Repurposing Conversation: Experiments with the Continuous Speech Stream

Donald McMillan **Antoine Loriette** **Barry Brown**
Mobile Life Centre, Stockholm University, Sweden
[don, antoine, barry]@mobilelifecentre.org

ABSTRACT
Voice interaction with mobile devices has been focused on hands-free interaction or situations where visual interfaces are not applicable. In this paper we explore a subtler means of interaction – speech recognition from continual, in the background, audio recording of conversations. We call this the 'continuous speech stream' and explore how it could be repurposed as user input. We analyse ten days of recorded audio from our participants, alongside corresponding interviews, to explore how systems might make use of extracts from this stream. Rather than containing directly actionable items, our data suggests that the continuous speech stream is a rich resource for identifying users' next actions, along with the interests and dispositions of those being recorded. Through design workshops we explored new interactions using the speech stream, and describe concepts for individual, shared and distributed use.

INTRODUCTION
The success of commercial speech interfaces – such as Apple's Siri, Google's Now, and Microsoft's Cortana – has highlighted new possibilities for speech interfaces, particularly on mobile devices. Yet HCI research into speech has been somewhat conflicted, with some claiming that HCI speech research is having a "middle-life crisis" [3], and that HCI researchers have been insufficiently opportunistic with new possibilities for speech interfaces. In this paper we attempt to explore one new approach for speech interfaces making use of always-on ambient sound recording as an input to system behaviour.

Using the "continuous speech stream" (CSS) we explore if and how systems could use speech recognition to pre-empt user actions, support activities without disrupting them, and draw on shared repositories of others' recognised speech. Most current commercial speech systems focus on a *dialogic* model of interaction, where users enter a dialogue with a system, where speech results are recognised for immediate system activity, with spoken responses, and the possibility for further dialogue. Yet recent advances in battery life and signal processing offer the possibility of systems that continually listen to natural conversation, and can conduct at least partial recognition on that speech to detect keywords, or even attempt transcription. With the opportunities of technologies such as smart watches or other wearable devices (such as the Motorola Hint earpiece), clear audio recording of everyday conversation is also easier to foresee (although with some serious privacy issues). Technically, this offers opportunities for systems that could modify and change their behaviour based on a users' implicit spoken interaction with others. For example, phones could automatically carry out actions based on spoken audio – searches could be opportunistically run and displayed on lock screens, adverts could be better based on user activities, and models of user behaviour could make use of richer descriptions of users' lives.

We explore currently available open source speech recognition software with which we were able to recognise keywords from a previous recorded corpus, suggesting potential for future implementation. The paper then moves on to study system opportunities for using a 'continuous sound stream' (CSS). We recorded day-long audio streams from ten participants, and manually partially transcribed the stream to explore pre-emptive system functions. Interviewing participants we explored possible automatic system actions from the CSS. Two design workshops produced a set of potential design concepts based around individual, collective and distributed use of the speech stream. In conclusion we discuss the relevance of our findings for speech as an interface more broadly.

BACKGROUND WORK
Automatic speech recognition (ASR) is broadly defined as the translation of spoken words, or an acoustic waveform containing speech, into a string of words [8]. In modern ASR speech is modelled as a mixture of acoustic and language properties [2]. The acoustic models attempt to move from audio samples of speech to potential phonemes being spoken, and from these phonemes to possible words. While speaker independent speech recognition is possible, acoustic model accuracy increases by up to three fold if they are trained on a specific speaker [26]. These have traditionally been implemented as Hidden Markov Models [20] however there is an increasing interest in using Deep Neural Networks [7] for greater accuracy and performance.

CHI 2015, April 18 - 23, 2015, Seoul, Republic of Korea.

ACM 978-1-4503-3145-6/15/04…$15.00.
http://dx.doi.org/10.1145/2702123.2702532

The language model provides the probability of a specific sequence of words occurring given a particular form of speech – such as news, lectures or conversation. This is used in conjunction with the output by the acoustic model to identify the spoken phrases which are most likely to be correct [8]. Constraining the vocabulary of the model is used to increase the identification rate for systems like telephone interfaces [4] and controlling subsystems in a car [13]. Larger, yet still constrained, models are used for tasks such as Internet search and calendar management [22]. The advantages of such a system are that with a tightly defined language model recognition rates can be greatly increased. Constraining the vocabulary can improve the accuracy between 50% and 80% [4].

A much more difficult task for speech recognition is to consider less constrained speech, such as inter-personal conversations, where the language model is expected to encompass the totality of the language being spoken and the speech itself is more natural in both content and context – meaning that the speech will contain fillers, such as 'um', words that run together, and the audio stream itself will include more background noise [2]. In such situations the error rate rises considerably – however it is improving with Kingsbury *et al* [11] reporting an accuracy of 32% in 2001 and Povey *et al*, reporting 55.7% in 2011 [19].

Speech in HCI

Speech has been used as an input in HCI in a number of ways. The most common are dialogic systems where the user interacts with the system as part of a conversation, such as Siri. These are now commonly available on consumer devices. There has been research on improving the interface for speech dictation, by making it more immediate and interactive [14] and controlling a traditional touch and click based UI using speech input has also been tried in systems such as SpeechAct [28].

