Dual-Type Automatic Speech Recogniser Designs for Spoken Dialogue Systems

Jason Littlefield and Michael Broughton
Command and Control Division,
Defence Science and Technology Organisation
PO Box 1500
Edinburgh, Australia, 5111
Jason.Littlefield@dsto.defence.gov.au
Michael.Broughton@dsto.defence.gov.au

Abstract

A Dual-Type automatic speech recogniser (ASR) is a multi-pass ASR system that incorporates both a speaker-independent (SI) and a speaker-dependent (SD) ASR. The purpose of this approach is to improve the robustness of spoken dialogue systems for a broader range of applications. This paper identifies feasible Dual-Type multi-pass ASR system designs that are intended to overcome limitations arising from the use of a single type of ASR. Implementation issues are also discussed.

1.Introduction

Current implementations of Spoken Dialogue Systems (SDS) are developed around a single ASR. This design limits the overall system's recognition accuracy to the performance of the installed ASR, while the useability is determined by the individual ASR type. ASRs can be categorised into two types, either SD or SI, with each having their own strengths and weaknesses.

SI ASRs have the advantage of not requiring a prior enrolment or customised training session for their end users, thereby allowing any user of a given regional dialect to effectively use the system. These systems rely on an underlying grammar that typically needs to be relatively small, or at least, only have a small portion of the grammar active at any point in time. Due to requiring limited grammar size for optimum recognition accuracy, these systems are often used in a system led interaction, whereby the machine asks questions of the user that elicit simple responses. These responses may be single words, or small continuous strings.

SD ASRs, on the other hand, require training for each individual user. This training is relatively short, generally less than ten minutes, and involves the speaker reading aloud a prepared text, which is analysed by the ASR for generation of the speaker's acoustic model. A speaker profile is created by combining this acoustic model with a vocabulary and a regional language model. SD systems are adaptive and have the advantage of being able to recognise free speech or constrained grammar speech, although extracting semantic content from the free speech is more difficult than with a formalised grammar.

In addition to the fundamental differences between the two ASR types, there are also individual differences within each ASR type. The various commercial and research implementations are developed from different algorithms and techniques, typically providing varying output for the same paragraph of spoken text.

ASRs are also further classified by their speech continuity as well as grammar and vocabulary size. Speech continuity describes whether words are spoken in isolation, as connected speech or as continuous speech (Zue et al., 1997). Connected speech ASRs require pauses between multiple word phrases, whereas continuous speech ASRs do not. The grammar and vocabulary size refers to the number of phrases and words that can be spoken and recognized. A grammar can be characterised by the number of plausible alternatives (perplexity), the number of rules and the number of words (Gibbon et al., 1997).

A prototype Multimodal Dialogue System, incorporating an SDS, has been developed for the Future Operations Centre Analysis Laboratory (FOCAL) at Australia's Defence Science and Technology Organisation (DSTO). FOCAL is a collaborative environment that is exploring new paradigms for situation awareness and command and control in military command centres (Wark et al. 2004). An SDS was initially implemented for FOCAL to enable natural dialogue with its Virtual Advisers (Broughton et al. 2002) using an SD ASR and later using an SI ASR (refer section 3). These Virtual Advisers are real-time animated talking heads that can deliver briefs or be queried for additional information. Figure 1 shows some of FOCAL's Virtual Advisers on the main display during an interactive briefing session.

Proceedings of the Australasian Language Technology Workshop 2005, pages 78–86,
Sydney, Australia, December 2005.

SDSs rely on accurate speech-to-text transcription (spoken utterance decoding) from their ASR to perform well. State-of-the-art ASRs perform optimally in quiet environments but are sensitive to interference from ambient noise, overlapping speech and reverberation (Littlefield et al., 2002). Due to the 3.6 metre radius 150° spherical screen, the reverberation characteristics of FOCAL are less than ideal, particularly near the focal point of the screen. This causes degradation in performance of ASRs used in this environment.

Figure 1: Photograph of FOCAL with screen.

To overcome these limitations, we are interested in the development of multi-pass systems, those requiring two or more ASR engines to improve robustness and overcome deficiencies in single ASR based SDSs. The ASR engines can either be of the same or different type, with the overall aim of improving recognition accuracy in a broader range of applications, by utilising the best features of each ASR in the SDS system. An example of an existing multi-pass system is SpeechMAX™ (Custom Speech USA, 2005), a dual-engine system that utilises two SD ASRs, in this case Scansoft's Dragon NaturallySpeaking and IBM's ViaVoice (ScanSoft, 2005). Pellom and Hacioglu (2003) incorporated two passes in the University of Colorado's SONIC ASR to improve robustness in noisy environments. Furthermore, Pérez-Piñar López and García Mateo (2005) use a multiple-pass ASR system where the ASRs have language models adapted from distinct topics.

