Does Gender and Accent of Voice Matter? An Interactive Voice Response (IVR) experiment

Eman Mubarak
University of Michigan
emanm@med.umich.edu

Tooba Shahid
Massachusetts Institute of Technology
toobash@mit.edu

Maryam Mustafa
Lahore University of Management Sciences
maryam_mustafa@lums.edu.pk

Mustafa Naseem
University of Michigan
mnaseem@umich.edu

ABSTRACT

We explore the impact of the gender and accent of the voice recording in Interactive Voice Response (IVR) systems in low-literate and patriarchal contexts. We conducted a small randomized control trial (RCT) with 62 participants, prompting them to identify myths and factual statements from a list of 10 prompts. One of four sets of recordings were randomized for each participant: Male formal (MF), Male Informal (MI), Female Formal (FF), and Female Informal (FI). We found that (a) male participants found male voices as providing more accurate information, (b) formal voice made myths seem accurate to male participants, (c) there was a significant impact of participant education level on correctly identifying myths, and (d) female participants were more knowledgeable about maternal health facts. Our study provides some basic guidelines on the potential characteristics of the voice used in IVR systems when deployment is in low-literate, and patriarchal communities.

CCS CONCEPTS

• **Computer systems organization** → **Embedded systems**; *Redundancy*; Robotics; • **Networks** → Network reliability.

KEYWORDS

Interactive Voice Response (IVR); gender; speech; ICT4D; accent

ACM Reference Format:
Eman Mubarak, Tooba Shahid, Maryam Mustafa, and Mustafa Naseem. 2020. Does Gender and Accent of Voice Matter? An Interactive Voice Response (IVR) experiment. In *Information and Communication Technologies and Development (ICTD '20), June 17–20, 2020, Guayaquil, Ecuador.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3392561.3397588

We explore the impact of the gender and accent of the voice in Interactive Voice Response (IVR) systems in low-literate and patriarchal contexts. Although in recent years there has been a push to create and deploy IVR systems for the dissemination of health information amongst low-literate populations [16, 20], the impact of the gender or accent of the voice in such a system on the trust in the disseminated information is still not clearly understood. Access to accurate and timely health information is important in promoting healthy behaviors and reducing health inequities [1]. Improved health literacy allows individuals to be more comfortable engaging with healthcare services in their community and ultimately experience better health outcomes. Thus, the empowerment of individuals within a community is inextricably tied to their ability to access, understand, and appropriately use health information.

In Pakistan, approximately 30 percent of men and 50 percent of women are low-literate [2]. Mass media is often used to convey health messages, however access is limited for those who are female, uneducated, living in rural areas, and/or impoverished. Mobile phone ownership also differs significantly by gender, with 92.7 percent of men and 39.2 percent women owning a mobile phone [2]. With health outcomes of women, including maternal mortality rates, continuing to fall short of intended goals despite counter efforts, it has become increasingly important to understand how to improve health literacy and healthcare access for women in Pakistan, particularly those of low-income and low-literate backgrounds [18]. While mobile health has shown promise in improving health knowledge and motivating behavior change in low-access communities, there are inherent limitations to implementation in these settings [7, 19]. For example, SMS-based interventions require literacy of their users. Other technologies have also been explored in the realm of healthcare, including pushed voice messages and other call-based formats [4, 7].

One important solution to this challenge is to deploy IVR systems on simple feature phones to reach the target populations. These systems are then used to deploy health information along with entertainment features to motivate use [14]. However within low-literate populations that have limited exposure to technologies and have patriarchal set-ups, the impact of the gender of the voice used to deliver key information is not clearly understood. The design of any localized IVR platform requires attention to how certain audio profile characteristics may impact user satisfaction and disclosure of sensitive information within the population it aims to cater to. We conduct a study to explore the co-relation between the trust and belief that low-income, low-literate users have on the given information and the characteristics of the IVR voice.