Using speech as a secondary input for information retrieval, for example automatic transcription and indexing of voice mails [25] and meetings [24] has received considerable research. Beyond transcription, problems such as topic extraction, detection of conversational 'hot spots', and recording speaker decisions have been addressed [23]. Speaker diary recording and role recognition can provide information on the number of people in a conversation and show the flow of conversational control from one to the other. Speech audio can also be examined to label instances of fillers and laughter. Systems where the speech is used in the background are rare, however one such example is the Ambient Spotlight [10] where supporting documents were retrieved to match the current topic in an ongoing meeting. It is also possible to perform tasks such as 'mining' sociometric data [27] to understand the social connections of subjects. This has been done with continual listening on mobile devices to extrapolate the audio levels and sound characteristics of social activity and to detect patterns of social interactions between users [6].

Working with Recognised Speech

Building on work that process the audio into text, and annotating it with other features, it is also possible to employ natural language processing to analyse the text extracted. There are many approaches that can be used, in particular text summarisation approaches, covered in detail by Lloret & Palomar [15], which support extraction of key phrases and words from text. One simple approach is to extract the *subject* term from sentences. This can be done with a fairly high accuracy – over 90%.

THE CONTINUOUS SPEECH STREAM (CSS)

As noted by Aylett [3] there is an apparent divide between HCI practitioners and speech technologists – however they also note that there are opportunities to improve user experience by using speech technologies. One identified area of research that has had less focus is how a system might make use of users' conversations in the background. That is to say, using spoken conversations that are not directed at a device but that are picked up incidentally by listening to users' interactions with others.

Our interest in this was sparked by earlier research we conducted which used mobile phones to record audio and screen interactions with the phone. Even though our intent was to understand interaction with phones, we found that we also incidentally collected surprisingly legible person-to-person audio from conversations conducted around the phone. This led us to consider whether speech could be extracted from background audio and used as a resource for systems and applications. For example, could an application detect when a user might run a search on their phone from their conversation. This would require a speech recognition system that could detect *topical resources* from conversation, grammatical structures such as *objects, subjects,* and *nouns.*

Exploring the Technology

We conducted two tests – one based on previously recorded data of natural speech in the wild, and another on an open source speech corpus. The situational audio data we used in the first test was recorded by the microphone on a GoPro camera around the neck of a participant while in conversation recorded from a previous user trial. This audio was segmented into 2 - 10 second clips of speech, based on voice activity detection carried out using the Bob library [1]. Each of these clips was submitted to the, undocumented yet public, Google Speech API. Submitting our conversational clips, totalling 56 minutes of spoken audio, returned either an indication of no speech detected, or incorrectly recognised text. The clarity of speech in the audio for ASR needs to be much higher than could be processed from our source, which included environmental audio from the city as well as the speech. The very nature of the speech we were trying to recognise – spontaneous, bursty and often unstructured – was inherently difficult for ASR to transcribe accurately suggesting that a keyword spotting approach may provide better results.

For our second test we used an open source corpus of human speech, the VoxForge database (voxforge.org), with the Sphinx (speech.cs.cmu.edu) open source speech recognition system. We compared giving a list of keywords to the Sphinx recogniser (which were used in the recognition algorithm), with a keyword spotting system based on using Google Speech API and filtering the output to only include the keywords. This keyword list comprised of the 30 most common bigrams and trigrams (2 and 3 word phases) in the VoxForge dataset.

When processed through the Google Speech API we found 28% of the keywords in the source were detected, with a 83% precision. Using Sphinx gave performance in a range between a 63% detection rate with 24% accuracy and a 35% detection rate with 64% accuracy.

Speech recognition benefits greatly from one dedicated microphone per speaker, increasing the signal to noise ratio (SNR), which has a large impact on the accuracy of the recognition. Training on a particular speaker also increases accuracy, suggesting that a personal approach should be taken excluding others' speech. However, with the large amount of investment going into applications such as Google Now, Siri and Cortana, we are confident that the technical challenges for improving recognition of human-human communication are being met with advances in fields such as noise-robust and missing-feature ASR.

Defining the stream

While a CSS system could be possible, it is still beyond the publically available state of the art. To explore the design space we experimented with a two stage process for working with audio. In a working system ASR would provide a transcript of the spoken words and phrases. As noted above, the quality of the audio, volume of the speech and training on a particular person's voice has a great impact on the accuracy of the automatic speech recognition. Taking this into account we decided that only the owner of the device would have their speech transcribed, background conversations and conversation partners would be ignored in their stream. For the second stage we determined that natural language processing would be used to extract *subjects* and *objects* of speech, filtering out personal pronouns, for inclusion in the stream. We also decided to include *time phrases* and *numbers* wherever they appeared grammatically given their utility. Such processing of an ASR transcript we deemed to be computationally feasible. We note that the attributes of the ASR and the NLP stages could be tuned to include or exclude a number of attributes – grammatical, aural or statistical – however this definition was compatible with both our goals for design and privacy. Not providing a full transcription, not promising accuracy of recognition, not providing timestamps – all of these were to introduce ambiguity, uncertainty and therefor deniability into the system.