We are interested in a new area of research that incorporates a Dual-Type ASR to improve SDS robustness. A *Dual-Type ASR* is a multi-pass ASR that incorporates both a SI and a SD ASR. As discussed, SD and SI ASRs have differing advantages and disadvantages to each other, and the aim of this proposed research is to exploit the benefits of these systems to improve recognition accuracy in situations that would normally be detrimental to these systems if used in a traditional single-pass design. Hockey et al. (2003) have developed an SDS that uses two ASRs, a grammar-based SI primary ASR a Statistical Language Model ASR

Section 2 introduces components of an SDS while section 3 describes the past and present SDS in FOCAL. Components of a Dual-Type ASR are identified and explained in section 4. Section 5 describes the alternative designs for a Dual-Type ASR and issues common to all the designs are examined in section 6. Section 7 describes future implementation and experimentation. Finally, the conclusion is provided in section 8.

2.Components of an SDS

The components of an SDS include a microphone, an ASR, a grammar and a Dialogue Manager.

The *microphone* physical design, directionality, frequency response and electrical output are characteristics that help describe different microphone types and aid in the correct microphone selection for specific applications. The microphones that are used in the FOCAL environment include analogue and digital super-cardioid headset microphones, and analogue super-cardioid gooseneck microphones.

An *ASR* decodes audio from spoken utterances into one or more recognition results in the form of text. By default ASRs often display only one speech-to-text interpretation, the most probable interpretation. However, ASRs can produce a list of alternative interpretations, each with a confidence score expressed as a probability or percentage. It is useful to use more than one interpretation in an SDS when another component, such as the Dialogue Manager, has more contextual information than the ASR to select the most likely recognition result.

A speech recognition *grammar* is a list of rules and symbols that can be spoken and recognised by an ASR, often represented as a context free grammar (CFG). The format of the CFG used by an ASR is usually a standard format, such as Nuance Grammar Specification Language (GSL) (Nuance, 2001) or Java Speech Grammar Format (JSGF) (Sun Microsystems, 1998).

The *Dialogue Manager* controls the flow of dialogue with the user and coordinates system responses. The Dialogue Manager, as implemented within FOCAL, can receive one or more recognition results from the ASR system. The additional recognition results are compared within the current dialogue context to improve likelihood of correct recognition.

3.FOCAL's Current SDS

FOCAL's initial SDS (Broughton et al. 2002) was developed around the SD ASR Dragon NaturallySpeaking™, chosen because of its high recognition accuracy, availability and developer support. Additional software for natural language understanding and dialogue management was developed using Natlink (Gould, 2001). Natlink enabled the development of macros and grammars for Dragon NaturallySpeaking. This initial concept system demonstrated the ability to interact with FOCAL's Virtual Advisers. However, the major limiting factor of SD ASRs meant that only those trained with the ASR could use the system. More sophisticated grammars also needed to be implemented to enable scalability of the system.

To address these issues, a second SDS was developed using a SI ASR and a more sophisticated Dialogue Manager based on an agent-based architecture (Estival et al., 2003). In this system, Nuance 8.0 (Nuance, 2001) was chosen as the SI ASR as it provided high reliability, a developer's toolkit, and an Australian-New Zealand acoustic language model. Regulus (Rayner et al., 2001, Regulus, 2005) was incorporated for language processing, enabling the development of typed unification grammars and their compiling into Nuance compatible context-free grammar language models. The agent-based dialogue management system was incorporated into the larger FOCAL agent architecture (Wark et al., 2004) to enable broader application within FOCAL. Currently this system has been implemented to enable users to dialogue with the Virtual Advisers during their presentation of a brief. It enables any one of four Virtual Advisers to be asked questions relevant to their presented information.

The SI ASR in FOCAL's current SDS has two functions. Firstly, it detects, records and saves the audio from spoken utterances as wavefiles. Secondly, it decodes the audio input from the spoken utterance into recognition results. The recognition results are a set of text strings that most closely match rules in the ASR's small CFG.

The Queensland University of Technology Universal Background Model (QUT-UBM) Speaker Identification System (SID) (Pelecanos and Sridharan, 2001) which recognises a person from the sound of their voice, has also been integrated into FOCAL. The SID performs acoustic analysis of the audio from a spoken utterance and tries to match the pattern with that of a trained target user model. The system's response is either the name of the matched target user model or "unknown".