1 RELATED WORK

The characteristics of a voice, such as dialect, gender, or perceived age, have been shown to impact the listener's assessment of the personality traits the speaker possesses [5]. Interestingly, literature

ICTD '20, June 17–20, 2020, Guayaquil, Ecuador

ACM ISBN 978-1-4503-8762-0/20/06.
https://doi.org/10.1145/3392561.3397588

assessing gender bias in recorded speech has generated a conflicting body of evidence. Research led by MacDorman in 2011 showed that men and women find a female voice to be "warmer" than a male voice [12]. Indeed, many everyday voice assistants, such as Apple's Siri, Amazon's Alexa, and Microsoft's Cortana feature female voices. Another study has shown that both genders have a preference for adult male voices and still others report no difference in user preferences for gender at all [6, 10, 11]. One possible reason for these discrepancies is that the preference for the gender of the voice likely depends on the user population and the gender norms that dictate a society. To our knowledge, none of the studies in the field have examined the impact of gender or any other voice characteristic in the Urdu language.

Although Urdu is the national language of Pakistan, it is the first language of less than 8 percent of the country [3]. Punjabi is the largest mother tongue by population. While most Pakistanis are familiar with at least two languages, urban areas of Punjab have seen a decrease in the proportion of Punjabi speakers over the last several censuses [9]. Additionally, the Punjabi lexicon has evolved to include more Urdu words over the past few decades, possibly because Punjabi may be associated with lower prestige as well as economic, social, and educational value compared to Urdu [9, 15]. Research aimed at understanding how different accents of Urdu may be perceived among Pakistanis is lacking. The goal of this study is to determine whether low-literate men and women within an urban setting in Lahore, Pakistan have a preference for certain voice characteristics including gender and accent of Urdu when discussing reproductive and maternal health information. This preliminary research is intended to inform future projects related to developing an IVR system to disseminate women's health information in Pakistan.

2 METHODOLOGY

We performed a small randomized control trial, with data from 62 interviews used for analysis. The experiment was designed to evaluate the effect of gender and accent of a voice recording used in an IVR helpline, specifically in the context of sharing information about maternal health. The primary goal of this study was to determine if low-literate, low-income users had a preference for a specific gender or accent for the voice of the IVR recordings. To determine a preference for gender, we compared a male voice recording versus a female voice recording. For accent preferences, we compared a formal Urdu accent versus an informal Urdu accent with a Punjabi dialect. The study was conducted in the province of Punjab and the local language of our target population is Punjabi.

A specially curated, prerecorded script, consisting of five factual pieces of information (F1-F5) and five commonly believed myths (M1-M5) about maternal health were used in this experiment. The facts were created based on guidelines that the World Health Organization has set for maternal health. These 5 facts were specifically selected due to our previous understanding of their contentious nature in Pakistan. Some of them e.g. F1 (folic acid) and F4 (marrying young), were facts that most people knew to be true but were not acted upon by people due to contradicting cultural beliefs. The myths around maternal health were curated via a prior field study [13]. The 5 most commonly held beliefs were selected for the script. The Urdu script is attached alongside in the supplementary files, but the English translation can be found below.

- F1: Pregnant women should take folic acid supplements daily. Folic acid supplements are a form of medication taken by pregnant women to prevent anemia.
- F2: Pregnant women should have at least one ultrasound scan during the first trimester of the pregnancy. Ultrasounds let you make sure that the baby is healthy.
- F3: Breastfeeding should continue for up to two years after birth and beyond. This is essential to help the baby maintain a healthy weight and fight off diseases.
- F4: A mother with a child at the age of 19 or lesser increases the risk of complications for both newborns and young mothers. Complications associated with adolescent pregnancy include preterm delivery, low birth weight, and infant mortality
- F5: Births that occur without the assistance of a doctor are at a higher risk for complications. That is why giving birth at the hospital is better than home-births.
- M1: Pregnant women are more vulnerable to evil forces and should therefore be more careful. To avoid evil eye, pregnant women should go to a spiritual healer for an amulet.
- M2: Eating garam (hot) foods during pregnancy can cause miscarriage. Some examples of garam foods are nuts, eggs, dates, fish, and meat.
- M3: After delivery, water intake should be restricted because it causes bloating and makes it harder to lose weight.
- M4: Having back to back babies is beneficial for maintaining a happy family. This practice is healthy for the mother as well.
- M5: The sex of the baby is determined by the mother.

