MS Connect: A fully featured auto-attendant. System Design, Implementation and Performance

David Ollason, Yun-Cheng Ju, Siddharth Bhatia, Dan Herron and Jackie Liu

Microsoft Speech Technologies, One Microsoft Way, Redmond, WA 98052 USA
[dollason, yuncj, sbhatia, dherron, jackiel]@Microsoft.com

Abstract

In this paper we describe our solutions to the problems faced in building the fully featured auto-attendant application that currently provides telephone access to approximately 50,000 Microsoft employees. The paper focuses on issues that the application developer can address directly. Some of the problems tackled in building the grammar include the handling of homonyms, several people with the same name, and data requiring name normalization. Our voice UI helped with this high perplexity task by providing multiple ways to specify the required employee, including spelling their email address or selecting them from a list of employees. Over 70% of the calls analyzed were successfully completed. The system was built using the Microsoft Speech Application SDK and runs on the Microsoft Speech Server 2004.

1. Overview

The system connects people to nearly 50,000 Microsoft employees in North America by phone. Callers speak the name of the employee they want to contact. The system performs confirmation steps and, if successful, connects to the employee, otherwise the caller is connected to the operator. Figure 1 shows some of the possible dialog interactions.

Name Confirmed

S: Good morning, Who would you like to contact?
U: *Yun-Cheng Ju*
S: Are you looking for <Yun-Cheng Ju>?
U: *Yes*
S: OK, Calling the office

Name selected from NBest list

S: Good afternoon, Who would you like to contact?
U: *Siddharth Bhatia*
S: Did you say [Dan Herron]?
U: *No, Siddharth Bhatia*
S: I think you want one of the following three people. Number 1 <Jackie Liu>, Number 2 <Siddharth Bhatia>, NuÖ
U: *Thatís him*
S: Got it. <Siddharth Bhatia>. Calling the office.

E-mail alias used

S: Good evening, Who wouÖ
U: *Dave Ollason*
S: <Ben Miller> Is that right?
U: *No, David Ollason*
S: Are you looking for <David Campbell>?
U: *No*
S: OK, letís try a different way. Do you know the email alias of the person youíre looking for?
U: *Yes, D O L L A S O N*
S: Are you looking for <David Ollason>?
U: *Yes*
S: OK, calling the office.

The <> indicates that the system has a personal recording for this name and [] denotes TTS.

Figure 1 Example Dialogs

The challenges in building systems to do this for a large number of employees fall broadly into two main areas: building the grammar to recognize the names and creating effective dialog strategies for confirming and disambiguating the required employee.

For MS Connect, employee details are stored in a corporate database. The grammar-building problem is made more complex by the fact that employees can be referred to by multiple name variants. In addition to the required first and last name, the database optionally allows for nickname (e.g. Dave for David), maiden name and designation (e.g. Jr. or III) for each employee. Table 1 shows the percentage of people that have used these optional name parts. Also presented are the percentages of collisions i.e. full names and last names shared by 2, 3, 4 etc employees.

<table>
<tr><td colspan="3">Nickname</td><td colspan="3">Maiden</td><td colspan="2">Designation</td></tr>
<tr><td colspan="3">23.7</td><td colspan="3">1.29</td><td colspan="2">0.05</td></tr>
<tr><td colspan="8">% Full names with <b>N</b> collisions</td></tr>
<tr><td><b>2</b>
2.4</td><td><b>3</b>
3.7</td><td><b>4</b>
1.1</td><td><b>5</b>
0.6</td><td><b>6</b>
0.3</td><td><b>7</b>
0.2</td><td><b>8</b>
0.1</td><td><b>9+</b>
~0</td></tr>
<tr><td colspan="8">% Last names with <b>N</b> collisions</td></tr>
<tr><td><b>2</b>
24.9</td><td><b>3</b>
6.3</td><td><b>4</b>
1.9</td><td><b>5</b>
1.7</td><td><b>6</b>
1.1</td><td><b>7</b>
0.7</td><td><b>8</b>
0.6</td><td><b>9+</b>
4.9</td></tr>
</table>

Table 1 Name Statistics

Currently there are 46838 employees in the system. Although many have colliding names, thus reducing the number of distinct names, the extra combinations provided by nicknames and maiden names outweigh this and the final number of unique First/Last name pairs is 57342. The distinction between names and employees is important. The speech recognition is responsible for recognizing a name, whereas the voice UI is responsible for enabling the caller to select a particular employee.

A primary goal of the system design was to minimize the number of calls routed to the human operator by maximizing the task completion rate. Secondary goals included having as fast a transaction time as possible and a dialog experience that was both feature-rich and pleasant to use. In order to achieve these goals the system relies heavily on accurate recognition, but also on an effective voice user interface design.