In addition to our current SDS with the Virtual Advisers, we are also exploring multimodal input with an immersive geospatial application (Wark et al., 2005). This system builds on our current SDS, to enable deictic referencing from pointing devices.

4.Components of a Dual-Type ASR

The components of a Dual-Type ASR include a microphone, spoken Utterance Recorder, speech recognition grammar, SI ASR, SD ASR coupled with an SID, and recognition result Error Detector and Reconciler. Configurations of these components are described in section 5.

The *Utterance Recorder* is used to detect, record and save the audio from spoken utterances as wavefiles. Although ASRs are capable of recording spoken utterances, we propose that the use of an independent spoken utterance recorder will lead to a more scalable and flexible system. This is important because more than one of the components requires the audio from spoken utterances at the same time. However, this incurs a delay, since the ASRs cannot begin to decode a spoken utterance until that utterance has finished and has been saved as a wavefile. It takes roughly as long as the duration of an utterance to decode an utterance from a wavefile.

Because there is an independent Utterance Recorder, the *SI ASR* is only required to decode the audio input from spoken utterances into recognition results.

The *SD ASR* also decodes the audio input from the spoken utterance into a set of recognition results. However, because this ASR has a more accurate model of a speaker's voice pattern than the SI ASR, it can use larger grammars. Hence, SD ASRs can operate in at least two different modes: large vocabulary continuous speech (dictation mode) or small vocabulary connected speech (command mode). The dictation mode uses a large vocabulary of 20000 words or more (Zue et al., 1997). The command mode employs a user-defined CFG in a standard format.

Since the SD ASR needs to know the speaker's identity, we couple a *SID* system with the SD ASR in an attempt to automate this process.

The *Error Detector* will select the best recognition result interpretations, measure agreement between the best interpretations, and identify erroneous segments of interpretations.

The recognition results from each ASR include an ordered list of possible interpretations within the grammar, with each interpretation having a confidence score associated with it. The best interpretations will be selected by examining the confidence scores and choosing those above a

predefined threshold.

The interpretation with the highest confidence score from each ASR will be compared for agreement. The assumption here is that if the ASRs produce the same recognition result and this recognition result receives a high confidence score, then it is likely to be correct. In this case a second ASR reinforces the best result of the first ASR. This comparison will be accomplished using Sclite (NIST, 2001), a software tool from the US National Institute of Standards and Technology, that provides word error rate between the two strings, a reference and a hypothesis. If the word error rate is zero, the Error Detector will flag agreement. Since there is only one recognition result in this case, the Reconciler is not required, and is bypassed. If the word error rate is greater than zero, the recognition results are aligned and compared again using Sclite. Sclite aligns the strings and identifies substitution, insertion and deletion misalignments. Part of an example report from Sclite for insertion, substitution and deletion misalignments follows.

```
REF: the brown ** fox JUMPED over THE lazy dog
HYP: the brown IN fox LUMPED over *** lazy dog
Eval:          I      S           D
```

Note that the reference (REF) is only the best recognition result based in confidence scores, not necessarily a correct recognition result. Hence, the substitution, deletion and insertion misalignments are only possible sources of errors.

The degree of agreement (word error rate produced by Sclite) and the location and type of misalignments will be passed on to the Dialogue Manager which will lead to a response to query the user about the error. The best recognition results will be passed on to the Reconciler to process.

It is expected that the Error Detector will require minimal processing for smaller grammars due to the high recognition accuracy achievable with them. The high recognition accuracy will provide identical outputs from both the SD and SI systems and therefore minimal work for the error detection system. As the grammars become more complex however, variation between the two ASRs is expected and providing the correct output in this situation is one of the aims of this research.

The *Reconciler* will receive a set of recognition results from more than one ASR and produce the most probable interpretation. The Reconciler will use an existing system in the speech and language technology domain that makes a selection from multiple output strings. Multi-engine machine translation systems require a component similar to the recognition result Reconciler presented here. DEMOCRAT is an example of such a component for deciding between multiple outputs created by automatic translation (van Zaanen and Somers, 2005).

The best two or three interpretations from each of the ASRs will be sorted in order of confidence and passed on to DEMOCRAT. However, the relationship between the confidence scores from one ASR to another is unknown. The Reconciler will need to take this into account when selecting the best candidate interpretations. DEMOCRAT will produce a consensus interpretation by taking the best segments of each interpretation (van Zaanen and Somers, 2005).

