TapBack: Towards Richer Mobile Interfaces in Impoverished Contexts

Simon Robinson[1], **Nitendra Rajput**[2], **Matt Jones**[1],
Anupam Jain[2], **Shrey Sahay**[2], **Amit Nanavati**[2]

[1] Future Interaction Technology Lab
Swansea University, SA2 8PP, UK
{ s.n.w.robinson, matt.jones } @swan.ac.uk

[2] IBM India Research Laboratory
Vasant Kunj, New Delhi, 110070, India
{ rnitendra, anupamjain, shrsahay, namit } @in.ibm.com

ABSTRACT
Much of the mobile work by HCI researchers explores a future world populated by high-end devices and relatively affluent users. This paper turns to consider the hundreds of millions of people for whom such sophistication will not be realised for many years to come. In developing world contexts, people will continue to rely on voice-primary interactions due to both literacy and economic reasons. Here, we motivate research into how to accommodate advanced mobile interface techniques while overcoming the handset, data-connection and user limitations. As a first step we introduce TapBack: back-of-device taps to control a dialled-up, telephone-network-based voice service. We show how these *audio gestures* might be recognised over a standard telephone connection, via users' existing low-end devices. Further, in a longitudinal deployment, the techniques were made available on a live voice service used by rural Indian farmers. Data from the study illustrates the desire by users to adopt the approach and its potential extensions.

Author Keywords
Developing regions, mobile HCI, audio gestures, back-of-device interaction, spoken web.

ACM Classification Keywords
H.5.2 [User Interfaces]: Input devices and strategies; Interaction styles; H.5.1 [Multimedia Info. Systems]: Audio I/O.

General Terms
Design, Experimentation, Human Factors.

INTRODUCTION

For hundreds of millions of people in developing, rural regions, the mobile phone is the primary – if not only – interactive technology available. Already pervasively used for calling, these devices are increasingly set to become access terminals for remote information services.

Unlike the state-of-the-art, future-looking devices often studied by HCI researchers, though, a large proportion of these mobiles are likely to remain relatively dumb-phones with only a low proportion being routinely served by a data connection. Furthermore, the users themselves add additional challenges to the goal of universal access: many have a low level of textual literacy, and their prior exposure to computing technology is often very limited.

To meet these challenges, a class of network-level audio-based services have been proposed. These often combine automatic speech recognition (ASR) and touch tone dialling (DTMF) to allow people to create and browse through spoken content. The Spoken Web [4], for example, is a collection of interconnected *voice sites*. These interactive audio applications provide content on topics such as farming or health information over the public telecom network. Individual voice sites are accessed using any type of phone by dialling unique telephone numbers (analogous to URLs).

Although both ASR and DTMF allow a level of control and interaction with audio content, we believe there is still much work to be done in terms of improving the expressiveness and range of interactions. As a first step towards richer mobile voice interfaces, we present TapBack: an extended interaction method for voice sites that aims to allow callers to smoothly navigate through and control the content they are listening to without having to unnecessarily interrupt its playback. Our approach uses simple back-of-device interactions – *audio gestures* – on the phones users already own.

While there has been previous research on back-of-device and non-speech natural audio input (e.g. [5, 6]), this has involved state-of-the-art custom-built devices, and users with high levels of literacy and technology experience. In contrast, the majority of our target audience use relatively low-end mobile phones, so we have focused on providing these additional interaction features without requiring users to own a specialised device. The users themselves are also from very different backgrounds to those often studied by other researchers, bringing additional insights and challenges.

The contribution here, then, is an exploration of ways that impoverished platforms and their users can be afforded the sorts of advanced interactions being imagined for people living in the 'developed' world.

BACKGROUND

ASR & DTMF Interaction

Automatic speech recognition promises intuitive, low cognitive load interaction with audio content, with the benefit that no base level of literacy or numeracy is required. However, the pauses or cues that are needed to prompt speech input and detection can easily upset the interaction flow, especially for short inputs such as 'yes' or 'no' [7].

DTMF key tones are quick to enter on keypads, designed to be unambiguous when recognising, and offer many possible input sequences. However, because the tones generated are echoed in the phone speaker (confirming input, but drowning out incoming audio), interactions and responses are often fragmented, unlike in our design. A further key issue with DTMF is that it is often necessary for the caller to take the phone away from their ear to see the keypad and respond to any input cues. In our design we have concentrated on removing this disruption – instead we allow callers to interact on the back of their device during the normal call flow.

