ASR performance analysis of an experimental call routing system

Thipe Modipa[1], *Febe de Wet*[1,2], *Marelie Davel*[1]

[1]Human Language Technologies Research Group
Meraka Institute, CSIR, South Africa
[2]Stellenbosch University Centre for Language and Speech Technology (SU-CLaST)
tmodipa@csir.co.za, fdw@sun.ac.za, mdavel@csir.co.za

Abstract

Call routing is an important application of Automatic Speech Recognition (ASR) technology. In this paper we discuss the main issues affecting the performance of a call routing system and describe the ASR component of the "AutoSecretary" system: an experimental call routing system performing automated secretarial functions. Specific attention is paid to the ASR pronunciation dictionary, and the impact of different dictionary configurations on recognition accuracy is investigated. While ASR accuracy, when evaluated off-line, is found to be high, accuracies recorded during on-line usage are much lower. Reasons for the lower than expected call completion rates are investigated through a detailed analysis of session log files, and improvements to the current system are proposed.

Keywords: Call routing, proper name recognition, automatic speech recognition, spoken dialogue system.

1. Introduction

Call routing systems are an important application of automatic speech recognition (ASR) technology. Internationally, millions of calls are handled by these systems on an hourly basis, with calls being routed to the appropriate destination based on a carefully designed spoken dialogue. As human labour costs can be reduced drastically by using call routing systems on a 24–hour basis [1], research that results in more effective call routing systems can support improved service delivery at a reduced cost.

An experimental call routing system for the South African context was developed at the Meraka Institute. The system performs automated secretarial functions and is known as the AutoSecretary system. This paper describes the ASR component of the AutoSecretary system: its design, implementation and evaluation.

The next section provides some background on issues that affect recognition accuracy in call routing systems, and introduces the AutoSecretary system. In section 3, the design and development of the ASR component are described. An off-line performance evaluation is first conducted and reported on in section 4, before the usage logs are analysed in section 5. These results are compared and evaluated in section 6.

2. Background

2.1. Call routing

In a typical call routing system, a voice-based application is provided with a spoken name in order to make a call to a requested individual. Using a spoken name is the most natural way of interacting with or accessing services, without a user requiring training on how to use a system. Issues that affect the development of call routing systems include proper name recognition, confidence scoring, spoken dialogue design and directory size.

Proper names are difficult to pronounce due to their diverse sociolinguistic origins [2]. Proper name recognition systems often employ explicit pronunciation dictionaries developed by manually creating pronunciations of all possible names to be recognized. These dictionaries can become very large, and are costly to develop and maintain. Techniques that are used to increase proper name recognition in large vocabulary name recognition systems include: speaker clustering, massive adaptation, pronunciation modelling [3] and syllable-based recognition [4].

Confidence scoring techniques can be used to determine how close a recorded utterance is to the ASR hypothesis generated. Acoustic and contextual (language model-based) information can be useful in this regard. The spoken dialogue can be crafted to respond differently to low and high confidence scores. Typically, when the confidence score is high, a call is routed directly to the desired individual, otherwise the caller may have to confirm his or her request by providing further spoken utterances (or DTMF input) to reduce the ambiguity of the ASR hypothesis. In order for call routing systems to be more robust, confidence scoring considers uncertainties from different system components by applying a statistical search [5]. Furthermore, confidence scoring can also assist in predicting the accuracy of proper names in call routing systems [6].

The effectiveness of a call routing system is heavily influenced by the quality of the spoken dialogue design. Considerations such as speaker guidance (guiding a user in providing information in the correct format, at the correct time) and error correction strategies are of particular importance. As the accurate retrieval of proper names is particularly error-prone, a well-designed SDS can compensate for repeated ASR errors. For example, in [7], it has been observed that requiring a combination of spoken and spelled names in the spoken dialogue can increase proper name recognition accuracy.

As can be expected, the recognition accuracy of proper name recognizers decreases as the directory size increases [8]. This is due to the fact that, as the database increases in size, there are more variations in the dictionary which increases the complexity of the search space. Typical call routing systems are able to handle thousands of names.