METHODS

For our experiment we had three research questions:

1. Can information be extracted from the continuous speech stream which is representative of the users day?
2. Are there common system actions (such as search) that could make use of terms from the speech stream?
3. What applications more broadly could we build that would make use of the speech stream?

Our first step was to obtain a collection of continuous speech streams, recorded from a variety of individual's ordinary days. Participants wore a lapel microphone mounted onto their clothing, with small digital audio recorders worn on the belt or in a pocket. We recruited ten participants who agreed to carry audio recorders over the period of a day, recording their spoken conversations. For privacy reasons we allowed participants to stop and start the recorder at will, and also offered to delete any sections from the audio that the participants did not want to share with us. This provided us with a small but rich dataset of naturalistically recorded speech. Six of our participants were recruited via a dedicated participant recruitment website and compensated with a cinema gift card, the remaining four were recruited from our extended social network. Participants ranged from an architect, an elementary school teacher, a brand consultant, a special needs carer and a mystery shopper. All participants were requested to inform those around them that they were recording and to wear the microphone openly during the day. Four participants recorded a day during which they were not working, six participants recorded a weekday comprising of work and after work.

The worn microphone setup meant that we did record talk that was not by the participants themselves, although for the experiment we focused on the speech of the main participant – simulating the accuracy enhancement of a speaker dependent system.

We manually simulated two technical stages – speech recognition and subject extraction. It was likely that our manual approach provided a much more accurate extraction that could be obtained automatically so we would be experimenting with a 'best case' scenario of speech recognition and language abstraction.

This produced a word list from the day for each participant, covering all the topics spoken about over this period. This let us provide a summary that could be used by our participants to reflect upon their day and the contents of the speech stream without having to read a complete transcription. These words also represent an automatic analysis of the speech stream that systems could potentially be built upon.

We interviewed the participants using the word list as prompts. This let us explore participants' reactions and how the list summarised their day by asking them to describe

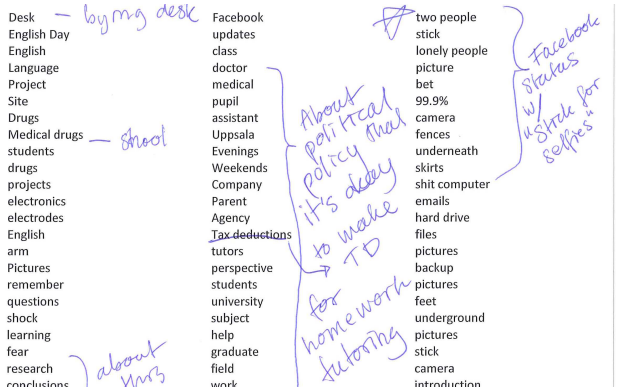

Figure1: Sample CSS annotated during participant interview

their actions and interactions first without, then with access to the list. We also explored any system actions they might want to be automatically performed on those words. We covered how the participants would feel about a system listening to their speech – and how this related to the recordings we were carrying out for the trial. We then discussed the participants experiences in using a mobile device in conversation, and the relationship between their spoken conversations and any actions they carried out on their phone (such as searches, diary entries or phone calls).

Lastly, for seven of the participants we asked them to look through a sample of other participants' word lists and to describe how they perceived the events taking place and the personality and demographic of the subject.

Trial results

The longest day recording was of 12 hours, with a mean of 7.4 and a median of 8.5. The number of words processed ranged from 287 to 1,405 words, and from 228 to 1,291 unique words (median 525, average 585) recorded and transcribed into the sample streams.

The transcribed items were then coded for topical and semantic content. An initial set of 500 randomly selected items was independently coded by two researchers using a data centric technique allowing for classifications to emerge from the data naturally. A coding scheme was then agreed upon containing 11 categories, into which all unique items were independently coded. Mismatches were discussed and resolved after the coding was completed.

The categories, with their percentage of unique keywords in parentheses, are *Action (7%), Body (6%), Communication (4%), Food (9%), Leisure (8%), Numbers (6%), People (15%), Place (11%), Items (10%), Time (7%), Work (12%), and N/A (7%).* Just from the initial spread of items we can see some patterns emerge. In talking about people they, understandably, each have their own names which are unique to our list and increase the size of that category. Work as well contains a lot of domain specific items.

Looking at the wordlists individually we can see contextual domination of categories, for example one participant who recorded a day of picnicking with friends had a wordlist containing 24% in food and leisure, a second participant on a workday in his architecture firm produced a list containing 23% work related terms, with a further 9% related to places. Our recruitment agent participant produced 32% on work and people, and one participant who recorded a romantic evening returned a list with 37% of the 1,405 items taken up with leisure, people and places.

We counted a total of 2,502 unique English words and while this is somewhat large for a speech recognition keyword dictionary, dictionaries in the thousands of words have been shown to be successful in recognition [17]. This suggests that building a keyword list for each user alongside a model of their speech would lead to increased precision over time. The word lists give an overview of what was talked about by the participants on these days.