5.Proposed Dual-Type ASR Designs

The following Dual-Type ASR designs we have proposed incorporate one or more ASR to decode the audio input from spoken utterances into recognition results. Each iteration through an ASR is a recognition pass, and therefore, a design using one ASR is a single-pass system, and a design using two ASRs is a two-pass system and so on.

The four proposed Dual-Type ASR designs are:

1. Single-pass ASR
2. Two-pass ASR in parallel
3. Two-pass ASR in parallel with error detection
4. Three-pass ASR in parallel with error detection.

The first Dual-Type ASR system design being proposed is a single-pass ASR which includes a Utterance Recorder followed by a SID system where the speaker's identity is decoded. This design is illustrated in figure 2.

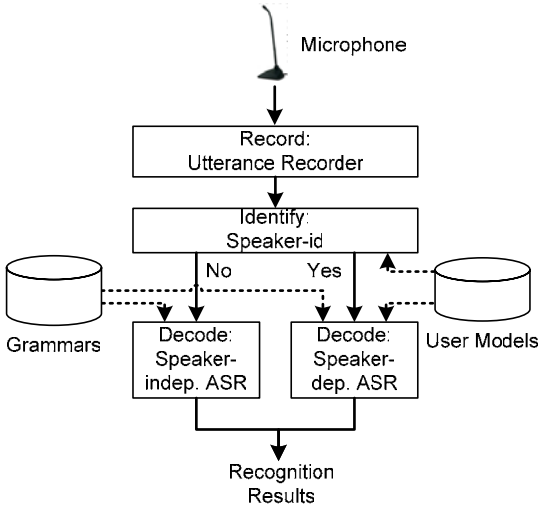

Figure 2: System design of the Dual-Type single-pass ASR.

If the speaker is identified, then the wavefile for the utterance is passed to the SD ASR for decoding into a recognition result. If the speaker is not identified then the wavefile for the utterance is

passed to the SI ASR for decoding. This assumes that the SD ASR is at least as accurate as the SI ASR for large vocabularies as referred to by Merino (2001). This system does not require a Reconciler or Error Detector component.

The second system design, shown in figure 3, proposes a two-pass ASR in parallel where two ASRs decode all spoken utterances concurrently. A spoken utterance recorder detects and records utterances as wavefiles, which are then decoded simultaneously using a SI ASR and a SD ASR incorporating a SID system. The Reconciler compares the recognition results and provides a reconciled result for the SDS Dialogue Manager.

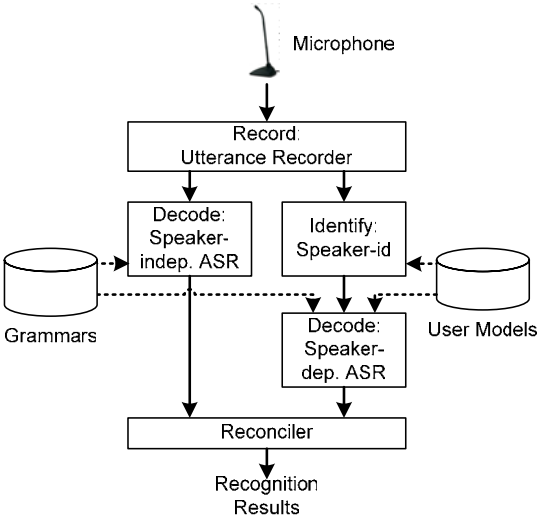

Figure 3: System design for the Dual-Type two-pass ASR in parallel.

The third Dual-Type ASR system design being proposed is an extension of the second system. It is a two-pass system, where two ASRs decode all spoken utterances in parallel, with the addition of an Error Detector. In an effort to be more efficient, a first-pass using a SI ASR will be used every time, whereas the second-pass using a SD ASR will be used only if the Error Detector decides it is required.

As before, a spoken utterance recorder detects and records utterances as wavefiles and the spoken utterances are decoded by the SI ASR and the SD ASR incorporating a SID system. However, the SI ASR recognition result is assessed for errors. If an error is detected, then the result is passed to the Reconciler and compared to the SD ASR recognition result. The Reconciler then passes a reconciled result to the SDS Dialogue Manager. If no errors are found, then the Reconciler is bypassed, and the result from the SI ASR is passed on directly to the SDS Dialogue Manager. Figure 4 illustrates the Dual-Type two-pass ASR in Parallel with Error Detection system design.