2.2. The AutoSecretary system

The AutoSecretary system which was developed at the Meraka Institute has two main functions:

1. *Call routing:* a caller can be transferred to either a land line or mobile phone by mentioning the name of the staff member they would like to speak to.

2. *List details:* a caller can obtain staff details such as an office number, building number, group name, and email address from the system.

In the past, these functions were mainly the responsibility of administrative officers. In order to relieve staff of such tasks, the AutoSecretary system was developed with the aim of providing these functions automatically. The service is provided via a dedicated number, which is open for anyone to use.

3. ASR in the AutoSecretary system

This section provides an overview of the AutoSecretary system, focusing specifically on the ASR component.

3.1. Call-flow design

Figure 1 shows the flow of the call routing system. The telephone interface is associated with the caller's speech input. The recognizer decodes the speech input using an acoustic model, and an acoustic score is generated to determine how close the speech input is to this model. If the score is not high, the caller is requested to confirm (or signal a failed recognition attempt) by pressing either 1 or 2 on the phone. A call is only transferred when the caller has confirmed the correctness of the recognition result.

Figure 1: *ASR component of the AutoSecretary system*

3.2. ASR system design

The ASR component was implemented using Hidden Markov Models (HMMs) modelling individual context-dependent triphones and a 3-state left-to-right topology. Within the HMMs, Gaussian mixtures are used as density estimators.

As features, 13-dimensional Mel Frequency Cepstral Coefficients (MFCCs) are extracted every 10 ms for a 25 ms frame. In addition, 13 delta vectors and 13 acceleration features are also calculated per frame. Energy features are not normalized since the recognizer is trained for live audio. Cepstral Mean Normalisation (CMN) is applied, with a global mean (calculated based on the training data) used to initiliase CMN during on-line recognition.

The Hidden Markov Model Toolkit (HTK) version 3.4 [9] is used to implement the ASR component.

3.3. ASR training data

A small speech corpus (referred to as the AutoSecretary data set) was collected over a telephone line to be used as training data. The AutoSecretary data set consists of 31 speakers, each producing 31 utterances. These utterances consist of specific names of individuals in the target user group and are identical in format to the utterances that the AutoSecretary system is required to recognize. The data set is of limited size, consisting of only 961 utterances. There are 61 unique entries made up of first and last names of the individuals in the group. While this data set is very small, it was intended for the development of seed models for initial data collection only. As the system is used, this data set can be increased with operational data collected.

3.4. ASR pronunciation dictionary

Three different ASR pronunciation dictionaries are created: (1) a baseline dictionary, (2) a multiple pronunciation dictionary and (3) a reordered dictionary.

The baseline dictionary consists of a single pronunciation per name. It was created by ordering all the names alphabetically and manually generating each name's pronunciation using English phonemes.

In order to model the pronunciation of proper names accurately, speech recognizers need all possible pronunciations of proper names [10]. In an effort to improve the recognition accuracy we therefore extended the pronunciation dictionary by including multiple possible pronunciations for every name. These possible pronunciations were obtained by asking individuals how they would pronounce a particular name. Figure 2 shows an example of typical entries in the multiple pronunciation dictionary. Interestingly, this exercise revealed that people had different strategies to come up with a pronunciation for a name with which they are unfamiliar.

```
aby          ei b i
aditi        a d i t i
aditi        a d @ t i
alta         a l t a
alta         A: l t @
badenhorst   b A: d @ n h o s t
badenhorst   b A: d @ n h o r s t
barnard      b a r n a r t
barnard      b a r n a t
barnard      b A: n A: d
boitumelo    b Oi t u m e l u
boitumelo    b Oi t u m e l O:
brian        b r ai @ n
```

Figure 2: *Typical entries in the multiple pronunciation dictionary*

In a third experiment, the pronunciation dictionary was reordered by identifying which variants were preferred by the recognizer according to the speech input. The variants were reordered according to the most frequent pronunciation from the recognizer. For each name, the most frequent variant appeared first, followed by the less frequent variants up to the least frequent one.

Dictionary	% Accuracy
Baseline dictionary (without variants)	83.6
Multiple pronunciation dictionary	95.6
Reordered dictionary	95.5

Table 1: *Effect of different pronunciation dictionaries on recognition accuracy*

3.5. Logging

Every time a call is made to the AutoSecretary system the details of the session are captured in a log file. The information captured in the log files includes caller identification (phone number), number of attempts made by the system to recognize the name, ASR responses, caller's responses and a summary of the number of ASR failures and successes. In addition, the raw audio is also captured and stored with the log file, for further analysis.

