MODALITIES AND DEMOGRAPHICS IN VOICE SEARCH: LEARNINGS FROM THREE CASE STUDIES

S. Chang, S. Boyce, K. Hayati, I. Alphonso, B. Buntschuh

Tellme, a Microsoft Subsidiary
1310 Villa Street, Mountain View, CA 94041
{shchang, suboyce, katiah, issaca, brubunt}@microsoft.com

ABSTRACT

We present case studies of three different approaches to voice search for business information: Premium DA, Fully Automated Business Search, and Multimodal Search. In each case we describe the user demographics, business economics and technology limitations that drive user interface and system design. Findings from statistical data analysis of real user utterances are highlighted with discussion to their implications in speech recognition, back-end search and voice user interface design.

Index Terms— directory assistance, voice search, multimodal, category search, speech recognition

1. INTRODUCTION

Broadly speaking voice search has been around for many decades, ever since the beginning of directory assistance (DA) services offering by telephone companies. However, it is only in the past few years that the paradigm has started to shift towards increased automation and new modalities, as a result of fast development in technologies such as speech recognition, search, and computing power in general. This paradigm shift has posed a significant challenge to the design and development of automated voice search applications [1][2][3]. One aspect of that challenge has been balancing the needs of users with limitations in the technology while satisfying the overall economics required by the business. In this paper we highlight our findings from three different approaches to voice search for business information; Premium DA, Fully Automated Business Search (FABS), and Multimodal Search (Phonetop). Each of these search products has dictated differences in user interface design and recognition strategies, while relying on the same underlying search back-end.

In the following sections our experiences with each of the approaches to voice search are discussed in three case studies. We first describe the user and business characteristics that drive our interface and system design. Some findings from analyzing a large amount of user utterance transcriptions are presented. More general implications to speech recognition and back-end search are also discussed.

For each application speech utterance data were randomly collected roughly in the same period in the summer of 2007, and include the first locality and listing requests from about 10,000 unique calls or user sessions. These data were used to generate all tables and figures in this paper, unless otherwise specified.

2. PREMIUM DA

Tellme provides automated directory assistance for phone companies in both the landline and wireless business. In these deployments virtually all calls are first routed to an automated dialog that prompts the caller for city and state, and then for the listing name. If the automation fails to successfully find the requested listing, the call is routed to an operator for further assistance.

There are many characteristics of the DA business and caller population that have shaped the current Premium DA user experience. First, most callers are paying for DA on a per call basis. Charges vary across landline/wireless and across companies but they can approach $2 per search. Hence, Premium DA callers demand that their requests be handled quickly and accurately. Second, since DA is offered ubiquitously across phone companies' region, the caller population is highly varied. DA is a service that phone companies have been offering for more than 100 years. Older Americans are extremely familiar with its use. As such automated Premium DA services need to be designed to meet the needs of a highly varied caller population. The callers vary widely in their experience with technology and their willingness to tolerate automated dialogs over the phone. Landline DA callers tend to be older, are more likely to be women, and on average tend to have a relatively low adoption rate of technology. Figure 1 shows a comparison of gender distribution between Premium DA (landline callers only) and other voice search applications. The aspect of Premium DA that callers value most is its convenience. They can dial *411* and get their answers back quickly.

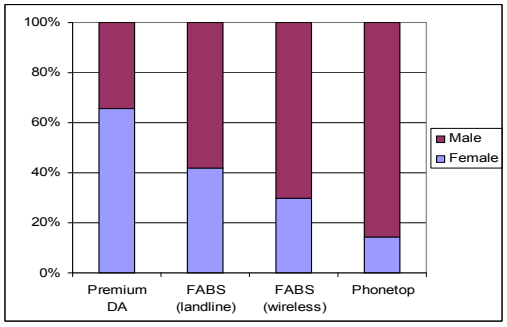

Figure 1 *Gender distribution in Premium DA, FABS (landline callers), FABS (wireless callers) and Phonetop.*