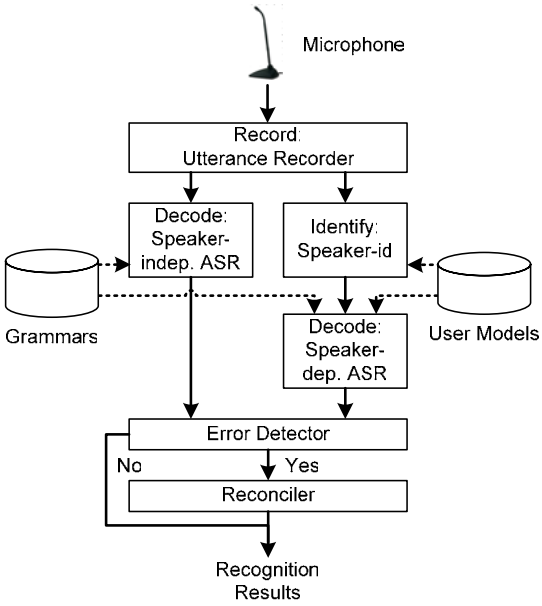

Figure 4: System design of the Dual-Type two-pass ASR in Parallel with Error Detection.

The last proposed design shown in figure 5 incorporates three recognition passes. As in proposal 3 (figure 4), the SI and SD ASR will be used in parallel with error detection. The third pass in this proposal is another SD ASR in dictation mode without a constrained grammar. This would be a useful approach in situations where there are out of vocabulary errors using constrained grammars. The SD ASR in dictation mode has a much larger vocabulary, which could help overcome out of vocabulary errors with constrained grammars and potentially provide a more accurate recognition result.

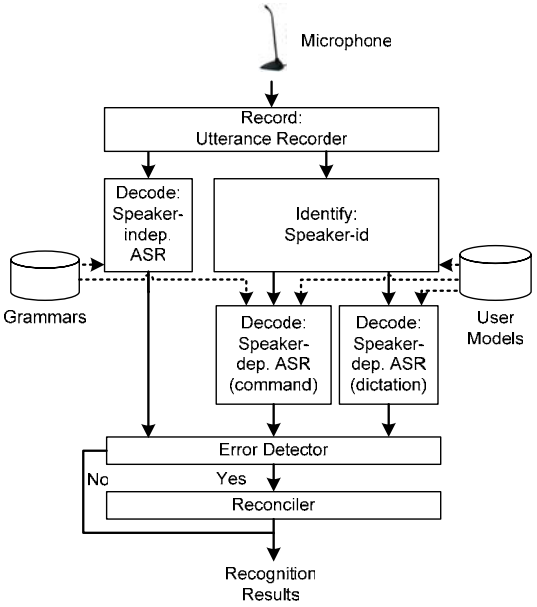

Figure 5: System design of the Dual-Type three-pass ASR in Parallel with Error Detection.

6.Pro and Cons of Each Proposed Design

The speed, accuracy and complexity of each proposed design and its effect on SDS robustness will be compared to determine the most promising approach. A breakdown of the time delay overall is discussed in section 7.1. In the single-speaker situation, the speed of each proposed Dual-Type ASR is estimated to be $2t + 2$ seconds, where t is the length of the spoken utterance in seconds. This assumes the Error Detector and Reconciler incur a negligible time delay.

The accuracy and robustness of each proposed design will be determined through experimentation. Proposed designs 3 and 4 are expected to perform better due to the use of the Error Detector component. This is due to the agreement of recognition results with high confidence scores between ASRs. Also, the additional alignment data enables the Dialogue Manager to query the user when conflicting recognition results occur. That is the ability to query the user for clarification of an utterance segment when required.

The complexity of each of the proposed designs can be described in terms of the number of ASR passes and the number and type of required components. Table 1 shows these terms for each of the proposed designs, 1 though 4. The more complex the design, the more effort required to build and maintain.

Design	No. of Passes	SI ASR	SID	SD ASR	Rec.	Err. Det.
1	1	✓	✓	✓	×	×
2	2	✓	✓	✓	✓	×
3	2	✓	✓	✓	✓	✓
4	3	✓	✓	✓	✓	✓

Table 1: Required components for each proposed Dual-Type design.

7.Foreseeable Design Issues

Before implementing these proposed designs for experimentation, there are some design issues that need to be considered.