4. Results: Off-line analysis

We first evaluated the effectiveness of the ASR component in isolation, by training and testing it on partitions of the AutoSecretary training data. We evaluated each of the three dictionaries described in Section 3.4.

4.1. Pronunciation dictionary analysis

The effect of the three different configurations of the pronunciation dictionary was analysed by performing three experiments: obtaining a baseline result; investigating the effect of multiple pronunciations; and investigating the effect of variant reordering. For the baseline result, 21 speakers were included in the training set and the remaing 10 were used as a test set. For the other two experiments, 10-fold cross-validation was implemented. (The set of 31 speakers was divided into 10 folds. We then trained the system with 9 folds and tested on the remaining fold. This was repeated 10 times.) The results of these three experiments are presented in Table 1.

As Table 1 indicates, adding multiple pronunciation variants to the baseline dictionary improves the recognition accuracy. Reordering the variants has no effect. (This may be influenced by the limited amount of training data, as well as the very small directory size, currently limiting the number of confusable names.)

5. Results: Usage logs

Once a sufficiently high off-line accuracy had been obtained, the on-line system could be implemented and actual usage logs could be recorded. The analysis presented here is based on the analysis of 239 log files recorded over a period of approximately 6 weeks.

5.1. Overview of usage statistics

Table 2 shows the number of user attempts per call in relation to the ASR success rate, with the highest number of user attempts being 8 at present. Of 230 calls made, 86 calls were successful and 144 calls were unsuccessful (failures). Of those 86 successful calls, 46 calls were successful on the first attempt, and 25 calls were successful on the second attempt. Furthermore, 21 of the unsuccessful calls did not reach the ASR input as shown in Table 2. The remaining 123 unsuccesful calls did reach the ASR but failed to recognize the required person. The overall call completion rate of the AutoSecretary system was 37%, and the overall ASR recognition rate (of all audio, whether recognizable or not) was 22%.

Nature of error	No. of Attempts	% of Attempts
No audio	158	48.2
Cut names (barge-in errors)	88	26.8
Good quality audio (incorrectly recognized)	34	10.4
First or last name only	13	4.0
Background noise	8	2.4
OOV words	27	8.2
Total	328	100.0

Table 3: *Possible reasons for ASR recognition errors*

5.2. Categorisation of errors

Clearly, on-line performance is much poorer than off-line recognition accuracy. In order to determine the reason for this, a detailed log analysis was performed. Table 3 shows different categories of errors encountered during the recognition of proper names. Almost half of all errors (48.2%) were due to audio that did not contain any speech content whatsoever. Some logs did not contain any audio while others contained sounds such as the phone line being dropped or music playing.

The second largest category of errors (26.8%) was caused by "cut names" where the whole name was not presented to the ASR system, but the start or end of the audio was lost during pre-processing. This is probably due to buffering errors in the larger application. Additional errors were caused by out-of-vocabulary (OOV) utterances (words spoken that do not form part of the recognition grammar), background noise and the use of first or last names only.

Clearly audible audio, incorrectly recognised, contributed to only 10.4% of the errors. When ASR accuracy is calculated as a percentage of all recognizable audio (including first or last names only and audio containing background noise) an accuracy of 70.3% is achieved. When only those utterances containing fairly clear audio and full names (both name and surname) are evaluated, the accuracy increases to 85.7%.