To design an effective user experience the needs of the business and needs of the users had to be balanced. The resulting design is one that is fast and efficient in its prompting – mimicking the words that have been traditionally used by operators ("city and state please", "what listing"), but passes the caller to an operator for difficult requests. The automated system has one chance (or at best two) to recognize correctly before the call is handed off to an operator. If an operator is required he/she finds the listing and then hands the caller back to the automation to read out the number. A typical dialog from a landline deployment is as follows:

System: City and state please?
User: Brooklyn New York
S: Say the name of the business you want, or say residence
U: Edward's Shoes
S: One moment while I get an operator to assist you
Operator: (operator gets data passed from automation). Is that Edward's Shoe Repair?
U: yeah
O: Please hold for your number…
S: That number is …

One consequence of the diverse caller demographics in Premium DA is a relatively high proportion of unclean utterances. Figure 2 compares various forms of unclean listing utterances between Premium DA (landline data only) and other business search applications. Garbage refers to non-listing utterances such as side-speech, noise and echo; fragments are utterances containing fragmentation of certain words, which can be due to caller speech issues, audio transmission problems or other system-side difficulties; unintelligible utterances are determined by human transcribers as possible listing requests that are not fully intelligible; fillers are non-essential words included in a listing utterance in addition to the salient portion of a listing name. These can include either a prefiller, such as "I want the number to …", a postfiller, such as "… please", or both. Overall, Premium DA has a significantly higher percentage of unclean utterances than other business search applications, particularly in the filler usage rate. This information can be critical to filler modeling in speech recognition and listing search.

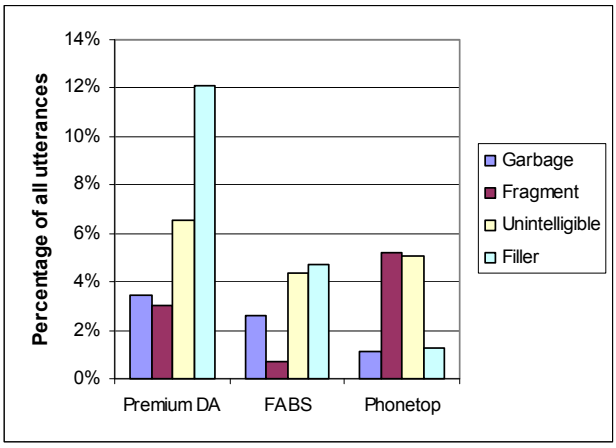

Figure 2 *Comparison of various forms of unclean listing utterances among three business search applications.*

In order to ensure that Premium DA result sets have high precision, we must constrain the listing search space. We only use a full phrase match in the search back-end: we will only match exactly what we thought the caller said. This strategy helps precision at the expense of recall. To regain coverage, we must anticipate the many different ways that users can ask for the same business. This is accomplished through an offline synonym generation phase, which uses a combination of human-generated rules and pattern matching to generate an average of about six synonyms per listing.

3. FULLY AUTOMATED BUSINESS SEARCH

A second approach to DA that we've taken is our Fully Automated Business Search (FABS) product available for free on *800-555-TELL*. This application allows callers to search for business listings either by name or by business category (such as 'shoe store'), with no operator back-up. Since this service is offered for free it draws a somewhat different user demographic with different goals.