Back-of-Device Interaction

As mobile devices have continued to offer more features in increasingly-compact form factors, researchers have recognised the need for interactions beyond the screen and keypad. This has culminated, recently, in touch-based back-of-device interaction with almost no need for a device at all [1]. However, although these designs offer touch interaction anywhere, they also require users to own specialised hardware. We believe it is possible to offer a subset of these input methods to users who might not have the most modern devices.

Li et al. [5] noted the problems that can result when callers attempt to use a phone's keypad without taking the device away from their ear, addressing these issues by using a modified keypad on the back of a phone with audio cues to assist. However, this was aimed at allowing use of the phone's functions during a call, unlike our approach, which uses back-of-device interaction to control the call itself.

Tap and Scratch Interaction

We build upon previous research into interaction that appropriates a device's surface as an input channel. Murray-Smith et al. [6] used a custom-built sensor pack shell with exterior textures that produced distinct sounds when scratched. A microphone inside the device captured the sounds, allowing complex scratches to be recognised as distinct commands. Our approach, although not capable of the same diversity in inputs, affords similar interaction on a normal phone.

Harrison and Hudson [3] built on this work to allow scratches and taps to be used as inputs on any solid surface, capturing sounds using a modified stethoscope. They found that inputs were reliably recognised by a fast, lightweight recognition engine – an approach we adopt in our system.

Mobile possibilities for these interactions have recently been demonstrated as a commercial prototype[1] that can detect tapping locally on a dumb-phone. In contrast with these types of approaches, however, TapBack also separates the source of the audio gestures (the handset) and the system that interprets them (a server accessed over a normal telephone call).

[1] *TouchDevice* – www.inputdynamics.com

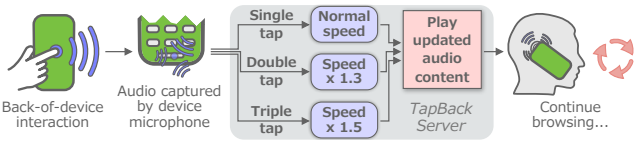

Figure 1. TapBack: Audio gestures are generated by the caller locally on the back of their handset, but analysed remotely (shaded region).

THE TAPBACK SYSTEM

The TapBack system allows callers a richer experience with interactive voice sites by enabling *audio gestures* to be used at any time during a call. By using the back of users' phones as an input surface while a call is in progress, we remove the interruptions of ASR/DTMF and allow users to keep the phone by their ear throughout the call. Unlike previous back-of-device methods, we use the phone's inbuilt microphone to pick up the sounds generated on the back of its case. These sounds are loud enough to be transferred to the other party in the call, but, unlike DTMF tones, are not so loud on the caller's end that they drown out the audio being played.

For a simple user introduction to audio gestures we chose to apply tapping recognition to the control of audio playback speed. Previous voice site analyses [2] have shown that callers would appreciate finer control of playback, so this was a natural application for our system. In our implementation, when users tap two or three times, the time compression is 25% and 35%, respectively, while still retaining intelligibility. Tapping once returns playback to its normal speed.

Implementation

The TapBack system is installed on a remote server, monitoring low-level network packets to track incoming phone calls to individual voice sites. When a call is established, real-time audio capturing, decoding and analysis is initialised. The audio is first filtered to remove frequencies below 3KHz, greatly reducing the problem of ambient noise. The stream is then windowed using a 512-sample Hamming window with an overlap of $\frac{7}{8}$. Tap recognition itself is very unsophisticated, simply searching each window for short, high-intensity, high-frequency sounds. Detected tap events are then fed to a higher-level audio gesture classifier which uses timeouts and basic heuristics to classify each tap type.

When an audio gesture is found, the system sends a command to the Spoken Web server, which adjusts the voice site playback speed in response to the request. Users are also free to control the speed of playback by using DTMF inputs instead of taps. In this case, keys 4, 5 and 6 correspond to single, double and triple taps, respectively.

EVALUATION

The TapBack system was evaluated in three ways, focusing primarily on the viability of the approach, rather than the sophistication of the recogniser. Recognition accuracy was

measured and refined using a test voice site. A live deployment of the system over an extended period was used to assess the usefulness and usability of the approach. Finally, we explored alternative audio gestures with participants to understand the potential for extending such techniques.

RECOGNITION ACCURACY TEST

We conducted a user study to measure and improve the recogniser's accuracy over a standard telephone connection. 18 users of an existing, popular farming information voice site based in a rural region in India were recruited (see [7] for more detailed user population demographics). To ensure a cross-section of user expertise, participants included people who access the voice site very regularly and also those who are only casual users. All users were male, and the average age was 32. The set of phones used by the participants consisted of 14 different (low-end) handset types produced by four manufacturers. All participants had already used the DTMF speed control methods detailed in [2].