Interviews

The interviews were conducted for between 40-90 minutes and explored their use of mobile devices, the interpretation of the stream with regards to their memory of the day in question, any likely straightforward system actions they would want to carry out on stream words, and their interpretation of the stream of another person. Participants were encouraged to annotate their CSS as they discussed their recorded day with us, the results of which can be seen in Figure 1. All the interviews were fully transcribed.

We look at the results of the interviews in five sections by topic. First we examine their experiences being recorded. In the second section we focus on *using the stream.* Following that we discuss the implications for and effects of *memory and recollection* and the different *activities and people* that the stream helps to identify. We finish this section with an analysis on how the *streams of others* are interpreted.

Experiences with being recorded

Due to the self-selection inherent in any trial such as this it may be unsurprising that the general consensus of the participants was that being recorded was not a great cause for concern. However, on reflection a number of our participants reported some lingering awareness of the recording taking place, with a couple reporting a change in their activity as a direct result. One participant consciously didn't say the name of a confidential client during the recording, even if he was working on this client's project for the whole day.

"I didn't say the name, because it is, the client is a secret client, but Oliver [his co-worker] said it all the time."

There was also some concern as to the projected image of themselves that would be produced by a transcript of everything they say in a day.

"I was a little apprehensive, because I thought, Oh God, everyone's going to know how stupid I am all the time."

But most of our participants reported some degree of normalisation to the process, with half of them reporting that in some way the recording device was forgotten at least

Action	Number	Examples
Search	53	Blister, Suits
Add to calendar	24	Sushi, happy birthday
Show map	1	Mariatorget
Remind later	23	Calligraphy, check email
Others	3	Skybar, Humlegarden

Table 1: Desired Actions Based on Wordlist

for part of the time. Most of our participants discussed the recording taking place with the people they were talking to throughout the day, partly as a way of explaining the visible lapel microphone and partly as a disclosure that the other party may be recorded – although the speech from anyone but the participant was not transcribed it was at least partly listened to in the production of the stream. Only one participant reported that they were requested to stop the recording, the reactions from others were mostly ambivalent.

It was also noted by our participants when reviewing their streams that the periods of time without any keywords, when they were working or alone, were longer than expected.

Using the Stream

Our initial design ideas for the use of the CSS centred around better supporting interaction with the mobile device during speech. We ask participants if there were any terms of the list that they would like to have their phone automatically process in some way – such as showing a search result from a term that was spoken. We also asked for items to potentially place in their calendar, for places that could be viewed on a map, for items to be reminded of, and lastly any for other automatic functions they would want to be performed based on an item in the list. An overview is shown in Table 1.

All of our participants expressed a hesitance to use their mobile devices in conversation. While they had all experienced the phenomena of either using their device or having another party use a mobile device in a conversation which they were a part of it was universally described as something that they didn't want to do – let alone encourage or make easier.

The concept of delayed interaction with the device, queuing up topics for later enquiry, was more acceptable to our participants. However, presenting the list of words and asking for likely actions resulted in a fairly small set of selected terms. Moreover, many of the terms that were selected for phone-based operation were ambiguous and it is difficult to see how they could clearly map to automatic actions. For example, "happy birthday" refers to a date that the participant would have wanted recorded in their calendar, but in the wordlist itself there is no record of when that date would actually be. That said, around 10 items were selected to be searched on per day by each participant, representing a potential use case.

Memory and Recollection

There has been a longstanding interest in systems that can help with memory formation, but also in recollection of events, notably with the work carried out using the SenseCam [9] – an automatic wearable camera. In the interview the participants were all asked to first describe their day including who they talked to and what they talked about, they were then given the same task with access to their personal ambient speech stream from that day. Every participant drew on the words to illustrate and narrate their day, describing events that they said they would have missed without the list of terms. However this was not a straightforward reading of the keywords and describing what had happened, there was considerable work done by the participants to reconcile what was placed in front of them with their memory of the day in question.

Again, this is similar to the SenseCam [9] studies, and our participants reported that they remembered more about their activities during the day simply as a result of the novelty of being recorded and taking part in the trial. Our participants also recalled more events when recreating their day through their stream than before they were exposed to it.

"I missed so much in my original summary of my day."

Much of the stream was recognised and supported the participants' memories of the day in questions. There were three different types of supported recall that we noticed in our participants. The most simple was the recollection of an interaction that had been forgotten.

"...even ones that like 'enormous worm' and I was like what is that? But then I obviously remembered the funny moment."

We also saw examples of participants correcting their memories in various ways. The order of events, as noted by Sellen, was something that was plastic in the memory of our participants. One participant reported surprise that a particular conversation happened on that day, and using this information was able to also correct her recollection of the person to whom she had told that particular story.

"Oh yeah, hang on, I think I was telling, was that that day?"

The length of time certain topics were discussed was also something that we saw corrected as our participants re-interpreted their day through the lens of the keywords. One participant in particular reported a conversation over lunch where he was talking in derogative terms about his boss alongside discussing 'life, and everything' with his friend. When he examined the keyword stream of this particular conversation he was forced to re-evaluate the conversation:

"I cannot believe it. It's crazy, we talk so much about work."