The typical FABS caller is young (we estimate under 40), mobile, and more technology savvy than the landline DA callers. There is also a greater proportion of male callers in FABS than Premium DA, as shown in Figure 1. The key characteristic of the FABS callers is that they are extremely price sensitive; they like FABS because it's free. Industry statistics indicate that callers to free DA services are unlikely to call Premium DA services due to their high cost. Our experience shows that these callers are tolerant of longer, more complicated dialogs with more re-prompts and disambiguation dialogs. If a caller cannot find their listing, they are encouraged to use category search as an alternative method. A typical dialog is as follows:

System: City and state, please
User: New York, New York
S: What business or type of business are you looking for?
U: Coffee shop
S: In what neighborhood, or you can say 'I don't know'
U: Gramercy Park
S: Alright, when you hear the one you want, just say its number
One, 71 Irving Place Coffee and Tea
Two, Starbucks
Three…
U: Starbucks
S: What street is it on, or you can say 'I don't know'
U: I don't know
S: Ok, I found 5 locations. When you hear the one you want just say its number…

This relatively extensive automated dialog would be unacceptable to Premium DA callers since it is perceived as slower than an operator lookup, however in the context of Free DA this dialog is well tolerated by callers. The key is that they can find their listing for no charge.

In this context we are much freer in the search back-end. We can afford to match queries more loosely and provide information to the front-end that will help it present more results to the user or conduct a disambiguation prompt. For FABS we broaden the search to use a full boolean conjunction of the query terms, excluding at most stopwords like "the", "on", "in", etc.

This allows us to recover from cases where the caller only specified parts of a listing name (e.g. "pete's" for "pete's hardware store"), while maintaining a very high precision.

There are both differences and similarities between FABS and Premium DA in terms of the listing queries callers make. Figure 3 shows the cumulative percentage out of 10,000 random listing utterances from each application, accounted for by top-ranked listing queries. In FABS the most popular listing queries account for a significantly greater proportion of all listing utterances than in Premium DA. For example, top 200 listing queries account for roughly 19% of all listing utterances in Premium DA, but over 25% in FABS. This is at least partially due to a fairly large proportion of category searches in FABS. On the other hand, FABS and Premium DA appear to have a similarly long tail, with a very large number of infrequent listing queries.

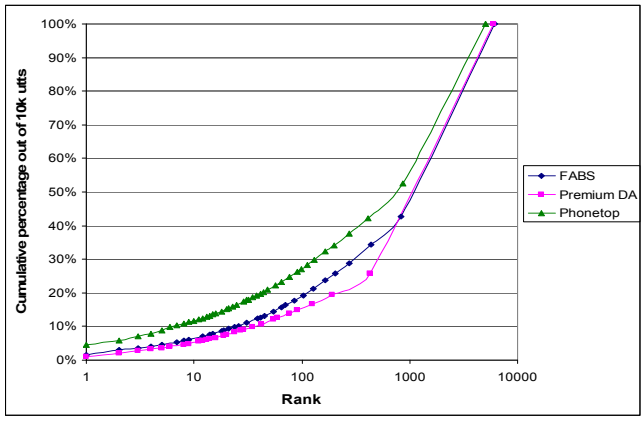

Figure 3 *Cumulative percentage of listing utterances accounted for by top-ranked listing queries, computed on a random sample of 10,000 utterances for each application.*

4. MULTI-MODAL SEARCH (PHONETOP)

Our third example of business search is our multi-modal client, Phonetop [4]. Users must first download and install the client application to their cell phone. Then, with the application running they can step through the DA dialog with a push-to-talk interface and business information is presented visually on their handset screen.

The demographic of Phonetop users is markedly different than that of our other DA services. These users have demonstrated their relatively high understanding of technology by seeking out client applications for their phone. They are likely to be young, urban and technology savvy, and there are significantly more male users than female users (ca. Figure 1). They're more likely to have a high-end cell phone and are familiar with the features of their cell phone. For these callers, "cool" and convenience are the key drivers.

Having a visual interface fundamentally changes the nature of the user experience. In the previous two DA services, what information is presented to the caller has to be carefully considered since the linear, non-permanent nature of an auditory-only experience is such a narrow communication channel. The visual interface of the cell phone screen presents its own limits, most notably its size. However, the user's ability to scroll, scan and potentially save information for later dramatically increases the possibilities. Lists of disambiguation choices can be displayed simultaneously and the user need only to scan the list, scroll and click to move to the next stage of the dialog (ca. Figure 4). The display of recognition results and the availability of the "back" button, allows callers to easily correct recognition errors. Finally, the visual interface allows for easy access to previous searches, via pop-up menus, so information (such as city and state) only needs to be entered a single time.

