRP pilot, using a laptop to communicate through the MRP system. The experimenter sat in a separate room, playing the role of a local, interacting with the pilot through the MRP system. See Figure 4. Table 1 shows a summary of this experimental procedure.

To simulate a quiet office environment, we mixed an approximately $45dB$ background typing noise into the signal that participants heard. This background noise varied in volume over time.

For each sidetone level, participants went through a round of three exercises—nine exercises in all. Participants filled out a questionnaire about their experience between each of the three rounds.

For the first exercise, the experimenter asked the participant to read a list of words from the Modified Rhyme Test (MRT) collection [1] (*wordlist* exercise). These were semi-scripted interactions in that the experimenter read instructions and worked from a script.

For the second exercise, the experimenter asked the participant to answer five questions that were designed to free participants from excessive attention to their environment. Two sample questions were, "If you were an animal, what would it be, and why?" and, "When did you last write a handwritten letter?" (*warmup* exercise). These questions were designed to be more conversational in nature; thus, they were modified from a set of coffee table cards that are typically used to start conversations.

For the final exercise of each round, the experimenter named an object, and asked the participant to list as many non-standard uses for this object within 60 seconds. We repeated this question for the objects "shoe" "key" and "bed sheet" (*creativity* exercise). This approach is adapted from the Alternate Uses Creativity Task [6].

The order of exercises within each round was fixed, and no variables other than sidetone were modified for any one participant. These three tasks were used to cover a variety of interaction scenarios. The word list was reading aloud without time constraints. The warm up was revealing information about oneself and was very personal in nature. The creativity task was about generating divergent thinking ideas and had hard time constraints.

From the perspective of a participant in this study, he or she would sit at a desk in a relatively quiet office space. An open laptop sat upon the desk, displaying the user interface for controlling the MRP system, which included a live video feed to the Local Room. The participant would then fill out a study agreement form. The experimenter explained the procedures for the study, then left the Pilot Room to sit in the Local Room. The participant then saw and heard the experimenter through the laptop. The participant did each of the three tasks (word list, warm up, and creativity), filled out a paper questionnaire, and repeated those two steps two

more times. Upon completing the study, the experimenter debriefed the participant and answered the participant's questions.

As we will describe, we recorded all speech. From the recordings, we extracted the portions where participants were speaking—as opposed to thinking about answers or listening to the experimenter. On these extracts, we then computationally compared each participant's speech pressure levels in sidetone conditions with their levels without sidetone.

The following three subsections describe the acoustics and psychoacoustic terms that are necessary for understanding the experiment set-ups, the headset experiment conditions, and the loudspeaker conditions.

Defining Terms

The ultimate goal of this study is to decrease inadvertent loudness of MRP pilots, who are projecting their voices too loudly through the MRP systems. "Loudness" is a term that applies to the subjective psychological experience of sound. In order to decrease the experience of MRP pilots being too loud, the current experiments explored ways to get MRP pilots to self-regulate their loudness, but we chose to use an objective measure rather than a subjective one.

We needed to measure how much each participant's sound pressure level changed across the experiment conditions so we chose to use the objective measure of "sound pressure level." This is a logarithmic measure of sound pressure (averaged over time, using root mean square) relative to some reference level (typically 20μPa in air, a standard set by the American National Standards Institute). Sound pressure levels are often measured in decibels ($dB(SPL)$). dB is an abbreviated term for $dB(SPL)$. L is an abbreviated term for "sound level." Frequency weighting is a process of scaling loudness measurements to match the human perception of sound, which is frequency dependent.

When you record the sound pressure levels in a room, you get a combination of many different sources, including atmospheric pressure, sidetones (in the case of Experiment 2), and the voice of the participant. To isolate the sound pressure levels of only the sidetones or only the voice of the participant, we had to measured the sources of extraneous sound pressure. The following sections explain the experiment set ups that we created to (1) create the sidetones, (2) measure the sound pressure levels of the sidetones, and (3) to record and calculate the participant voice sound pressure levels, isolating it from the other sources of sound pressure.