7.1.Time Delay

The time delay between the end of spoken utterance and the SDS executing an action is crucial to user satisfaction. A brief investigation was conducted into the duration of spoken utterances and SDS response times in FOCAL. Short sentences, such as 'Yes' or 'No', were about 0.5 second long, while the longest sentence was about 4 seconds long. The corresponding SDS response times were estimated to be between 5 and 10 seconds depending on the complexity of the sentence and resulting action.

The proposed Dual-Type ASR designs incur further time delay. The Utterance Recorder takes the duration of the spoken utterance, which is between 0.5 and 4 seconds, to detect, record and save an utterance. The SID system takes about 2 seconds to identify speakers, while both types of ASRs take about the length of the utterance to produce recognition results. However, if the SD ASR needs to load a different speaker profile, there is an additional delay. Dragon NaturallySpeaking 8 takes about 6 seconds to do this. The recognition result Reconciler and Error Detector have not been implemented yet, however for small grammars their duration is expected to be negligible.

Hence, for spoken utterances of between 0.5 and 4 seconds the estimated overall delay for an SDS incorporating a proposed Dual-Type ASR design will be between 8 and 26 seconds. Note that the Dual-Type ASR is responsible for between 3 and 16s of this estimate. In a multi-speaker environment, the SD ASR will only need to load a speaker profile if the speaker changes. This is not the case all the time. Table 2 shows the breakdown of estimated time delay for 0.5 and 4 second long utterances with and without a change in speaker. The time delay will be measured in future experimentation.

Utt & UR	***SID***	***SD. ASR***		***DM***	***Total***
		Loading	***Transcribing***		
0.5s	2s	-	0.5s	5s	**8s**
4s	2s	-	4s	10s	**20s**
0.5s	2s	6s	0.5s	5s	**14s**
4s	2s	6s	4s	10s	**26s**

Table 2: Estimated time delay for responses with an SDS incorporating a proposed Dual-Type ASR.

The breakdown includes the utterance duration (Utt.), the utterance recorder (UR), the SID system, the SD ASR (loading and transcribing) the Dialogue Manager (DM) and the total.

7.2.Speaker Identification Accuracy

A preliminary trial was conducted by Zschorn (2005) testing a SID system across different spoken utterance lengths. The results demonstrated that for utterance lengths of 0.5, 2.0, 4.0 and 8.0 seconds, the error rates were 57%, 18%, 10% and 6% respectively. Initial investigations into the length of typical spoken utterances using a question-answer (QA) SDS were between 0.5 and 4 seconds. Hence, the accuracy of the SID for very short utterances is expected to be

poor. Both the length of spoken utterances and SID error rates will be measured in future experimentation.

7.3.Grammar Compatibility

In a system where SD and SI ASRs are used in parallel, a single grammar format would be ideal. However there are many different grammar formats. SI ASRs such as Nuance uses GSL and Sphinx 4 uses JSGF. SD ASRs such as Dragon NaturallySpeaking use the Microsoft Speech API 4 (SAPI 4) BNF grammar format and Microsoft's Speech Recognition Engine (MSRE) 5.1 uses GRXML.

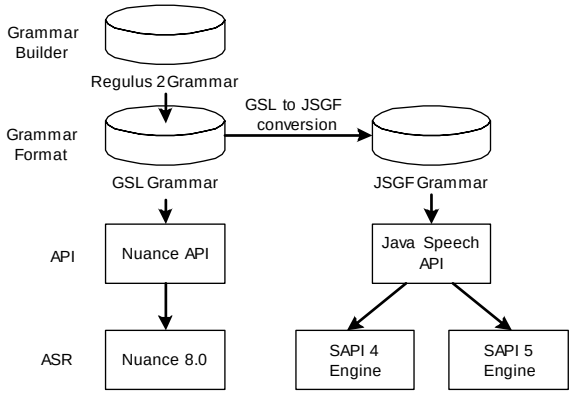

Figure 6: Grammar format and APIs for leading commercial ASRs.

The grammars for the SDS in FOCAL are generated using a grammar building tool called Regulus (Rayner et al., 2001). Regulus can build grammars in GSL format for Nuance. For SD ASRs, any SAPI 4 or SAPI 5 compliant ASR can use JSGF via the Java Speech API (JSAPI), including Dragon NaturallySpeaking, IBM ViaVoice and Microsoft's SAPI 5.1 engine. GSL and JSGF are commonly used standard grammar formats endorsed by World Wide Web Consortium (W3C) in the VoiceXML 2.0 specification. Figure 6 illustrates the relationship between grammar format, API and ASR engines. A GSL to JSGF grammar conversion tool would simplify integration of a Dual-Type ASR. This will be the topic of a Summer Vacation student project at DSTO during the summer 05-06.