To test participant preferences for voice, we generated audio recordings of the same script in four different voice formats: Male Formal (MF), Male Informal (MI), Female Formal (FF), and Female Informal (FI). The distinction between formal and informal accent was determined by two native Punjabi speakers who also are fluent in Urdu. Some examples of how the accents differ are as following:"zaada" (informal) versus "zyaada"(formal), p'bundi (informal) versus pa'abundi (formal), baho (informal) versus bohat (formal) etc. These audio files were recorded using a smartphone application in the voice of two staff members at a university in Pakistan who volunteered for the role, and were selected by the research team via auditions prior to recording sessions.

One of the authors conducted short (10-15 minute), in-person structured interviews, where the author randomized and played back one of the four sets (MF, MC, FF, FC) of audio recordings to the participant. Before each interview, oral informed consent was obtained, and each interview was audio recorded. The audio recordings were stored and played from a smartphone app, and the participant was asked to answer each of the ten audio prompts as True or False. A total of 71 interviews were attempted, of which 62 interview recordings were usable. The remainder of the recordings were not transcribed due interviews that were cut short or technological malfunctions. Additionally, 6 of the interviews did not contain responses to all ten questions. These 6 interviews were used in analysis only if they contained a complete set of facts or

	Correctly identified myths (mean ± SE)
No schooling	0.937 ± 0.292
B/w Grade 5 and 10	2.000 ± 0.312
Tenth Grade or Higher	3.875 ± 0.413

Table 1: Relationship between correctly identified myths and education level of participant.

myths. The study was IRB approved as exempt at both the US and local university where the authors are affiliated.

At the end of the audio prompts, demographic information was collected. Any unprompted qualitative responses were transcribed and structurally coded by two of the authors. For quantitative data, two of the authors annotated the interviews into excel, and conducted ANOVA tests to evaluate the impact of site location, participant age bracket, and education status on interview responses. When significant results were detected, Tukey's HSD post hoc comparisons were conducted. The effects of gender and accent on interview responses were analyzed using Student's *t*-test . Fisher's exact test was used to determine statement-specific differences in responses among participants. All analyses were run using JMP Pro 15.0.0. Statistical significance was set at $p \leq 0.05$ for all tests.

3 FINDINGS

We recruited low-income, low-literate individuals using convenience and snowball sampling. Our data sample consisted of 23 male and 33 female participants. While we did not ask our participants for their income level, the recruitment strategy was specifically geared towards lower-income individuals: we recruited three sets of participants (a) janitorial staff at two universities in a large Pakistani city (b) a vocational training center in the same large city, and (c) a free maternity clinic and a local factory in a small Pakistani town. Our analyses did not reveal a significant effect of site location on the number of statements participants answered correctly, indicating that participants from these sampling locations can be assumed to have similar levels of baseline knowledge on the subjects of interest. The most common age group was 18-25 years of age. Participants were stratified into three education level groups: no schooling (40%), between primary and secondary school (35%), and 10th grade or higher (25%).

3.1 Finding # 1: Impact of Education Level on Correctly Identifying Myths

There was a significant effect of participant education level on the number of statements out of ten that were answered correctly ($F(2,29) = 11.31$, $p = 0.0002$). Post-hoc comparisons using Tukey's HSD test revealed that the mean score for participants with an education level of matric or higher (M = 8.33, SE = 0.51) was significantly higher than both the mean score for participants with no schooling (M = 5.47, SE = 0.32, $p = 0.0002$) and with an education level up to middle school (M = 6.55, SE = 0.38, $p = 0.02$). Interestingly, while the effect of education level on the number of statements answered correctly was not significant when examining only the five facts from our set of ten statements, education level was a significant factor for the five myths ($F(2,35) = 16.87$, $p < 0.001$). Post-hoc analyses demonstrated that this difference is significant among all three groups of education level: matric or higher vs. no school ($p < 0.0001$), matric or higher vs. between primary and secondary school ($p = 0.003$), and no school vs. between primary and secondary school ($p = 0.046$). Therefore, higher levels of educational attainment were consistent with higher numbers of correctly identified myths, as shown in Table 1.