This could be that the more enjoyable, interesting and uncommon conversation topic became the most memorable part of the lunch, superseding the work conversation. This

is also an example of how access to a record of the conversation caused a re-evaluation of the experience. Up until this point the conversation over lunch was described positively, a pleasant memory of a wide-ranging conversation with a friend. After the speech stream was interrogated this experience became one of reflecting on the amount of time and energy spent on work related topics even out-with the working environment.

"Oh, now I remember where 'delicious kids' came from […] we were talking about eating their artwork, 'Delicious, kids.'"

This final example shows the ambiguity of meaning in the CSS where the removal of large parts of the speech and timing has changed the meaning at first read.

Activities and People

We noted that certain terms were repeated by a number of the participants, on a global level these were related to their dominant activity from their work, study or hobby – individual words like *mascara* for the cosmetics brand manager or the word *facade* for the architect, or cooking topics for the school teacher (who cooked as a hobby).

These terms could also be used to identify conversational partners – participants talked about how particular topics went together with those they were talking about:

"I was back to Eva because of the botox […] that was the discussion with Joel that always turns around the same thing."

Conversations surrounding work were littered with jargon or codes for some while for others revolved around specific topics – yet for all the participants that used the recording device at work the identification of these portions of the stream was easy. They were also able to assign differing clusters of keywords to the same anecdote recited to different people. Specific pieces of information relating to activities were identified, with features such as starting times, durations and locations easily tied to a particular event – be that one in the planning stage, or one that took place on the day of the recording. A clear example of this was before meal preparation or eating, where the stream would contain numerous food related words.

Lastly, lists of options were also easily identifiable from the stream. The particular conversation pattern consisting of listing a number of options of the same genre for consideration by the other party produces a dense and informative section of the stream showing the options, then their narrowing until a choice is made.

These lists are different from the patterns of keywords that identify the pre-purchase activities of evaluating the relative merits of one purchase vs. another. This follows work in marketing which has discussed how products and advertisements can come to have an involvement in conversation [21]:

"if I'm talking about something to have a good offer come up […] usually when I Google and those first sponsored links, I use a lot."

Streams of Others

From the second round of interviews the participants were shown the stream of others, what was most interesting about this exercise was the manner in which our participants approached the problem of interpreting the partial streams of others.

"This must be a girl that right here. It's the first person. It looked like it's a person that is a bit stressed about the age because talking about Botox all the time and her legs."

They worked to simultaneously build a model of the user who created the stream and to understand the context around the words presented to them through that model.

Lessons

From these wide ranging observations and analysis we distilled a number of important characteristics of the CSS to be fed back into our design dialog.

- Activities and interests were identifiable where prospective actions such as search were not.
- Both people and activities have identifying keyword clusters.
- Lists of options produced by planning future activities are easily recognizable.

DESIGN PHASE

While these interviews and the topic lists gave us some insight into the speech stream itself, our goal was to explore future design opportunities using the stream. Accordingly, we ran two design workshops based around the data. In the first we asked five professional designers (1 digital design team leader, two post-doctoral industrial designers and 2 junior designers) to generate design concepts, using the topic lists as input to imagined systems. This workshop involved: (1) reading through the word lists to understand the technology, (2) comments on the word list (3) concept generation around systems that could use the lists (4) discussion of possibilities for sharing lists. We selected stories from the interviews as prompts to the design, printed on cards, to seed the designs in the workshop. In the second design workshop we repeated this method ourselves, again going through the word lists and interview data to suggest potential systems drawing on the interviews and further refining the design concepts that arose in the first design workshop.

We categorized the concepts these workshops generated into three different categories based on who the stream of the user is shared with. *Personal* designs are those where the stream is kept between the user and the application or service provider. *Shared* designs are ones where a stream is published, uni- or bi-directionally. Lastly, *distributed* designs are those in which a large number of streams are aggregated, transformed or shared to provide a service.

Personal Stream Designs

We outline three sections of this area and give examples of applications of services in each to better make our case that the proposed ambient speech sensor stream is an interesting

design proposal and worthy of future research and development.

Pre-Input

The first ideas we generated were around the use of these streams as a resource for pre-input, that is using the words heard in conversation to prepare for device use by either pre-launching applications, pre-searching for information or images, or preparing calendar entries for use when the user turns their attention to the mobile device. Terms used in speech could also be used for auto-complete fields used when entering text. For example, a navigation application could listen for addresses spoken and offer these as options when entering a destination address. Words spoken could also be used to manage the text entry dictionary on a phone keyboard, and to offer to autocomplete words that have been spoken recently.

Memory Aids

One of the most straightforward uses of this information would be in the area of memory aids, life logging and quantification of self. The amount of conversation taking place, the topics (work related or personal), and the counts of certain words can be used for a number of self-reflective purposes. The use of images recorded though out the day using a wearable camera as a memory aid have been shown to be successful [9], extending this to the conversations people have would be an interesting comparison. Such word counts and graphs can also be used in language learning, to quantify progress and suggest areas of vocabulary for further study. However in this space there are also more performative and playful applications for this technology that can be explored. Extending the language learning example to provide real time synonyms and antonyms could be used as an enabling technology for the production of amateur rap music, and – possibly more practically – the stream surrounding taking of a particular photo could be embedded within the photograph in tags, or even visualised to give additional context to the photo.