8.Implementation

Development of a Dual-Type ASR and integrating it with the SDS in the FOCAL agent-based architecture has already begun. The components of the Dual-Type ASR will be integrated in the current agent-based framework (Estival et al., 2003) so that each of the proposed designs can be tested for experimentation. An agent will be created for the Utterance Recorder, SID system, SD ASR and SI ASR. These agents will interact via a Dual-Type ASR Speech Input agent that will direct data as required. The Dual-Type ASR Speech Input agent will also handle the functions of the recognition results Error Detector and Reconciler as required and interact with the existing Multimodal Input Processor (MIP) in the SDS. The MIP fuses input from multiple modalities and forwards this to the Dialogue Manager (DM). Figure 7 shows the overall design of the Dual-Type ASR and the SDS.

The Utterance Recorder agent will detect spoken utterances, record begin and end timestamps and saves the audio as a wavefile independently.

The SID agent uses the QUT-UBM SID system to decode the identity of the speaker.

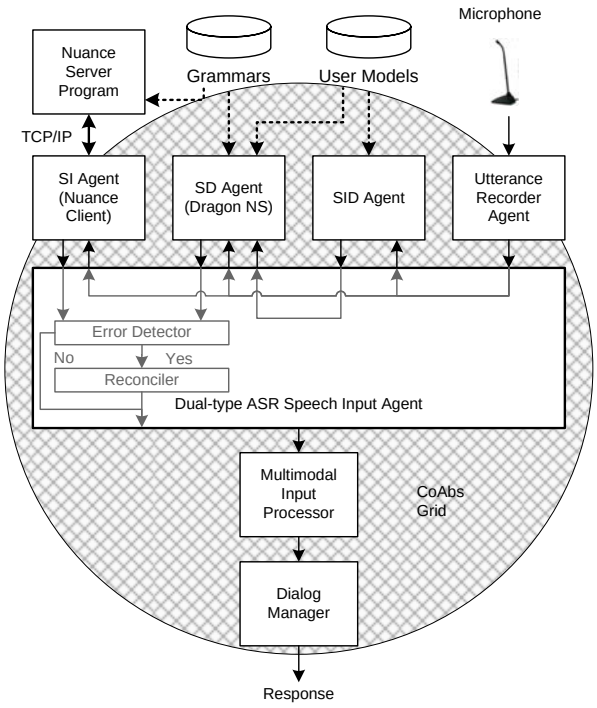

Figure 7: Proposed Dual-Type ASR Speech Input agent with independent Utterance Recorder.

The SD agent will return recognition results. The SD agent consists of the JSGF Grammar and a SD ASR such as Dragon NaturallySpeaking, IBM ViaVoice or MSRE 5.1 integrated using CloudGarden's TalkingJava JSAPI (CloudGarden, 2005). In a multi-speaker environment, the time delay incurred by the SD ASR switching user profiles can be eliminated by using one computer per participant, each with a SD agent. The SI agent consists of the GSL Grammar and Nuance Client. The Nuance Client communicates with the Nuance Server via TCP/IP. This implementation will allow each of the proposed designs will be tested by modifying the routing procedure within the Dual-Type ASR Speech Input agent.

9.Conclusion

In this paper we have presented four alternate designs for a Dual-Type ASR, a system that combines both SI and SD ASRs. The motivation is to provide a more robust SDS system than is currently achievable with a single ASR of either type. We aim to achieve improved robustness through the provision of alternate recognition results from different types of ASR. The designs enable improved user flexibility over a SD system by also providing a SI alternative.

The final design of the Dual-Type ASR system may incorporate several of the proposed designs to maximise the available advantages. These will be reported after the planned development and evaluation of these initial designs.

Acknowledgements

We wish to thank Dr Dominique Estival for her continued guidance and Andrew Zschorn for his consistent contribution to spoken dialogue systems. We would also like to thank Nuance Communications for providing Research membership support for their products.

References

Broughton, M., Carr, O., Taplin, P., Estival, D., Wark, S. and Lambert, D. 2002. Conversing with Franco, FOCAL's Virtual Adviser. *Proc. Virtual Conversational Characters (VCC) Workshop, Human Factors Conference (HF2002)*, Melbourne, Australia.

CloudGarden. 2005. TalkingJava SDK with Java Speech API implementation, http://www.cloudgarden.com/JSAPI/, last accessed 2 November 2005.

Custom Speech USA.. 2005. SpeechMax, http://www.customspeechusa.com/, last accessed 20 August 2005.