The push-to-talk interface on Phonetop provides users significant control over their speech input. For example, a user may elect to delay his/her speech input action when there is a train passing by in the background, which on a telephone-based IVR could have triggered a false activation of the speech detector. As a consequence, as shown in Figure 2, Phonetop enjoys a smaller garbage utterance rate. The figure also shows a significantly smaller filler usage rate on Phonetop than the other applications. This can most likely be attributed to the non-linear, multimodal user interface that induces more salient, less conversational-style speech elements. It can also be attributed to the fact that Phonetop users tend to be more technology savvy and more comfortable with advanced human-machine interaction. The relative succinctness and cleanness of Phonetop utterances is consistent with previous findings by other studies on multi-modal interface [5]. The relatively high fragment rate on Phonetop in Figure 2 is due to improper synchronization between button push and start/end of recording on certain models of mobile phones, not an inherent issue with the multimodal interface.

Similar to FABS, Phonetop users are encouraged to use category search if they have difficulty finding a specific listing directly. Table 2 below compares the proportion of listing queries that are category searches, in each of FABS landline calls, FABS wireless calls and Phonetop sessions. It is interesting to observe that category request rate is the highest in landline calls, lowest in multimodal sessions, with wireless calls coming in between. Although we do not fully understand why there is such a marked difference, one conjecture is that landline callers generally have more time for and tolerance to prolonged search interactions than mobile users. Another interesting factor appearing to influence category search rate is whether a caller sounds native (i.e. a fluent American English speaker with no or very little foreign accent). Non-native speakers are more likely to request a category search than native speakers, particularly in wireless and multimodal environment.

Figure 4 *Screenshots from Phonetop client on a typical mobile phone. Left: locality input; middle: a list of listing results after the user said "Indian restaurant" at the listing input; right: a detail page after user selected "Indian Oven Restaurant."*

Table 1 *Percentage of listing requests that are category searches for each of FABS landline callers, FABS wireless callers and Phonetop, as well as conditioned on the nativeness of a caller's speech utterance.*

Category request%	Native	Non-Native	*All*
Landline (FABS)	29.0%	31.5%	29.1%
Wireless (FABS)	20.5%	30.4%	20.6%
Phonetop	16.0%	27.9%	16.6%
All	20.8%	28.7%	21.1%

5. DISCUSSION

In Figure 2 we have seen a large difference in filler usage rate between different business applications. Among utterances from the same application, there can also be a significant variation of filler usage rate, depending on the salient information content. Figure 5 shows filler usage rate by the number of salient words in a listing utterance. For example, in "I want *Starbucks*," only *Starbucks* is considered a salient word. Both Premium DA and FABS have a decreasing filler usage rate as the number of salient words increases from one to four. However, filler usage increases sharply when there are more than four salient words. Even though Phonetop has a significantly smaller filler usage rate overall, the relationship between filler usage and number of salient words is similar to that in Premium DA and FABS.

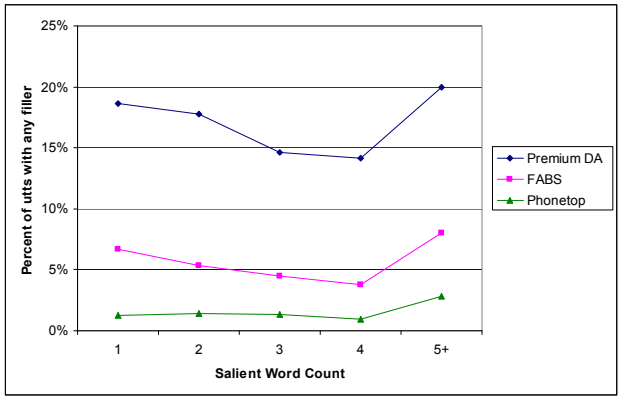

Figure 5 *Filler usage rate by the number of salient words in a listing utterance. The filler usage rate is the percentage of utterances that have either a prefiller, a postfiller, or both, in addition to a salient listing entity.*