	Gender of participant	Gender of recording	Accent of recording
Total correct statements	$t(47) = 0.045$, $p = 0.964$	$t(47) = 0.935$, $p = 0.354$	$t(47) = 1.301$, $p = 0.200$
Correctly identified facts	$\mathbf{t(54) = 2.106}$, $\mathbf{p = 0.040^*}$	$\mathbf{t(54) = 2.380}$, $\mathbf{p = 0.021^*}$	$t(54) = 1.030$, $p = 0.308$
Correctly identified myths	$t(53) = 0.394$, $p = 0.695$	$t(53) = 0.210$, $p = 0.835$	$t(53) = 1.681$, $p = 0.099$

Table 2: Pooled *t*-test demonstrating relationship between correctly identified response and Gender of Participant, Gender of Recording, and Accent of Recording.

3.2 Finding # 2: Female participants are more knowledgeable about Maternal Health Facts

As pointed out in table 2, we did not see any differences in the total correct responses out of ten among participants overall based on the gender of the participant, gender of the recording, or the accent of the recording. When examining only facts, female participants correctly identified more facts out of five (M = 4.82, SE = 0.08) than male participants (M= 4.57, SE = 0.09), $p = 0.04$.

3.3 Finding # 3: Male participants perceive male voices as providing more accurate information

Participants that listened to a male recording correctly identified more facts out of five (M = 4.88, SE = 0.09) than those that listened to a female recording (M = 4.56, SE = 0.08), $p = 0.02$. When analyzing participants separately by gender, male respondents scored significantly higher in identifying facts when listening to a male voice (M = 4.889, SE = 0.147) than a female voice (M = 4.36, SE = 0.12), $p = 0.01$. This difference did not exist for females. Table 3 shows a male-only pooled *t*-test demonstrating these findings.

3.4 Finding # 4: Myths and Gullibility to formal voice

Table 3 shows that male respondents correctly identified more myths out of 5 when listening to a recording with an informal accent (M = 3.10, SE = 1.20) rather than a formal accent (M = 1.57, SE = 1.87), $p = 0.03$. This difference did not exist for female respondents.

The Fisher's Exact test (Table 4) for both male and female participants revealed that the proportion of participants that responded correctly to statements M1 and M4 significantly differed by accent of the recording ($p = 0.02$, $p = 0.002$ respectively). For both statements, a higher proportion of participants answered correctly

Gender of recording	Accent of recording	
Total correct statements	$t(20) = 1.094$, $p = 0.287$	$t(20) = 1.819$, $p = 0.084$
Correctly identified facts	**$t(21) = 2.816$, $p = 0.010^{*}$**	$t(21) = 0.442$, $p = 0.663$
Correctly identified myths	$t(22) = 0.619$, $p = 0.542$	**$t(22) = 2.267$, $p = 0.034^{*}$**

Table 3: Male-only pooled *t*-test demonstrating relationship between correctly identified response and Gender and Accent of Recording.

when the recorded voice had an informal accent compared to a formal accent (66.7% vs. 36.1% and 81.5% vs. 43.2% respectively). No differences were found when analyzing the effect of accent or gender of the recording on any of the other statements.

Our interview method allowed us to obtain limited qualitative information from our respondents as well. The researcher did not prompt respondents for qualitative feedback in between playing the 10 recorded prompts, however, sometimes participants volunteered qualitative feedback after a particular prompt was played. In response to M1, four of the participants that answered that while they believed in the concept of evil eye (nazar), they did not feel that spiritual healers and/or amulets were effective in warding off the evil eye:

"We don't get any amulets. Whatever will happen, will happen with God's will anyways." (P71)

Several other interviews revealed that participants practiced a spectrum of homeopathic and religious healing methods:

"Women should pray more when they are pregnant. They should say their prayers, read the Quran. Ask for Allah's blessings for themselves." (P59)

"...I ate Black pepper and carom seeds after getting them blessed, and thankfully I had two healthy baby boys delivered (P60)

In response to M4, several participants that answered "false" mentioned various intervals between pregnancies, ranging between 2-5 years. When one participant was asked why they answered "true" to this statement, their response was:

"Every parent wishes that they have a large family" (P2)

Notably, participants that requested further clarification of a statement by the interviewer, and subsequently changed their answer, only did so for M1 and M4. To remain consistent, our analyses takes into account the initial responses for these questions to best capture the effect of the voice recording rather than any bias the interviewer may introduce during clarification.