Estival, D., Broughton, M., Zschorn, A. and Pronger, E. 2003. *Spoken Dialogue for Virtual Advisers in a semi-immersive Command and Control environment*, 4th SIGdial Workshop on Discourse and Dialogue, Sapporo, Japan.

Gibbon, D., Moore, R. and Winski, R. 1997. *Handbook of standards and resources for spoken language systems*, Walter de Gruyter & Co., pp. 839-852 (Glossary).

Gould, J. 2001. Implementation and Acceptance of NatLink, a Python-Based Macro System for Dragon NaturallySpeaking. The *Ninth International Python Conference*, March 5-8, California.

Hockey, B. A., Lemon, O., Campana, E., Hiatt, L., Aist, G., Hieronymus, J., Gruenstein, A., and Dowding, J. 2003. Targeted help for spoken dialogue systems: intelligent feedback improves naive users' performance. In *Proceedings of the Tenth Conference on European Chapter of the Association For Computational Linguistics - Volume 1*, Budapest, Hungary, April 12 - 17, 2003, European Chapter Meeting of the ACL. Association for Computational Linguistics, Morristown, NJ, pp. 147-154.

Littlefield, J. and Hashemi-Sakhtsari, A. 2002. *The Effects of Background Noise on the Performance of an Automatic Speech Recogniser*, Research Report DSTO-RR-0248, Defence Science & Technology Organisation.

Merino, D. 2001. Speaker Compensation in Automatic Speech Recognition. In J.-C. Junqua and G. van Noord (Eds.), *Robustness in Languages and Speech Technology*, pp. 47-100. Telefónica. Netherlands.

NIST. 2001. NIST *sclite* version 2.2, part of Speech Recognition Scoring Toolkit (SCTK) version 1.2, http://www.nist.gov/speech/tools/, last accessed 11 August 2005.

Nuance. 2001. Nuance Speech Recognition System, Version 8.0: Introduction to the Nuance System, Nuance Communications, Inc.

Pelecanos, J. and Sridharan, S. 2001. Feature Warping for Robust Speaker Verification, in *Proc. ISCA Workshop on Speaker Recognition – 2001: A Speaker Odyssey*, June 2001.

Pellom, B. and Hacioglu, K. 2003. Recent improvements in the CU SONIC ASR system for noisy speech: The SPINE task, *in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2003, pp.14-17.

Pérez-Piñar López, D. and García Mateo, C. 2005. Application of confidence measures for dialogue systems through the use of parallel speech recognizers, In *Interspeech 2005*, pp.2785-2788.

Rayner, M., Dowding, J., Hockey, B.A. 2001. A Baseline Method for Compiling Typed Unification Grammars into Context Free Language Models, *in Proc. of Eurospeech 2001*, Aalborg, Denmark.

Regulus. 2005. Sourceforge Project: Regulus, https://sourceforge.net/projects /regulus/, last accessed 11August 2005.

ScanSoft. 2005. Dragon NaturallySpeaking and IBM ViaVoice, http://www.scansoft.com/, last accessed 5 September 2005.

Sun Microsystems. 1998. Grammar Format Specification, http://java.sun.com/, last accessed 10 September 2005.

van Zaanen, M. and Somers, H. 2005. DEMOCRAT: Deciding between Multiple Outputs Created by Automatic Translation, To appear *in Proc. of 10th Machine Translation Summit*, Phuket, Thailand.

Wark, S., Broughton, M., Nowina-Krowicki, M., Zschorn, A., Coleman, M., Taplin, P. and Estival, D. 2005. The FOCAL Point - Multimodal Dialogue with Virtual Geospatial Displays, *in Proc. SimTecT 2005*, Sydney, Australia.

Wark, S., Zschorn, A., Broughton, M., and Lambert, D. (2004) *FOCAL: A Collaborative Multimodal Multimedia Display Environment. Proc. SimTecT 2004.*

Zschorn, A. 2005. Speaker Identification Test Results, *DSTO informal report.*

Zue, V., Cole, R.A. and Ward, W. 1997. Spoken Language Input: 1.2 Speech Recognition, In Cole, R.A., Mariani, J., Uszkoreit, H., Zaenen, A. and Zue, V. editors. *Survey of the State of the Art in Human Language Technology*, Cambridge University Press, Cambridge, UK, 1997.

Phillips, S. and Rogers, A. 1999. Parallel Speech Recognition, *International Journal of Parallel Programming*, Volume 27, Issue 4, Aug 1999, pp. 257 – 288.