Experiment 1 Set Up: Headset

Figure 5 shows a block diagram of the headset condition. The participant's voice was captured in two digital recorders. Recorder 1 (R_1) was placed about 1m away from the participant, so that limited movement of the participant's head would not significantly impact the captured voice loudness. We used the R_1 recordings for our analysis.

Additionally, we captured the signal from the boom microphone in Recorder 2 (R_2) so that we could monitor the source from which sidetone was generated. The switch allowed us to choose between sidetone levels (S), which for the headset experiment were set to $S_{low} = -3db$ of the participant's voice level, and $S_{high} = 1.7db$. We chose these levels such that the low setting was just barely audible in the headset, while the high setting was clearly present. Sidetone was mixed with the voice from the experimenter seated in the locals room.

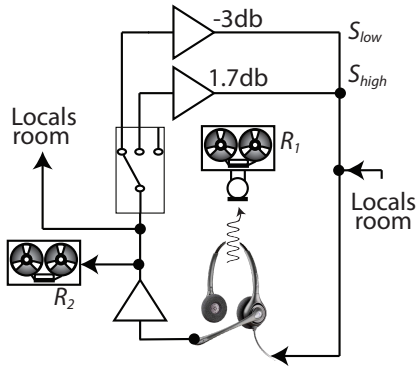

Figure 5: Experiment 1: Block diagram of the headset experiment set up.

Headset Data Preparation Procedure In order to analyze the impact of sidetone on participant voice levels we needed to extract from R_1 the relative loudness of each participant when sidetone was administered, versus when they heard no sidetone. We generated these ratios as follows.

From R_1's *.wav* file recordings we extracted the amplitudes of R_1's stereo channels. We computed the root mean squares *rms* of each channel, and averaged the results to obtain rms_{no}, rms_{low}, and rms_{high} for each of the sidetone levels (no sidetone, low sidetone, and high sidetone). We then computed the sound pressure level ratios (L) in decibels without frequency weighting:

$$L_{high} = 20 * log_{10}(rms_{high}/rms_{no}) \quad (1)$$
$$L_{low} = 20 * log_{10}(rms_{low}/rms_{no})$$

Note that sidetone generation (Figure 5) involved microphones, the switch, amplification, and so on, which are all components that impact audio signals as they pass through. The quantification of sidetone levels that we selected by ear thus needed to be quantified via a calibration process, which we describe next.

Headset Sidetone Level Quantification Figure 6 shows the headset quantification procedure's two stages. First (Figure 6*a*), one of the authors read aloud one of the word lists while wearing the headset. The resulting audio signal was passed through the entire signal chain, with the switch set to generate no sidetone. The output was recorded via cable into R_1.

Next (Figure 6*b*), the recording was transferred to R_2 and replayed three times through the signal chain, once without sidetone, once with the switch set to generate S_{high}, and once to generate S_{low}. Each output was again recorded via cable into R_1. We then computed the *rms* for each recording, and used Equation 1 to quantify the sidetone levels. The results

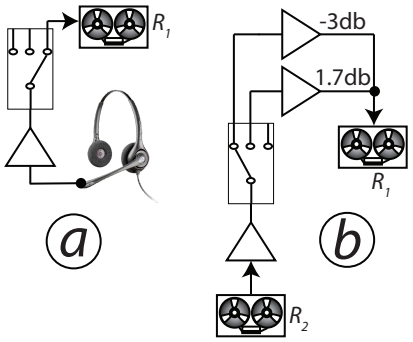

Figure 6: Experiment 1: Block diagram of the headset sidetone quantification procedure.

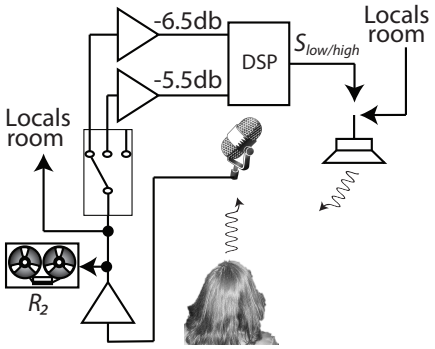

Figure 7: Experiment 2: Block diagram of the loudspeaker experiment set up.

were $S_{low} = -3db$, and $S_{high} = 1.7db$, a difference of approximately $5dB$.