Each participant was called by phone to explain the study method and the concept of audio gestures. The calls were made to participants when they were at locations from which they usually interact with the voice site services. The participant was then connected to a test voice site, which asked them to tap the back of the phone while holding it to their ear, in response to four sets of cues. Each cue set asked users to tap once; twice; then, three times. Each participant, therefore, provided 12 tap commands.

Recognition rates were: 1-tap: 93%; 2-tap: 78%; and, 3-tap: 56%. High accuracy rates for single and double taps were encouraging given: the minimal explanation of the concept to users; this form of interaction with a service was entirely novel in users' experience; a diverse set of low-end phones were involved; the audio channel was of standard telephone quality; and, the study was in a live, not laboratory setting.

A large proportion of the errors in recognition were due to participants tapping slower than the recogniser expected. This led to 2-taps being recognised as 1+1 taps (accounting for 50% of the 2-tap errors) and 3-taps being recognised as 1+2, 2+1 or 1+1+1 taps (60% of 3-tap errors). To deal with these errors, the tap classifier was modified to employ simple correction heuristics so that, for example, a 2-tap shortly followed by a 1-tap was interpreted as a 3-tap instruction. The remaining errors were caused by taps not being distinct enough for the recogniser to extract from the input.

DEPLOYMENT

The TapBack system was made available on a live farming information voice site [7]. This exploratory study aimed at measuring the adoption of audio gestures by logging any tap interactions and responses during normal use of the service.

Method

The system was deployed for 12 days, during which any of the 110 registered active users could call at any time. These users are geographically dispersed over a wide area of India, and are all farmers living in rural settings. When calling, users were given a brief automated introduction to the new method that explained how they could tap the back of their phone to control the playback speed. The system logged call details and any input actions (both tap-based and DTMF).

We supplemented this data by conducting detailed telephone interviews with 15 users. Ten of these were selected at random from the set of those who had used TapBack during the deployment period; the remaining five were randomly selected from callers who did not attempt to use the tap interaction. The average age of participants was 31, and all except one were male. During the interviews these users were asked about their reactions to the approach; how usable it was; and, any issues they had encountered in its use. Interviews were conducted in the participants' native language (Gujarati).

Results

286 calls to the voice site were recorded over the study period, from 52 unique callers. 1293 tap interactions were recorded in total. 36 callers used the TapBack feature (166 calls; 7.8 taps per call, on average).

Of the 36 participants that used tap interaction, 25 used the feature on more than one call. Two others called more than once but only used tap interaction on their first call; the remaining 9 TapBack users called only once over the study period. The 16 participants who did not use tap interactions did not use DTMF for speed control, either.

Tap interactions consisted of 772 single, 301 double, and 220 triple taps. Few of the callers that wanted to control the speed of the call used the DTMF method – 52 speed control DTMF events were recorded in total.

Participant Interviews

Considering first the ten callers who had used the TapBack feature. Of these, the majority were positive in their comments about the approach. Benefits mentioned ranged from those related to utility to those concerning the less-tangible 'user experience'. Several respondents commented on the tapping being easier to use and quicker than DTMF. Another interviewee talked of the 'fun' of the new interaction. Interestingly, one participant said, *"This is like having a touchscreen, this is a modern thing to use – it's cool."*

Negative comments from these 10 adopters included the predictable, such as frustration when a tap-event was not recognised: one respondent said he would always use buttons because, *"they always work – end of story."* However, there were also issues related to the use-context. Two interviewees worried about using the system regularly as the tapping, to their mind, might damage the phone. For one of these interviewees this was particularly worrying as they often lent their phone to others to use (a practice quite common in rural areas). Another respondent said they tended to listen to the service with a group of people using the speakerphone.

There were two explanations for the non-use of the approach by the five other interviewees. For some, their environments (as witnessed during telephone interviews) were too noisy;

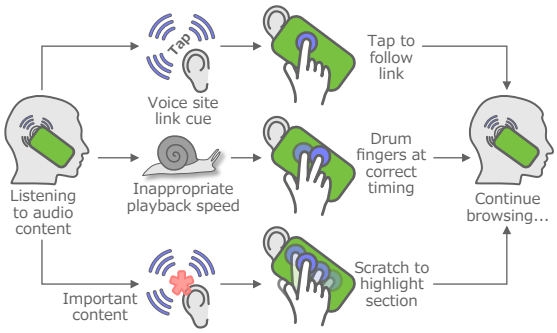

Figure 2. Potential future usage scenarios for audio gestures.

and, for others, they had not understood the new feature as explained by the voice site after call connection.