6. Evaluation of results

The off-line accuracy of the ASR component is high while the actual performance of the system is very low (see Tables 1 and 2). This is due to a number of reasons as shown in Table 3. Four categories of errors were identified:

1. *Errors relating to the larger speech processing system and/or its usage.* These include cases where no sound is recorded or names are cut i.e. not stated completely. This is probably a consequence of the fact that, in the current version of the AutoSecretary system, there is neither a beep sound to advise the caller when to start speaking, nor is barge-in implemented.
2. *Difficult ASR environments.* Some calls are made in very noisy environments, causing the recognizer to fail. Call routing systems typically have difficulty dealing

user attempts per call	# successful calls	# unsuccessful calls	# successful ASR attempts	# failed ASR attempts	Total # ASR attempts	% Accuracy
1	46		46	0	46	100
2	25		25	25	50	50
3	7		7	14	21	33
4	5		5	15	20	25
5	2		2	8	10	20
6	1		1	5	6	17
0		21		0	0	–
1		55		55	55	–
2		22	1	43	44	–
3		30		90	90	–
4		8	2	30	32	–
5		6	1	29	30	–
6		1		6	6	–
7		0		0	0	–
8		1		8	8	–
Total	86	144	90	328	418	22

Table 2: *Number of ASR attempts per call before a name is recognized*

with calls from noisy environments and/or speaker noise. These errors require an improved ASR system.

3. *Dialogue-related errors.* When only part of a name (such as first or last name) is supplied, the recognizer makes additional errors as confusability amongst tokens increases significantly. The dialogue can require that only full names be specified, but this restricts the flexibility of the system. Therefore a better solution to this problem may also be improved ASR.
4. *Standard ASR errors.* The models used in the current system were developed with minimal training data and much improvement is possible. Where sufficiently clear audio data is not recognized, channel mismatch may also be the cause of some of the failed ASR attempts. (Channel mismatch has not been investigated and no channel adapation is performed apart from CMN.)

7. Conclusion

In this paper, the ASR component of an experimental call routing system was evaluated. While off-line testing of the system indicates high ASR accuracies, the overall performance of the system itself is poor. This is due to a number of reasons relating to (1) the larger speech processing system; (2) difficult ASR environments; (3) dialogue-related errors; and (4) standard ASR errors.

In order to improve the call routing system, system-level improvement strategies such as barge-in detection and a careful evaluation of the speech processing pipeline are required. In addition, future work will include improving the ASR models through the use of additional training data as it becomes available from the recorded logs, possibly supplemented by data from the newly-released Lwazi corpus. Improved channel normalisation, speaker adaptation and more accurate noise modelling will also form part of our future work.

8. Acknowledgements

Various members of the HLT Research Group supported this project: Brian Khambane and Mpho Kgampe assisted with the dictionaries, Olwethu Qwabe with the spoken dialogue design, and Georg Schlunz and Tshepo Moganedi developed and incorporated different parts of the application.

9. References

[1] Lennig, M., Bielby, G. and Massicotte, J., "Directory assistance automation in Bell Canada: Trial results", Speech Communication, vol. 17, pp. 227–234, 1995.

[2] Jannedy, S. and Mobius, B., "Name pronunciation in German text-to-speech synthesis", Proc. Applied Natural Language Processing, pp. 49–56, 1997.

[3] Gao, Y., Ramabhadran,B., Chen, J., Erdogan, H. and Picheny, M., "Innovative approaches for large vocabulary name recognition", Proc. ICASSP, pp. 53–56, 2001.

[4] Chang, H.M., "Comparing machine and human performance for caller's directory assistance requests", Int. J. Speech Tech., vol. 10, pp. 75–87, 2007.

[5] Wang, Y., Yu, D., Ju, Y. and Acero, A., "Voice search – an introduction", IEEE Signal Processing Magazine, vol. 25, pp. 29–38, 2008.

[6] Schramm, H., Rueber, B. and Kellner, A., "Strategies for name recognition in automatic directory assistance systems", Speech Communication, vol. 31, pp. 329–338, 2000.

[7] Kellner, A., Rueber, B. and Schramm, H., "Using combined decisions and confidence measure for name recognition in automatic directory assistance systems" Proc. ICSLP, pp. 2859–2862, 1998.

[8] Kamm, C.A., Yang, K.M., Shamieh, C.R. and Singhal, S., "Speech recognition issues for directory assistance applications", IVITTA, pp. 15–19, 1994.

[9] Young, S., Evermann, G., Gales, M., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P., "The HTK book", Available, [Online], *http://htk.eng.cam.ac.uk/*, Accessed: Jan 2009.

[10] Spiegel, M.F., "Proper Name Pronunciation for Speech Technology Applications", Int. J. Speech Tech., pp. 419–427, 2003.