Experiment 2 Set Up: Loudspeaker

Figure 7 shows our arrangement of the loudspeaker setting. The participant was again seated about 1m from the microphone, which in this setup was not R_1's microphone, but a table microphone. In contrast to the headset experiment, the sidetone signal passed through a digital signal processing (DSP) stage consisting of notch filters that eliminated some frequencies that caused audio feedback ringing. The loudspeaker was placed about 2m away from the participant, to the right and below a table on which the microphone rested. The table surface thus prevented point-to-point travel of the audio signal between the sidetone source and the microphone.

This separation served two purposes. First, it helped avoid acoustic feedback problems, and second, the audio signal's multi-path travel ensured that the participant's voice and their sidetone arrived at the microphone incoherently. This incoherence assured that we could use power addition in our signal analysis.

Loudspeaker Data Preparation Procedure The procedure for extracting the treatment effect in the loudspeaker condition was more complex than in the headset case. Note that the microphone in Figure 7 was exposed to both the participant's voice, and any sidetone signal that managed to travel via echo paths from the loudspeaker to the microphone; (the direct path was blocked by a table). This sidetone component needed to be separated from the participant's voice.

For clarity, we describe the math using low sidetone. High sidetone is analogous. Recall that S_{low} is a decibel value comparing sidetone with the signal from which the sidetone was generated. Let L_{voice} be the unknown participant voice sound pressure level in decibel relative to the standard 20μPa. L_Σ is the combination of the (incoherent) participant's voice and the sidetone. At the microphone, the (incoherent) participant's voice and sidetone combine as follows:

$$L_\Sigma = 10 * log_{10}(10^{(L_{voice}/10)} + 10^{(S_{low}/10)})$$

$$Therefore:$$

$$L_{voice} = 10 * log_{10}(10^{(L_\Sigma/10)} - 10^{(S_{low}/10)}) \quad (2)$$

We computed L_Σ as usual:

$$L_\Sigma = 20 * log_{10}(rms_{low}/rms_{no})$$

We used Equation 2 for our analysis. It provides the sound pressure level of the participant's speaking voice, excluding the sound pressure level of the sidetone manipulation.

Loudspeaker Sidetone Level Quantification Sidetone levels S_{low} and S_{high} were more constrained in the loudspeaker experiment than for the headset setup, because audio feedback leads to whistling when sidetone is too high. We adjusted S_{high} to be the highest possible in the room.

Figure 8 shows our procedure, which is analogous to the headset quantification. We recorded a reference wordlist recitation into R_2 without sidetone (not illustrated in the Figure). We then injected the resulting recording electronically into the signal chain from R_2. The recording was played three times, once without sidetone, and once with each of S_{low} and S_{high}. No acoustic signal other than the room's natural background noise were present. The desk microphone recorded the sidetone into R_1. We then again computed sidetone levels from the files' *rms* values. The results were $S_{low} = -6.5dB$ and $S_{high} = -5.5dB$.

Note that while these values lead to accurate ratio calculations for our eventual data analysis, they are not good indicators of sidetone levels that participants actually heard at their location. In an effort to provide a better intuition for their experience we performed acoustic measurements in parallel to the above described procedure.

While playing just sidetone through the loudspeaker, we used a commercial grade sound pressure level meter to measure pressure levels where the participants' heads would be during the actual experiments. Those levels came to $43dBA$ at low sidetone, and $45dBA$ at high sidetone setting. The room's noise floor was $37dBA$ at the time. Sidetone was thus clearly audible to participants.