Qualified Self

While there has been recent interest in techniques for quantifying activity in an array of sense-able areas, in contrast the CSS offers possibilities for systems that can *qualify* that experience. We might call this the 'qualified self' – models and systems that make use of the concepts and terms spoken as part of our activity, rather than reducing this to what can be easily measured. So, for example, relationships can be examined for their qualitative attributes, such as the common topics of conversation and their similarities to other relationships, as well as their quantitative qualities of duration and frequency. A system could even offer suggestions for when a conversation became difficult, or was likely to engage with difficult topic areas. Since there is the potential with the CSS to track the semantics of activities one could also imagine a system attempting to intervene if a user started on activities they had declared they wanted to reduce. More playfully, we could imagine systems that record experiences to present personal collages of activity, or to expand ones social network through tracing and suggesting new activities.

Shared Stream Designs

Our second set of concepts focused on applications that could exploit sharing the stream with other users.

Awareness

We envisage two main ways in which the awareness of others can be enhanced using the ambient speech stream. First would be the active sharing of a length of time in the style of the "Glympse" (http://www.glympse.com) application for location sharing. A user could send the stream of their current situation to a loved one as part of an explanation of their lateness or to share a meeting with mutual friends with one who is not present. The second concept we explored was active reading, where the participants of a meeting or class could allow their streams to be accessible during that time to others on the class list. This would allow those arriving later to avoid topics that have been discussed, or prepare for what is currently under discussion in the room as they are arriving.

Subscription

The nature of the word lists we experimented with is that they condense and remove many details from what is spoken, which introduces some inherent ambiguity but also mediates some of the privacy concerns our participants had. This opens the possibility of subscriptions to the streams of famous people, giving a personal glimpse of their life that is filtered through the persona projected by them in the media. The sharing of not-so-famous people can also be envisaged, given the opportunity to self-publish feeds people could follow a group learning the same language for vocabulary tips or a fresh produce buyer for information on what is in season. Friends and family could also be subscribed to in order to increase awareness and connectedness, with simple filters set up to alert when two people are talking about the same topic or say each other's name for example.

Performance

The performative aspects of creating such a stream also offer opportunities for design. Along with the sharing of snippets of the stream from certain personal experiences such as birthdays or sports events and attaching them to other media users have the opportunity to manipulate their streams by altering their speech. This could be used in any number of ways, from showing your affiliation to a genre of music or a particular band to friendly, playful competitions to see who can manipulate their stream to match those generated by inputting the lines of a movie from their favourite character.

Distributed Stream Designs

Lastly, we discuss the opportunities afforded when a number of streams are shared and aggregated to provide a service. These are not only end user systems, but can provide value for the corporations producing or managing the streams.

Matching & Prediction

Chalmers [5] points out that the order and temporal structure of activities can provide a deep insight into the character of the user and pointers to predict use. While in that work the examples were generally confined to location, it is possible to expand this to include a historical vector of context built from the topics and keywords identified in an ambient stream. The main problem identified by Chalmers' was that sparsity of data could have a negative affect on the quality of recommendations. Using a corpus of many users' speech streams the comparison of vectors of topics and keywords would allow recommendations to be made of topics those with similar conversation trajectories had touched upon. This matching could also be used in a more playful, or possibly profitable, manner by introducing these sensors to the online dating community. An application could use the number of overlapping topics of conversation as a metric for the compatibility of a potential partner.

Perhaps the most obvious application for this sort of recommendation is advertising, in that the speech stream could be used to target or personalise advertisements. While few might be happy about advertisers having access to their private conversations, it might be possible to compensate a sample group who could be recorded. Through correlations with other data, such as websites visited, connections between the sample group's spoken topics and already widely tracked data could be found. We acknowledge that targeting advertising based on the contents of private conversations would be considered by many (ourselves included) as somewhat sinister, regardless of any potential benefits. We thus need to proceed with some caution around applications here, even if end users are properly informed or compensated.

Events tracking

A more straightforward application area is for the tracking and identification of events and users activities using their spoken conversation. So, one of the clearest signals in the word streams was before a visit to a restaurant where multiple food related terms could be detected in the stream. Similarly before activities (such as sport events) there would be a cluster of words related to the event itself. One obvious suggestion from this would be the possibility of offering suggestions or making recommendations for particular establishments at an appropriate time. With access to multiple users streams a system could even predict future congestion by multiple users attending the same event, or going to the same restaurant.

Similar to our earlier playful suggestion of being able to follow a friend or partner's stream we could imagine being able to follow the stream from a particular place. The text stream could aggregates the conversations taking place – providing for those who do attend a record of their interaction with a particular event and for those who are not in attendance a glimpse into the atmosphere through their projected experience of attending.

Talkthrough

One of the most interesting possibilities for even a small percentage deployment of such a system would be the ability to extract information on the topics of conversation in an area to form a new type of demographic. For advertisers this would lead to a new metric, the "*talkthrough"*. It would allow those serving adverts to be alerted to the number of situations where keywords from their advertisement, be they the brand name or a other salient feature of the product, have entered the conversation of the viewer. Moreover, the detectable pre-purchase routines discussed earlier would allow advertisements to be specifically targeted to those who are in the market for a competing product and to track if that consumer considered, talked about, or even bought the alternative.