Section 2 details the approaches taken to construct the grammar. The key aspects of the voice UI design are presented in Section 3 and Section 4 discusses performance on data from real calls to the system.

2. Grammar building

The task of recognizing one of 57342 different proper names pairs is a very hard one [1], from the point of view of both accuracy and computational load. Having recognized the name there is then the task of mapping it to the employee, or set of employees, that accurately reflects the acoustic match. In our implementation the grammar design helped with both of these tasks. The key challenges were to maximize the recognizerís ability to efficiently match all the names and effectively return the correct employee(s).

The grammar allows each person to be referred to by all reasonable name combinations using this basic structure:

[*First* | *Nickname*] [*Last* | *Maiden*] *(Designation)*

Some normalization was applied to the names. Accented and non-alphanumeric characters were replaced. Multi-word names, represented as a continuous string of characters with the word boundary indicated by capitalization, were split. For example, *Yun-Cheng, YunCheng* and *Yun Cheng* are all representations of the same name. The grammar builder also makes single letters optional. For example, first names of the form *F. Scott* or *David G.* are represented such that the names can be referred to with or without the inclusion of the letter.

The remaining tasks were to attempt to ensure that only the minimal set of acoustically variant paths is represented [2] and that employee information is encoded into the grammar [3].

2.1. Homonym merging

A homonym is where two words are pronounced the same but have different spellings. The problem posed by this is how to ensure that the caller is presented with all employees matching the spoken name. For example, the names *Kathy Smith* and *Cathy Smyth* have the same pronunciation, and if represented separately in the grammar, the recognizer will have to arbitrarily choose only one to present as the top choice.

The pronunciation(s) used by the recognizer for each name is either in a dictionary or, for less common names, generated by Letter To Sound (LTS) rules. The Grammar Builder uses both sources to identify all the homonyms prior to building the grammar. The criterion used to define two words, A and B, as a homonym is that they have identical sets of pronunciations. The resultant grammar has a single entry for the homonym-merged name and thereafter it is treated in the same way as other shared names that have a single spelling. The most common spelling of the homonym name is chosen to represent it, as this is likely to be most easily read by a voice talent (Section 3.1).

In total, 4.7% of the names (first, last or maiden) in the database constituted homonyms. Note that the data presented in Table 1 was compiled after the names in the database had been normalized and the homonyms merged.

2.2. Name grammar return values

For the names grammar the recognizer results contain the text of the recognized name together with either a single employee ID, for a unique name, or a list of IDs in the case of a collision. The ID also encodes whether the employee has a personal name recording (Section 3.1). Having this information encoded in the grammar provides the application with all the information required to complete the next dialog turn with the caller, the first phase of confirmation (Section 3.3), without incurring any potential database access delay. For unique names either the recording or TTS of the name text is played. For a collision the system plays the voice talentís recording of the name (Section 3.1).

The callerís perception of delay in the system can be greatly affected by where that delay occurs in the dialog. Having the database access occur after the confirmation better mimics the interaction that would happen with a human operator.

2.3. Spelling grammars

We use tree-structured spelled letter grammars for the email alias[1] and the last name spelling tasks. Email aliases are limited to 8 characters, containing only ëñë digits and letters. The format is both memorable and well suited for recognition. In the same way as for the names grammar, all the information required to complete the next dialog turn with the user is encoded directly into the grammar.

3. Voice user interface design

The voice UI features carefully designed prompts [4], recorded by a professional voice talent, and a dialog employing a variety of techniques in order to determine whom the caller wants to reach.

3.1. Name recordings

Successful confirmation of a name with the caller relies on two things; firstly that the name is correctly recognized and secondly that the presentation of the name is understood by the caller. TTS of names can be difficult to understand, especially in an application where the name is the only part in TTS, giving the caller little context with which to get ëtuned iní so to speak. Also, it is no trivial matter to have a voice talent record 50,000 names. Our approach was to use personal recordings from Microsoft employees, where available, and to put in place a mechanism for ongoing collection. We implemented an internal web site that allows employees to click a button and have a speech application make an outbound call to their office phone asking them to provide a name recording. Email feedback on this feature during rollout was extremely positive in the main.

Note that personal name recordings cannot be used during the confirmation phase for a full name collision. At this stage the system does not know which of the employees the caller wants. A voice talent recorded the collision names (a small fraction of the total number of names) for use in confirmation.

3.2. Presentation of lists

There are three points in the system where the caller is presented with a list: when the confidence score is such that the system presents a list of the top 3 recognition hypotheses (NBest) and when the full name or the spelled last name results in a collision. All three lists operate in the same way, differing only in the content presented. Each item in the list is numbered and the caller can access an item either by saying the number or by selecting at the appropriate time during the list. Take the full name collision list for example:

[1] Within Microsoft it is quite common to know the email aliases of many colleagues.