Equipment and Implementation

The switch and all amplifiers in Figure 5 and Figure 7 were implemented with a Tascam M-164UF (*Mixer1*) and a Behringer Eurorack M802A *Mixer2* mixing board. These devices were daisy chained for the experiment. The pilot room microphone were connected to *Mixer1*. That signal was routed

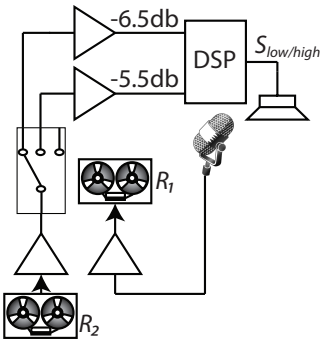

Figure 8: Experiment 2: Block diagram of the loudspeaker sidetone quantification procedure.

to the locals room for communication with the experimenter.

Depending on the sidetone condition, the microphone signal was also routed channel 1, 2, or 3 of *Mixer2*. These channels were pre-adjusted to generate S_{no}, S_{low}, or S_{high}, respectively. When changing condition the experimenter would visit the pilot room and replug the patch cable that connected *Mixer1* to *Mixer2*. This reduced the chance of making mistakes in adjusting a single potentiometer.

In addition to the microphone signal from *Mixer1*, *Mixer2* received audio from the locals room (i.e. the experimenter's voice and the typing loop). The resulting mix of sidetone and the incoming locals signal drove the headset or loudspeaker in the participants' (pilot) room.

The pilot headset/boom microphone was an inexpensive Plantronics set with boom electret microphone. The desktop microphone used for the loudspeaker experiment was an AKG 414. R_1 was a Zoom model H4 digital recorder, R_2 was a Zoom model H4n recorder.

A Behringer 31 band Ultragraph FBQ-Pro 3102 graphic equalizer and Behringer Feedback Destroyer Pro provided acoustic feedback control in the loudspeaker experiment. These devices correspond to the DSP box in Figure 7. The sound pressure level meter used during the loudspeaker sidetone quantification procedure was a class 2 CEM model DT-8851.

RESULTS

Experiment 1: Headset

We ran repeated measures ANOVA, using the three sidetone levels (none, low, and high), and three task type levels (word list, warm up, and creativity) as within-participant independent variables. We introduced participant gender (two levels: female and male) as a between-participant factor. The dependent variable was sound pressure level in $dB(SPL)$ level.

Figure 9 shows our measurement results as we varied sidetone. The vertical axis indicates the $dB(SPL)$ of participants' speech, averaged over all participants and all tasks.

Figure 10 shows our results by task, averaged across participants and sidetone levels.

We found that sidetone level affected people's vocal sound pressure levels, $F(2,36) = 7.21$, $p<.01$. Planned contrasts

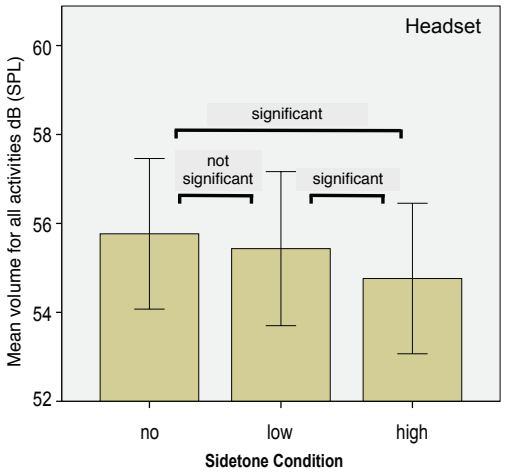

Figure 9: Speaking level averaged over participants across all sidetone settings in Experiment 1: Means and 95% confidence intervals.

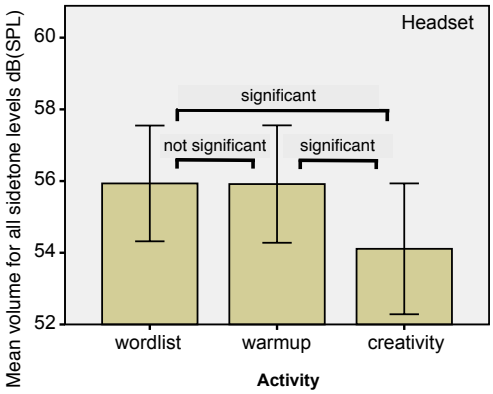

Figure 10: Speaking level averaged over participants across all tasks in Experiment 1: Means and 95% confidence intervals.

identified (*i*) a significant difference between the high sidetone condition and the no sidetone condition, $F(1,18) = 13.25$, $p<.01$, (*ii*) a significant difference between the high sidetone condition and the low sidetone condition, $F(1,18) = 6.59$, $p<.05$, and (*iii*) a non-significant difference between the low sidetone condition and the no sidetone condition. That is, people spoke more quietly in the high sidetone situation.