Barge-in is a frequently occurring phenomenon in telephone-based IVR applications. It happens when some acoustic event triggers the detection of start of speech before the audio prompt is finished. Such an acoustic event can be a normal caller utterance, which is often seen with power users who are familiar with the prompt. However, in many cases, a barge-in may be triggered by a side-speech, noise, or other unclean utterances that bears no salient information. Table 2 shows the garbage utterance rate (including side-speech, noise, echo, other non-speech utterances, as well as unintelligible speech) at the locality and listing prompt for Premium DA and FABS, conditioned on whether the first prompt of the call has a barge-in (note that the first prompt in both Premium DA and FABS is the locality prompt). For locality utterances, barge-in is clearly associated with high garbage rate. Interestingly, a barge-in at the first prompt of a call can also increase the chance of having garbage utterances further down the call flow. In Table 2, listing utterance garbage rate nearly doubles when the first prompt of the call is a barge-in. Since barge-in status at every prompt is known to the IVR system, it can be exploited to improve garbage modeling [6] in speech recognition.

Table 2 *Garbage utterance rate of locality and listing utterances as a function of whether there was a barge-in at the first prompt of the call, for Premium DA and FABS. Here garbage utterances include side-speech, noise, other non-speech utterances, as well as unintelligible speech.*

	1st Prompt Barge-in	Locality Garbage %	Listing Garbage%
Premium DA	True	38.5%	17.2%
	False	5.0%	9.8%
FABS	True	20.9%	11.2%
	False	3.0%	6.6%

It is hard to recover in search from a bad recognition. Some filler can be eliminated by a simple stopword deletion, but the effectiveness of this tactic can differ between various recognition technologies. However, if the recognition result is not semantically close to what the caller asked then the problem is much harder. Accordingly, when the recognition does not match the caller utterance the relevance of our search results is lowered by a factor of over 2.5.

6. CONCLUSION

Automated voice search is evolving quickly as new technology and business models are being developed. Through case studies of three approaches to voice search – Premium DA, FABS, and Multimodal Phonetop – we have highlighted the importance of understanding user characteristics, business economics, and modality challenges to the success of application design and development. We hope the lessons we have learned from comparing and contrasting the three different approaches can help advance the state of voice search, particularly in user-interface design, speech recognition and back-end search.

7. REFERENCES

[1] C.A. Kamm, K.M. Yang, C.R. Shamieh, and S. Singhal, "Speech Recognition Issues for Directory Assistance Applications," in 2nd IEEE Workshop on Interactive Voice Technology for Telecommunications Applications, 1994.

[2] S.J. Boyce and A,M. Mane. "Challenges of Designing a Large Directory Service," AVIOS, San Jose, CA, May 2002.

[3] P. Natarajan, R. Prasad, R.M. Schwartz and J. Makhoul, "A Scalable Architecture for Directory Assistance Automation," in Proceedings of ICASSP 2002, Vol. 1, pp. I-21,24, May. 2002

[4] http://www.tellme.com/products/tellmebymobile

[5] S. Oviatt, "Multimodal Interfaces for Dynamic Interactive Maps," in Proceedings of CHI 1996, pp. 95-102, 1996.

[6] J.G. Wilpon, L.R. Rabiner, C.H. Lee, E.R. Goldman, "Automatic Recognition of Keywords in Unconstrained Speech Using Hidden Markov Models," IEEE Trans. On ASSP, Vol. 38, No 11, pp. 1870-1878, Nov. 1990.