On a similar note, the action of polling an area or demographic for a sentiment could be augmented by allowing polling agencies to purchase the number of occurrences of certain words or phrases in a specific area or within a specific demographic group over a period of time. This could be used to measure statistics such as the dissemination of marketing campaigns or the political engagement of voters on a particular topic.

PRIVACY

There are a number of areas of concern in this regard when the deployment of the underlying technology that would enable the systems outlined above is considered. That our self-selected trial participants voiced some discomfort with being recorded suggests the problems that such systems might face with regards to the concept of being listened to at all times. However we note that systems listening continuously seems to be acceptable, at least to some, with this functionality now commercially available in both Android and iOS platforms. The recent release of commercial home appliances which continuously listen such as Amazon's Echo and The Ubi (theubi.com) not only signposts the acceptability of such listening services to consumers, but highlights the need for research in understanding the implications of their deployment. Confidential information being inadvertently disclosed during the use of this system was one such concern, with one of our participants modifying his speech to not include the name of a particular client. In mitigation of this risk a system could offer users control through word, concept filter or context [12] filtering, or by obscuring new or unexpected content in some way until it is approved.

There is considerable ambiguity in the keyword stream, so on one hand the potential for a user to be held accountable for a particular topic of conversation is lower than it would first appear. This, of course, is taking the continuous speech stream as an isolated data point. When combined with other sources of data it provides yet another point of triangulation. For example, the combination with location data and time could add enough context to a keyword to strongly suggest an action. The other side of the ambiguity

of the stream is the potential for topics, or even errors in recognition, to be taken out of context causing concern and embarrassment:

"Porn […] I think we were perhaps joking about something, but I can't remember exactly […] I really wouldn't want to share that word with family."

This could be to do with the novelty of such close and continuous surveillance, however it has been shown that these concerns can drop over time – even among those opposed to such surveillance – as familiarity with the technology increases [18].

Broader Concerns

More broadly, the prototypes suggested here are in some senses personal surveillance systems. The use of the distributed system not to help pollsters and advertisers, but to alert government authorities to the people and places involved when 'subversive' discourse is taking place is easily envisaged. It is even possible that such technologies are already used by security services. For example, the electronic privacy information centre used a freedom of information request to obtain a list of 'trigger words' that the US department of homeland security uses as signs of terrorism. There are also organisations, such as Stratign (statigen.com), that sell GSM intercept systems that both break GSM encryption on calls and watch for particular 'trigger phrases' in speech or text. These concerns do not stop with official access to the data, the service provider using the data for a purpose unforeseen by the end user is another risk that must be mitigated.

The everyday speech of the user is more personal a resource than either location or app usage. Not only must the lessons learned in giving users graduated control of there exposure be applied in this area, the underlying implications that the device, and through it the service provider, is continuously listening must be dealt with honestly and openly if there is to be any chance of these technologies being accepted. We plan to include a participatory design phase in future work to better understand both users' current understanding of their aural privacy, and how we can design to empower them when it is threatened by emerging technology. The idea of control and empowerment of users through ensuring their understanding of logged location data and use data has been positioned [16], and expanding this work to more potentially invasive systems is also a goal for the future.

CONCLUSION

In this paper we have explored the design opportunities presented by a continuous speech stream. By manually producing 10 such naturalistic day-long streams we explored how the topics of everyday speech relate to the actions of our participants. We then, though a process of design workshops with designers, developed three areas for further research and application development. In future work we propose to move further on the exploration of these areas, starting with the Personal Stream area of applications. For applications of this sort it would be possible to conduct a real-time Wizard of OZ user test of any produced applications, exploring in detail how the pre-input and pre-search applications would be used either during or after conversations.

We have discussed the feasibility and practicality of deploying such a system, and while it may seem difficult to justify the development effort for any one of the applications outlined above, the opportunities presented when they are taken as a suite of possibilities are much greater. The possibility to track the spread and flow of conversation topics would give a new and exciting metric for the understanding of how topics, brands and idioms are placed in the public consciousness. By leveraging this against the consumer facing applications of this technology the expense of development and deployment becomes a much more attractive opportunity.

We have opened up the design space by projecting a very ambitious goal for speech technology, however with regards to the call for HCI researchers to bound their expectations of speech technology put forward by Aylett *et al* [3] we note that each of the designs for systems put forward above could be build upon a specific, and therefor more tractable, speech technology solution.

Despite the privacy and ethical concerns of an entity having access to a significant proportion of the conversations spoken by their users on a regular basis, the opportunities presented by speech recognition technologies go far beyond the simple dialog-based interactive systems that are the mainstream today. In this stream there is a possibility to get access to and provide much more reflective and personal services in a way we have rarely seen before. With such a personal, nuanced and malleable source – the day to day conversations of the user – the envisaged speech stream would enable whole new genres of services and applications from the playful to the powerful.