We also found that task type affected people's vocal sound pressure levels: $F(2,36) = 5.61$, $p<.05$, Greenhouse-Geisser corrected to account for a violation of the ANOVA sphericity assumption. Planned contrasts revealed a significant difference between the creativity task and the other two task types, warmup ($F(1,18) = 23.40$, $p<.001$), and word list ($F(1,18) = 5.06$, $p<.05$). That is, people spoke most quietly when doing the creativity task. See Figure 10.

Gender and the interaction effects between task and sidetone were not found to be significant predictors of vocal sound pressure levels.

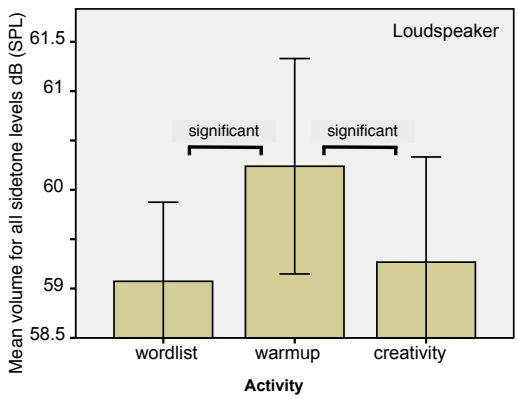

Figure 11: Speaking sound pressure level averaged over participants across all tasks in Experiment 2: Means and 95% confidence intervals.

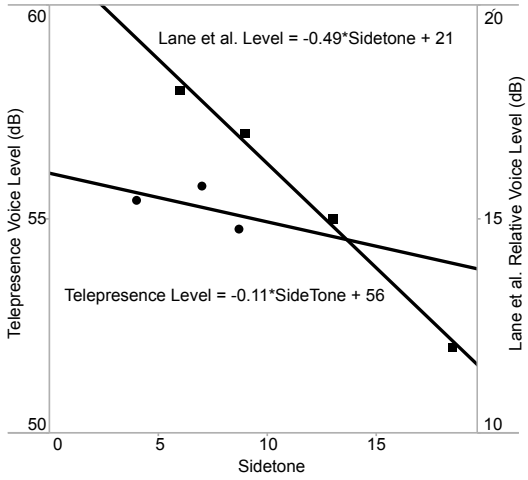

Figure 12: Comparison Lane et al. with this study's mobile remote presence sidetone effect.

Experiment 2: Loudspeaker

We performed the same ANOVA analysis on our loudspeaker data that we applied to the headset data. We found that task type affected people's vocal sound pressure levels, $F(2,24) = 3.63$, $p<.05$. Planned contrasts revealed a significant difference between the warmup task and the other two task types, creativity ($F(1,12) = 5.86$, $p<.05$) and word list ($F(1,12) = 5.76$, $p<.05$). That is, people produced higher sound pressure levels when doing the warmup task, which was relatively more socially oriented than the other two tasks. See Figure 11.

Sidetone level, gender, and interaction effects were not found to be significant predictors of people's vocal sound pressure levels.

DISCUSSION

For the headset condition we can compare our results against prior literature. The most striking aspect of our findings is that while we detected the expected statistically significant volume reducing impact of sidetone, the effect was smaller than was found in previous studies. Lane et al. [10] summarized ten sidetone experiments that mostly yielded slopes of around -0.5 for the linear function that describes the relationship between sidetone and speaker volume. Our observed slope is -0.1, a much reduced response. Only two of the ten studies found slopes as low as -0.25, and -0.3, respectively. Those studies administered sidetone monaurally, and they did not include any two-way communication tasks. Participants only read prepared passages.