3.5 Finding # 5: Trust in IVR or participant response bias?

Interestingly, 37.5% of the participants responded "true" to nine out of ten or all ten statements regardless of the characteristics of the voice recording they heard, suggesting that (a) either in this population, regardless of the characteristics of the recorded voice, an IVR-based health intervention elicits at least some level of trust from its users, or (b) this could be participant response bias due to interviewer demand characteristics and the role of social and demographic factors in influencing that bias. The researcher conducting the research is an educated individual who currently resides in the US, who may have been perceived as trustworthy by the participants [8].

	Male vs. Female Recording	Formal vs. Informal Accent
F1: folic acid	$p = 0.437$	$p = 0.437$
F2: ultrasound	$p = 0.462$	$p = 0.415$
F3: breastfeeding	n/a	n/a
F4: age of childbearing	$p = 0.160$	$p = 0.725$
F5: doctor-assisted delivery	$p = 1.000$	$p = 1.000$
M1: evil eye	$p = 1.000$	**$p = 0.023^{*}$**
M2: Hot food	$p = 1.000$	$p = 0.777$
M3: weight loss	$p = 0.572$	$p = 1.000$
M4: back-to-back babies	$p = 0.800$	**$p = 0.002^{**}$**
M5: gender of baby	$p = 1.000$	$p = 0.430$

Table 4: *p*-values for 2-tailed Fisher's Exact Test comparing (a) Male vs. Female Recording, and (b) Formal vs. Informal Accent.

4 DISCUSSION

Our study aimed to understand the impact of the characteristics of a recorded voice on the belief and perception of presented information in a low-literate, low-income audience within a patriarchal context. Our work shows that male participants found male voice recordings to be more believable than female recordings. This is a key finding to inform the design of future IVR systems, particularly in patriarchal countries, since previous IVR systems deployed in the Global South have not considered the gender of the voice [16, 17, 20]. We also found that participants inherently placed a certain level of trust in the information received through an IVR system, regardless of the voice characteristics. This places a greater burden on designers to ensure all knowledge and information being disseminated is accurate. This is particularly relevant for social voice based systems like Baang, which are open and allow posting of content without fact checking [16]. Additionally, because myths were more often correctly identified as false when an informal voice was used, a formal voice was perceived as more trustworthy. A limitation of our study was that our script contained only five facts and five myths related to maternal health. While this was enough to yield statistically significant results in our analyses, future studies may choose to expand on this subject matter. Our study provides some basic guidelines on the potential characteristics of the voice used in IVR systems when deployment is in low-literate, and patriarchal communities. Although a more detailed study is warranted to further analyze voice characteristics in these settings, we provide a foundation for future work to build upon.

ACKNOWLEDGMENTS

We would like to acknowledge the participants for their time, the speech artists who recorded the prompts in the four different options, as well as admin staff at the local universities who helped recruit participants.

REFERENCES

[1] 2017. The mandate for health literacy. https://www.who.int/healthpromotion/conferences/9gchp/health-literacy/en/

[2] 2019. *Pakistan Demographic and Health Survey 2017-18.*

[3] 2020. POPULATION BY MOTHER TONGUE. http://www.pbs.gov.pk/content/population-mother-tongue

[4] Stella Babalola, Caitlin Loehr, Olamide Oyenubi, Akinsewa Akiode, and Allison Mobley. 2019. Efficacy of a Digital Health Tool on Contraceptive Ideation and Use in Nigeria: Results of a Cluster-Randomized Control Trial. *Global Health: Science and Practice* 7, 2 (2019), 273–288.

[5] Brittni Elizabeth Bergstrom. 2017. Effect of Speaker Age and Dialect on Listener Perceptions of Personality. (2017).