REFERENCES

1. Anjos, A., El-Shafey, L., et al. Bob: a free signal processing and machine learning toolbox for researchers. In *Proc. of the 20th ACM international conference on Multimedia* (2012),1449-1452 ACM.
2. Anusuya, M. A. K., S. K. Speech Recognition by Machine, A Review. *International Journal of Computer Science and Information Security*, 6, 3 (2010), 25.
3. Aylett, M. P., Kristensson, P. O., et al. None of a CHInd: Relationship Counselling for HCI and Speech Technology. In *CHI '14 Extended Abstracts on Human Factors in Computing Systems* (2014),749–760 ACM.
4. Boyce, S. and Gorin, A. User interface issues for natural spoken dialog systems. In *Proc. of The International Symposium On Spoken Dialog*, 96(1996), 65–68.

5. Chalmers, M. A Historical View of Context. *CSCW Journal*, 13, 3 (2004), 223-247.
6. Chon, Y., Lane, N. D., et al. Understanding the coverage and scalability of place-centric crowdsensing. In *Proc. of the 2013 ACM international joint conference on Pervasive and ubiquitous computing* (2013),3-12 ACM.
7. Hinton, G., Deng, L., et al. Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. *IEEE Signal Processing Magazine*, 29(2012), 82-97.
8. Jurafsky, D. and Martin, J. H. *Speech and Language Processing (2Nd Edition)*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 2009.
9. Kalnikaite, V., Sellen, A., et al. Now Let Me See Where I Was: Understanding How Lifelogs Mediate Memory. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems* (2010),2045–2054 ACM.
10. Kilgour, J., Carletta, J., et al. The Ambient Spotlight: Queryless Desktop Search from Meeting Speech. In *Proceedings of the 2010 International Workshop on Searching Spontaneous Conversational Speech* (2010),49–52 ACM.
11. Kingsbury, B., Saon, G., et al. Robust speech recognition in Noisy Environments: The 2001 IBM spine evaluation system. In *Proc. of 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing* (ICASSP) (2002),I-53-I-56.
12. Klasnja, P., Consolvo, S., et al. Exploring privacy concerns about personal sensing. In *Proc. of Pervasive Computing,* Springer (2009), 176-183.
13. Kuhn, T., Jameel, A., et al. Hybrid in-car speech recognition for mobile multimedia applications. In *Proc. of Vehicular Technology Conference*, 1999 IEEE 49th (1999),2009-2013 vol.2003.
14. Kumar, A., Paek, T., et al. Voice Typing: A New Speech Interaction Model for Dictation on Touchscreen Devices. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems* (2012),2277-2286 ACM.
15. Lloret, E. and Palomar, M. Text summarisation in progress: a literature review. *Artificial Intelligence Review*, 37, 1 (2012), 1-41.
16. Morrison, A., McMillan, D., et al. Improving consent in large scale mobile HCI through personalised representations of data. In *Proc. of the 8th Nordic Conference on Human-Computer Interaction* (2014),471-480 ACM.
17. Nouza, J. and Silovsky, J. Fast keyword spotting in telephone speech. *Radioengineering*, 18(2009), 665–670.
18. Oulasvirta, A., Pihlajamaa, A., et al. Long-term effects of ubiquitous surveillance in the home. In *Proc. of the 2012 ACM Conference on Ubiquitous Computing* (2012),41-50 ACM.
19. Povey, D., Burget, L., et al. The subspace Gaussian mixture model—A structured model for speech recognition. *Computer Speech & Language*, 25, 2 (2011), 404-439.
20. Rabiner, L. and Juang, B.-H. An introduction to hidden Markov models. *ASSP Magazine, IEEE*, 3(1986), 4–16.
21. Ritson, M. and Elliott, R. The social uses of advertising: an ethnographic study of adolescent advertising audiences. *Journal of Consumer Research*, 26, 3 (1999), 260-277.
22. Starner, T. E., Snoeck, C. M., et al. Use of Mobile Appointment Scheduling Devices. In *CHI'04 Extended Abstracts on Human Factors in Computing Systems* (2004),1501–1504 ACM.
23. Tur, G. and Mori, R. D. *Spoken language Understanding: Systems for Extracting Semaintic Information*. John Wiley, 2011.
24. Waibel, A., Schultz, T., et al. SMaRT: the Smart Meeting Room Task at ISL. In *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing* (2003), vol.754.
25. Whittaker, S., Hirschberg, J., et al. SCANMail: A Voicemail Interface That Makes Speech Browsable, Readable and Searchable. In *Proc. of the SIGCHI conference on Human factors in computing systems* (2002),275–282 ACM.
26. Woodland, P. C. Speaker adaptation for continuous density HMMs: A review. In *Proc. of the ISCA Tutorial and Research Workshop on Adaptation Methods for Speech Recognition* (2001).
27. Wyatt, D., Choudhury, T., et al. Inferring Colocation and Conversation Networks from Privacy-sensitive Audio with Implications for Computational Social Science. *ACM Transactions on Intelligent Systems and Technology*, 2(2011), 7:1–7:41.
28. Yankelovich, N., Levow, G.-A., et al. *Designing SpeechActs: Issues in Speech User Interfaces* (1995), 369–376 ACM Press/Addison-Wesley Publishing Co.