In addition to the metastudy, Lane et al. conducted their own sidetone experiments [10]. We manually reproduced the sidetone portion of Lane's Figure 1, applied their x-axis shift to our sidetone levels, and reconstructed their original sidetone volume levels from their graph[1]. We then superimposed Lane's results with our mobile remote presence results.

Figure 12 shows Lane's results as squares against the right-hand side vertical axis. The current studies' results are the circles, and correspond to the left-hand side vertical axis. The vertical shift of the graphs is immaterial for two reasons. First, Lane added a base noise level to their audio. Second, the differences in microphone distance from all the different participants' mouths cannot easily be controlled in any of the experiments because it is very difficult to control without forcing study participants into very artificially constraining physical configurations. These distance variations will cause corresponding variations in absolute speaker volumes. All sidetone studies examine the *relative* variations produced by each individual in response to the sidetone manipulation, so the differences in absolute level are unimportant.

An implication of the shallow slope in our results is that even for high sidetone, which in contrast to levels used in telephones was clearly noticeable, the attenuation effect was $1dB$. That difference is (by definition of the decibel unit) the smallest difference still noticeable through the human ear.

At $2dB$, the effect of task type was stronger. The effect supports Lane's claim that actual communication tasks (as opposed to read-aloud tasks) will strengthen sidetone effects. In computer-mediated communication settings such as this one, we care about actual communication tasks.

In the loudspeaker experiment, only the task effect was significant, although the warmup task traded places with creativity. In addition, the effect was reduced from $2dB$ to $1dB$. We hypothesize that the sidetone effect is weaker with loudspeakers than headsets because (1) sidetone might be interpreted as ambient noise when it is heard through loudspeakers, (2) the speaker's own voice might sound dissimilar enough from his or her in-head experience, and/or (3) acoustic feedback limits the amount of sidetone that one can generate via loudspeakers.

The tasks may have been different for several reasons. We noticed that people seemed to be quietly thinking aloud to themselves in the creativity task, which was different from trying to speak loud enough to communicate in the first two tasks. The creativity task seemed to be the most cognitively

[1]This reconstruction was simply the reversal of Lane's complementing their sidetone vs. volume function, which they had undertaken to facilitate slope comparisons with manipulations other than sidetone.

demanding and engaging. We hypothesize that the task type results changed between Experiment 1 and Experiment 2 because people may be more sensitive to the sidetone manipulations during the first task out; it may have felt more strange to hear sidetone from a loudspeaker rather than a headset.

The main difference between our study and the prior work is that we introduced visual contact between the participants and the experimenter. The majority of earlier sidetone studies limited contact to binaural or monaural communication channels, like telephones or pilot-ground communication links. One study that did include a visual component was Siegel et al. [18], which worked with three and four year old children (with caretakers present), who were spoke about picture books they knew well. This situation is quite different from the current adult telepresence scenario.

We suspect that the high fidelity, almost head size screen presence of conversation partners in our application interferes with the sidetone effect. Thus, our hypothesis is that visual clues weaken the influence of sidetone.

LIMITATIONS AND FUTURE WORK

The two experiments presented explored a focused set of independent variables in the task domain of mobile remote presence. We chose the independent variables of sidetone levels and sound device (headset vs. loudspeaker) because previous literature suggested that they would influence how loud people would speak through the MRP system. Other levels of sidetones with other sound devices should also be explored if we are to rely upon the model presented in Figure 12, which requires more data points before we can have more confidence in its regression model.

To test the hypothesis that the presence of the video display influences the strength of the sidetone effect, a follow-up experiment could manipulate the video display presence vs. absence as an independent variable. To test the generalizability of these findings, other potentially useful independent variables would include the types of local MRP acoustic environments, ambient noise levels, degrees of experience with using MRP systems, and the types of conversational settings.

For any given user experience issue, there are a wide variety of design directions that one could choose from. Other design directions that could be explored include: more proactive visual displays to indicate when a user is speaking too loudly, ambient auditory feedback (e.g., alerts), less invasive interfaces for locals to decrease MRP volume levels, more directional audio outputs on the MRP systems, etc. Other effects to measure include the speaking volume of the locals, not just the pilots; we have noticed that locals align their speaking volumes to match those of the pilots, often without realizing it. This exacerbates the problem of annoying local bystanders and is an issue that also needs to be addressed.