[6] Mick P Couper, Eleanor Singer, and Roger Tourangeau. 2004. Does voice matter? An interactive voice response (IVR) experiment. *Journal of official statistics* 20, 3 (2004), 551.

[7] Jessica Crawford, Erin Larsen-Cooper, Zachariah Jezman, Stacey C Cunningham, and Emily Bancroft. 2014. SMS versus voice messaging to deliver MNCH communication in rural Malawi: assessment of delivery success and user experience. *Global Health: Science and Practice* 2, 1 (2014), 35–46.

[8] Nicola Dell, Vidya Vaidyanathan, Indrani Medhi, Edward Cutrell, and William Thies. 2012. "Yours is Better!": Participant Response Bias in HCI. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Austin, Texas, USA) *(CHI '12)*. Association for Computing Machinery, New York, NY, USA, 1321–1330. https://doi.org/10.1145/2207676.2208589

[9] Muhammad Din and Mamuna Ghani. 2017. A Sociolinguistic Study on the Diminishing Features of the Punjabi Lexicon. *International Journal of English Linguistics* 7, 3 (2017), p152.

[10] Rochelle E Evans and Philip Kortum. 2010. The impact of voice characteristics on user response in an interactive voice response system. *Interacting with Computers* 22, 6 (2010), 606–614.

[11] Pat Mirenda, Douglas Eicher, and David R Beukelman. 1989. Synthetic and natural speech preferences of male and female listeners in four age groups. *Journal of Speech, Language, and Hearing Research* 32, 1 (1989), 175–183.

[12] Wade J Mitchell, Chin-Chang Ho, Himalaya Patel, and Karl F MacDorman. 2011. Does social desirability bias favor humans? Explicit–implicit evaluations of synthesized speech support a new HCI model of impression management. *Computers in Human Behavior* 27, 1 (2011), 402–412.

[13] Maryam Mustafa, Amna Batool, Beenish Fatima, Fareeda Nawaz, Kentaro Toyama, and Agha Ali Raza. 2020. Patriarchy, Maternal Health and Spiritual Healing: Designing Maternal Health Interventions in Pakistan. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[14] Mustafa Naseem, Bilal Saleem, Sacha St-Onge Ahmad, Jay Chen, and Agha Ali Raza. 2020. An Empirical Comparison of Technologically Mediated Advertising in Under-Connected Populations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376683

[15] Barirah Nazir, Umair Aftab, and Ammara Saeed. 2013. Language shift–The case of Punjabi in Sargodha region of Pakistan. *Acta Linguistica Asiatica* 3, 2 (2013), 41–60.

[16] Agha Ali Raza, Bilal Saleem, Shan Randhawa, Zain Tariq, Awais Athar, Umar Saif, and Roni Rosenfeld. 2018. Baang: A viral speech-based social platform for under-connected populations. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–12.

[17] Waleed Riaz, Haris Durrani, Suleman Shahid, and Agha Ali Raza. 2017. ICT Intervention for Agriculture Development: Designing an IVR System for Farmers in Pakistan. In *Proceedings of the Ninth International Conference on Information and Communication Technologies and Development*. 1–5.

[18] Arjumand Rizvi, Zaid Bhatti, Jai K Das, and Zulfiqar A Bhutta. 2015. Pakistan and the millennium development goals for maternal and child health: progress and the way forward. *Paediatrics and international child health* 35, 4 (2015), 287–297.

[19] JA Unger, Keshet Ronen, Trevor Perrier, Brian DeRenzi, Jennifer Slyker, AL Drake, Dyphna Mogaka, John Kinuthia, and Grace John-Stewart. 2018. Short message service communication improves exclusive breastfeeding and early postpartum contraception in a low-to middle-income country setting: a randomised trial. *BJOG: An International Journal of Obstetrics & Gynaecology* 125, 12 (2018), 1620–1629.

[20] Aditya Vashistha, Edward Cutrell, Gaetano Borriello, and William Thies. 2015. Sangeet swara: A community-moderated voice forum in rural india. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 417–426.