SYS: *ì There are 3 people at Microsoft with that exact name. | Number 1 <Ben Miller 1> in building 119, Number 2 | < Ben Miller 2> in building 42 and last, Number 3 | < Ben Miller 3> in building 101Ö î*

Note the position of the | symbols which denote boundaries within the prompt. Before the first boundary the caller cannot barge in. Between the first and second, if the caller selects (e.g. ìthatís himî) then the first person is selected. Additionally, at any time after the first boundary, the caller can select a person e.g. by saying ìnumber 1î [5]. Instructions on how to select are provided at the end of the list, along with the ability to repeat. This imposes an initial burden on the naÔve caller, but allows experienced callers to access the list efficiently.

The NBest list is truncated to length 3. The presentation of both Last Name and Collision lists is limited to a maximum of 6 names. If longer the caller is transferred to the operator.

3.3. Spoken name confirmation

The opening prompt is *ì Good morning. Who would you like to contact?î*. The caller speaks a name and the system responds with one of three possible responses:

A:ì Are you looking for <David Ollason>?î
B: ì <David Ollason> Is that right?î
C: ì Did you say [David Ollason]?î

Prompt A is worded to refer to a person, rather than a name, and is played when there is a personal recording available. Prompt B is the same except that it is played in response to barge-in. The dialog adapts accordingly [6] under the assumption that this caller is more experienced and impatient. Presenting the recorded name at the start of the confirmation prompt allows the caller to barge-in earlier, thus minimizing the transaction time. Prompt C is worded to refer to a name and is played when there is no personal recording, regardless of barge-in. Because the name TTS can be confusing we decided that the caller should be prepared for hearing a name.

The first phase of confirmation is relatively simple and focused on the goal of speed. Note that although no NBest list is presented to the caller, it is preserved internally within the application for potential use later in the dialog. The second phase of confirmation is reached when the caller denies the name presented in the first. The caller is asked to repeat the name, if present the name denied in the first phase is removed from the NBest results [7] and then the confidence of the top choice is classified as High, Medium or Low. If High then the name is confirmed with the caller in the same manner as used in the first phase. If the confidence is Medium then the system presents the caller with the top 3 names in the NBest list e.g.

ì I think you want one of the following 3 people. Number 1 <David Campbell>, Number 2 [David Hamilton] Ö î

If Low then the name is unlikely to be in the list at all and so the dialog moves directly on to the spelling options.

3.4. Spelled alias and last name

Having failed to get the right name at two attempts the system asks the caller:

SYS: *ì OK, letís try a different way. Do you know the email alias of the person you are looking for?î*
This turn supports the mixed-initiative response of:
USER: *ì Yes, S B H A T I Aî*

If the caller does not know the email alias then the final phase is to ask for a spelling of the last name. In response, the system presents the list of people with that last name. An interesting point to note is that if people with the spelled last name appeared in the NBest lists from the previous phases then they are placed at the top of the list, and in the case where the list is too long those people alone are presented to the caller.

3.5. Cell phone and email features

Employees can enter their cell number via the web site mentioned earlier. Given that the office is the primary point of contact, coupled with the goal of providing a fast dialog, the system provides the following feature: Try the office first then offer the cell as an option if the office phone is not being answered. The dialog path is as follows.

SYS: *ì <Shu Ito> Is that right?*
USER: *ì Yesî*
SYS: *ì OK, Calling the office. RING RIÖ If youíd like me to try the cell then say Call the Cell nowÖ RING*

While the real office phone is ringing, the caller hears a recording of the ring tone sound overlaid with a prompt offering the cell as an option. Compared to a dialog that asks the caller to choose cell or office before connection this technique reduces the transaction time to the office phone by ~5 seconds.

The system blocks long distance calling and this affects ~15% of the employees in the database. In order to provide a level of service for this scenario the system offers the following:

SYS: *ì Iím afraid I canít call this person. Would you like to record a short message and send it as an email instead?î*

The caller records a message and is then offered the options: Send, Review, Start Over or Cancel.

4. System performance

We present the key performance statistics for the total of 11335 calls that were made to the live system from Feb 18th to Mar 22nd 2004. Microsoft employees span a broad range of ethnic origin, and are making calls to this system from their desk phones in both handset and speakerphone mode. The data has not yet been transcribed and we present results at the system level, as opposed to for the core recognizer. Also, a full analysis of each error category is the subject of future work.

Table 2 shows the rates for task completion. Over 70% of calls are successfully completed, the vast majority being transferred to employees. Those transferred to the operator, but counted as success, include requests for full names that exceed the collision limit of 6.

There are a number of failure categories (all resulting in a transfer to the operator). Too many errors i.e. the caller is not recognized, or they are silent, for three consecutive turns (any combination). The caller can deny all of the names presented by the system. Finally, a spelled last name that results in a list greater than 6 is also counted as a failure.