A very different approach to controlling speaking volume is to override the user's volume, using digital signal processing (DSP). DSPs use limiters and compressors [25], which are algorithms—implemented in hardware or software—that control the volume at which a sound signal will be reproduced through a loudspeaker or headphones. A compressor is set to a particular threshold, ensuring that an input signal does not excessively rise above that limit. Several parameters may be tuned to accommodate different types of sound materials. The most important is the *attack*, the speed with which the compressor will gain reduce an overly loud input signal. Another is the *release*, which is the amount of time the compressor will subsequently exert control over the signal. The third is the compression *ratio*, which is the amount of gain reduction the compressor will impose. A limiter is an extreme form of compression, in which a signal is guaranteed not to exceed the threshold at all. Many variations exist for these DSP approaches. As described in the introduction, all suffer from shortcomings when they are placed in acoustically highly unpredictable environments, but there is potential for exploration in this direction.

CONCLUSION

We set out to address the problem of mobile remote presence operators speaking too loudly through their MRP systems, which frustrated locals near the MRP. Using lessons from the telephone industry, we explored the effectiveness of using sidetones to provide auditory feedback to remote operators to subtly moderate their speaking volume levels. Prior studies examined sidetone under headset conditions, often using one-way read-aloud tasks. Our MRP scenario is more complex in that the roaming nature of the *ersatzperson* precludes the use of headphones for locals who interact with the remote pilots. In addition, the depth of conversational engagement through MRP systems is nearly comparable to face-to-face interactions so we used a set of more interactive tasks.

In the first experiment, we used headsets to check if sidetone effects would work at all through our MRP system. Indeed, it did. Remote pilots spoke more quietly when they heard high sidetones as opposed to low or no sidetones. Because many pilots actually do use headsets when operating MRP systems, this is a promising result for aiding those pilots in speaking at more appropriate volume levels through the MRP systems.

In the second experiment, we used loudspeakers instead of headsets because some remote pilots use loudspeakers when piloting MRPs. This is very different from prior work with telephone headsets. It was clear from the start that providing sidetone via loudspeakers would introduce audio feedback problems that would require significant engineering effort. Indeed, we had to change the sidetone levels to avoid screeching audio feedback problems. In this second experiment, we did not find evidence that these sidetones were helpful for aiding remote pilots (who use loudspeakers) in self-regulating their speaking loudness.

In both experiments we found that the task type influenced how loudly people spoke. This suggests that exploring a wider variety of communicative tasks could provide insight into the ways that mediating communication influences how people talk with one another.

In MRP and similar systems (e.g., video conferencing), the current studies provide support for the notion that sidetones could help with self-regulation of loudness by remote operators, who use headsets. However, we did not find evidence that sidetones would help remote operators, who use loud-

speakers. This work presents the first test of sidetones in the mobile remote presence task domain.

Despite the widespread use of sidetones in landline telephones, it is rarely used in computer mediated communication settings. These prototypes and studies point us in the direction of reviving historically effective (though often forgotten) technologies for new application domains. Of course, these technologies will not necessarily work exactly the same way in the context of these new technological settings so it is important to evaluate these solutions in the new contexts before re-adopting them.

ACKNOWLEDGMENTS

Thanks go to the MRP project team and the participants in studies. Thanks also to Bill Smart for his writing guidance.