Discussion

The logged data provides evidence that callers are willing to adopt the tapping method, with the majority of callers using the approach. It should also be noted that the functions controlled by TapBack – speeding and slowing the audio – are optional: users are able to listen to and navigate through content without employing them. We would expect, then, that some calls during the study would not show tap interaction. Furthermore, 93% of callers who used TapBack during their first call also used it again in their subsequent calls. Callers that used TapBack did so several times in each call.

The comments about the system's utility value are, of course, encouraging, especially when considering the accuracy of our simple recogniser. However, of more note, perhaps, are the responses relating to the user experience – a fun, 'modern' interaction is something that is not usually associated with the low-end devices these users have access to. The negative comments are spurs to improve the recognition engine and explanation of its use. The social issues raised suggest extensions to our approach – we will need to ensure the tap-models used are tuned not just to individual phone numbers but the set of users that might use that phone; and, in speaker-mode use, we might be able to consider a wider set of audio gestures as suggested in [3].

ALTERNATIVE AUDIO GESTURES

In order to further understand the needs of the target users, we conducted a study to gather more potential audio gestures and corresponding actions. The 15 participants questioned during the deployment were also asked to suggest additional back-of-device interactions they were comfortable with, and to discuss what actions they felt these gestures might initiate.

The majority of interviewees identified the value of two single-handed, back-of-phone interactions: drumming fingers and scratching. Many users suggested finger-clicking (using the non-phone holding hand). Most users also raised the possibility of making non-verbal utterances – e.g., *"I could make the noise I make when shooing away cows."*

Interviewees found it hard to make mappings from their gestures to controls; unlike many participants, who have extensive computer experience, these respondents had no notion of interface metaphors. However, one commented on the use of drumming to skip through voice content; and many others wanted a fast way to jump to particular sections of the audio. Fig. 2 shows how we might widen the set of audio gestures in response to the studies.

CONCLUSIONS

We have introduced TapBack to illustrate the potential of audio gestures to complement voice-based interaction over the phone. While the technique might appear unsophisticated or less exciting compared to the methods proposed for high-end mobiles, we argue that it is far more likely to have impact in the sorts of context that concern us.

Of course, there is much work still to be done to improve our lightweight recogniser. Future work could focus on refinements to the recognition algorithms to improve accuracy. Alternatively, increased robustness might be achieved by simplifying the gesture set to allow only single taps. In the current speed control application, this could be applied as a toggle between options so each individual tap would change playback to the next speed preset.

We have shown how tapping can be used as input by rural Indian farmers via their basic mobile handsets. Further, these users' responses to the method, along with their suggestions for additional gestures, indicate the viability of the approach for groups of people with very low exposure to computing.

ACKNOWLEDGEMENTS

This work was supported by IBM Research India, IBM Faculty Award (Matt Jones) and EPSRC project EP/H042857/1.

REFERENCES

1. P. Baudisch and G. Chu. Back-of-device interaction allows creating very small touch devices. In *Proc. CHI '09*, pages 1923–1932. ACM, 2009.
2. K. A. Dhanesha, N. Rajput, and K. Srivastava. User driven audio content navigation for spoken web. In *Proc. Multimedia '10*, pages 1071–1074. ACM, 2010.
3. C. Harrison and S. E. Hudson. Scratch input: creating large, inexpensive, unpowered and mobile finger input surfaces. In *Proc. UIST '08*, pages 205–208. ACM, 2008.
4. A. Kumar, N. Rajput, D. Chakraborty, S. K. Agarwal, and A. A. Nanavati. WWTW: the world wide telecom web. In *Proc. NSDR '07: Workshop on Networked Systems for Developing Regions*, pages 1–6. ACM, 2007.
5. K. A. Li, P. Baudisch, and K. Hinckley. Blindsight: eyes-free access to mobile phones. In *Proc. CHI '08*, pages 1389–1398. ACM, 2008.
6. R. Murray-Smith, J. Williamson, S. Hughes, and T. Quaade. Stane: synthesized surfaces for tactile input. In *Proc. CHI '08*, pages 1299–1302. ACM, 2008.
7. N. Patel, D. Chittamuru, A. Jain, P. Dave, and T. S. Parikh. Avaaj Otalo: a field study of an interactive voice forum for small farmers in rural india. In *Proc. CHI '10*, pages 733–742. ACM, 2010.