Uncompleted calls are those where the caller either hangs up or presses 0 to be connected to the human operator. This occurs at all stages of the dialog, and for a variety of reasons, other than simple frustration. For example, testing the system i.e. asking for their own name and hanging up when they hear it, or calling the person who is late to the meeting, and hanging up when they arrive. Requests other than names[1] can also result in hang up or the caller pressing 0 for the operator.

Success 8120 (71.6%)	Transferred Office (96%) Cell (4%)	97.3%
	Transferred to Operator	1.9%
	Email	0.8%
Failure 571 (5.0%)	Too many errors	58.5%
	Caller denied all names	32.4%
	Last name collision list too long	9.1%
Uncompleted 2644 (23.3%)	Caller hung up	81.4%
	Caller pressed 0	18.6%

Table 2 Task Completion Rates

Table 3 shows the rates for successful name confirmation at each of the major stages in the dialog. As mentioned earlier name confirmation relies on both recognition and playback and we count success when the caller confirms a name, or selects one in the case of a list. The first column shows the percentage of the total calls that were successful at each dialog stage. The sum (74.7%) is highly correlated with, but not identical to, the successful task completion rate. Callers who hang up in the full name collision list, for example, have confirmed the name, but failed to complete the task.

Table 3 also shows the average duration of each dialog stage up to the point that the caller confirms the name. As an example, the average call duration for first question confirmation is 10.3 seconds, and for picking a name from the last name list it is this 10.3 plus 12.3 (average for phase 2) plus 21 for last name spelling i.e. 43.6 seconds.

<table>
<tr><th>Successful confirmation stage</th><th>% of total calls confirmed at each stage</th><th colspan="2">Average duration at each stage (sec)</th></tr>
<tr><td>Phase 1 question</td><td>63.0%</td><td colspan="2">10.3</td></tr>
<tr><td>Phase 2 question</td><td>6.6%</td><td>11.7</td><td rowspan="2">Ave
12.3</td></tr>
<tr><td>Phase 2 NBest</td><td>1.4%</td><td>14.6</td></tr>
<tr><td>Email alias</td><td>2.9%</td><td colspan="2">15.6</td></tr>
<tr><td>Last name spelling</td><td>0.8%</td><td colspan="2">21.0</td></tr>
</table>

Table 3 Name confirmation and duration at each dialog stage for successful calls.

The overall average successful transaction time is 29 seconds, compared to 26 seconds for the human operator. This is a reasonably fair comparison.[1]

These are the objective system results. In addition we conducted a usability study with 20 randomly selected employees. The study was longitudinal, in the sense that participants attended 2 separate sessions in the lab. Unobtainable names were injected into the task set such that participants were exposed to the majority of dialog paths. The system achieved a Measure Of Satisfaction (MOS) score of 5.8 out of 7.

5. Conclusions

High accuracy recognition, coupled with the use of personal name recordings, work well together in MS Connect and result in successful confirmation at the first question for the majority of calls. The error recovery and disambiguation strategies boost performance to achieve a task completion rate of 71.6%. This certainly meets the service-level goal of significantly reducing the load on the human operator. Transaction times for these calls are comparable with those achieved by the human operator. The system also provides a highly satisfactory user experience, as indicated by the MOS score of 5.8.

With 23.3% of all calls being uncompleted, due to the caller either hanging up or pressing 0 for the operator, these are clearly issues that need further investigation.

This project has shown that there are a number of issues, some beyond the obvious, which need to be solved in order to create a world-class auto-attendant application delivering a quality service that is feature-rich, highly accurate and connects people quickly.

6. Acknowledgements

We would like to thank Kuansan Wang for his work on the usability study for the MS Connect application.

7. References

[1] B. Maison et al: ìPronunciation Modeling for Names of Foreign Originî; ASRU2003; pp. 429-434

[2] B. Buntschuh et al: ìVPQ: A Spoken Language Interface to Large Scale Directory Informationî; ICSLP98; pp. 2863-2866; Volume 7

[3] K. Wang, ìSemantically object synchronous understanding in SALT for highly interactive user interfaceî; Eurospeech September 2003

[4] Kotelly, Blade ìThe Art and Business of Speech Recognitionî, Addison-Wesley 2003

[5] Balentine, Bruce ìHow to Build a Speech Recognition Applicationî, EIG Press, 2001

[6] K. Komatani et al: îFlexible Guidance Generation using User Model in Spoken Dialogue Systemsî; 41st Annual Meeting of the Association for Computational Linguistics (ACL2003), pp.256--263, 2003

[7] B. Vromans et al: ìExtending the SUSI system with negative knowledgeî; EUROSPEECH99; pp. 2667ó 2670; Volume 6

[1] 85% of the human operator tasks were locating employees and the other 15% were locating departments etc, which the system does not currently handle.