REFERENCES

1. American National Standards of the Acoustical Society of America. American National Standard method for measuring the intelligibility of speech over communication systems, May 2009.
2. M. Argyle. *Bodily Communication*. Methuen, New York, NY, 1975.
3. J. M. Beer and L. Takayama. Mobile remote presence systems for older adults: Acceptance, benefits, and concern. In *Proceedings of Human Robot Interaction: HRI*, pages 19–26, 2011.
4. H. Clark. *Using Language*. Cambridge University Press, Cambridge, UK, 1996.
5. D.J. Goodman and J.D. Johnston. Sidetone expansion for the regulation of talker loudness. *Electronics Letters*, 15(16):492 – 493, 1979.
6. J.P. Guilford, P.R. Christensen, P.R. Merrifield, and R.C. Wilson. Alternate uses task manual, 1978.
7. E.T. Hall. *The Hidden Dimension*. Doubleday, New York, NY, 1966.
8. E. S. Jones. Providing Foldback with Out-Of-Phase Loudspeakers. *Journal of the Audio Engineering Society*, 19(4):306–&, 1971.
9. N. P. Jouppi. First steps toward mutually-immersive mobile telepresence. In *Proceedings of Computer Supported Cooperative Work: CSCW*, pages 354–363, 2002.
10. H. Lane, B. Tranel, and C. Sisson. Regulation of voice communication by sensory dynamics. *Journal of the Acoustical Society of America*, 47:618 – 624, 1970.
11. M.-K. Lee and L. Takayama. "Now, I Have a Body": Uses and social norms for mobile remote presence in the workplace. In *Proceedings of Human Factors in Computing Systems: CHI*, 2011.
12. E. Lombard. Le signe de l'elevation de la voix. *Ann. maladies oreille larynx nez pharynx*, 37:109–119, 1911.
13. D. Nestel, P. Sains, C.M. Wetzel, C. Nolan, A. Tay, R.L. Kneebone, and A.W. Darzi. Communication skills for mobile remote presence technology in clinical interactions. *Journal of Telemed Telecare*, 13(2):100–104, 2007.
14. D. A. Norman. Minimizing the annoyance of the mobile phone, 2005.
15. E. Paulos and J.F. Canny. PRoP: Personal roving presence. In *Proceedings of Human Factors in Computing Systems: CHI*, pages 296–303, 1998.
16. B. Reeves and C. Nass. *The media equation: How people treat computers, television, and new media like real people and places*. Cambridge University Press, 1996.
17. D. Sakamoto, T. Kanda, T. Ono, H. Ishiguro, and N. Hagita. Android as a telecommunication medium with a human-like presence. In *Proceedings of Human Robot Interaction: HRI*, pages 193–200, 2007.
18. G. M. Siegel, H. L. Pick, M. G. Olsen, and L. Sawin. Auditory feedback in the regulation of vocal intensity of preschool children. *Developmental Psychology*, 12(3):255–261, 1976.
19. G.M. Siegel and H.L. Pick Jr. Auditory feedback in the regulation of voice. *Journal of the Acoustical Society of America*, 56(5):1618 – 1624, 1974.
20. L. Takayama and C. Pantofaru. Influences on proxemic behaviors in human-robot interaction. In *Proceedings of Intelligent Robotic Systems: IROS*, pages 5495–5502, St. Louis, MO, USA, 2009.
21. L. Takayama, J. G. Sison, B. Lathrop, N. Wolfe, A. Chiang, A. Nielsen, and C. Nass. Bringing design considerations to the mobile phone and driving debate. In *Proceedings of Human Factors in Computing Systems: CHI*, pages 1643–1646, 2009.
22. D. Tannen. *You Just Don't Understand: Women and Men in Conversation*. Harper Paperbacks, 2001.
23. K. Tsui, M. Desai, H. A. Yanco, and C. Uhlik. Exploring use cases for telepresence robots. In *Proceedings of Human Robot Interaction: HRI*, pages 11–18, 2011.
24. G. Venolia, J. Tang, R. Cervantes, S. Bly, G. Robertson, B.s Lee, and K. Inkpen. Embodied social proxy: Mediating interpersonal connection in hub-and-satellite teams. In *Proceedings of Human Factors in Computing Systems: CHI*, pages 1049–1058, 2010.
25. Wikipedia. Dynamic range compression.
26. M. Wolff. *Peoples In Places: The Sociology Of The Familiar*, chapter Notes On The Behaviour Of Pedestrians, pages 35–48. Praeger, 1975.
27. N.H. Wolfinger. Notes on the behaviour of pedestrians. *Journal of Contemporary Ethnography*, 24:323–340, 1995.