Interacção com um Robô através de Linguagem Natural Falada [*]

Nuno M. Ferreira, Nuno F. Ferreira, Mário Rodrigues[*]
António Teixeira, Francisco Vaz, Luís Seabra Lopes

*Faculdade de Ciências da Universidade do Porto

***Resumo* - O desenvolvimento de um sistema capaz de interagir com um ser humano através da voz é algo que tem sido alvo de muitos estudos e projectos. Este tipo de sistemas são aliciantes pois permitem ao Homem, através de uma forma simples e intuitiva, comunicar com a máquina. A interface de voz com um robô – o Carl -, em desenvolvimento no âmbito do projecto CARL, é apresentada neste artigo. Apresenta-se sobretudo as tecnologias existentes, as ferramentas disponíveis e a interligação de todos os subsistemas constituintes de uma interface de voz: síntese, reconhecimento e processamento de linguagem natural.**

***Abstract* – Developing a system capable of using spoken language for interaction with humans has been pursued by several studies. Such systems are interesting by allowing humans to communicate with machines in a simple and intuitive way. The spoken language interface with a robot (named Carl), in development as part of the CARL project, is presented in this paper. We present, mainly, existing technologies, available tools and connection of all interface subsystems: speech synthesis, speech recognition and natural language processing.**

I. INTRODUÇÃO

A comunicação oral entre o Homem e a máquina é um assunto que tem interessado e fascinado engenheiros e cientistas ao longo de várias décadas. Para muitos, a possibilidade de conversar livremente com uma máquina representa o último desafio para perceber como se processa a comunicação por voz entre as pessoas. Além do desafio que este assunto proporciona, máquinas com interface de voz começam a tornar-se uma necessidade. Num futuro próximo, redes interactivas permitirão um acesso fácil a uma grande quantidade de informação e serviços, o que afectará a maneira como as pessoas trabalham e planeiam o seu dia-a-dia. Hoje em dia, tais redes estão limitadas ás pessoas que sabem trabalhar e têm acesso a computadores. Infelizmente estas pessoas representam uma pequena parte da população, mesmo nos países mais desenvolvidos. O avanço no processamento de voz é necessário para que o cidadão comum possa usar uma grande variedade de aparelhos utilizando somente a sua principal forma de comunicação, a voz. A necessidade de interfaces que usem a voz obriga a desenvolver métodos que permitam o seu reconhecimento automático.

II. PROJECTO CARL

O projecto CARL (**C**ommunication, **A**ction, **R**easoning and **L**earning in Robotics)[18] visa o desenvolvimento de um robô capaz de entender, através de um interface amigável, instruções sob a forma de conceitos familiares ao Homem. Conforme é ilustrado na Figura 1, construir um robô com esta capacidade é visto como um problema de integração de quatro domínios:

- interacção Homem-máquina;
- percepção e capacidades sensório-motoras;
- capacidade de decisão;
- aprendizagem.

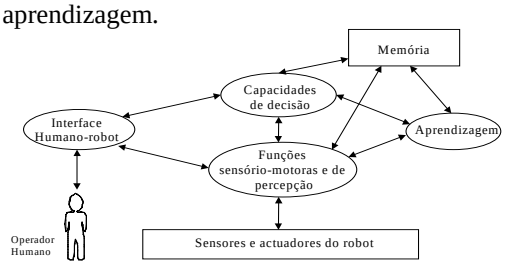

Figura 1 - Diagrama de blocos do projecto CARL.

Apesar destes quatro domínios já terem sido bastante estudados, a sua integração foi pouco tentada.

A filosofia do projecto CARL é baseada na hipótese de que é possível construir um robô inteligente, delineando a estrutura geral e organização da representação do mundo real e dos módulos de execução de tarefas, desde as funções de controlo de baixo nível até às capacidades de decisão de alto nível.

Numa arquitectura deste tipo, os símbolos desempenham um papel muito importante, pois estes vão ser a representação que a máquina terá do mundo. A tarefa de definir símbolos é um problema clássico na programação de robôs. A incorporação de mecanismos de aprendizagem em arquitecturas robóticas vai ser explorada com o intuito de simplificar o problema de programação. O projecto CARL visa uma aprendizagem do robô, não apenas ao nível de conceitos, mas também de tarefas.

É essencial que o robô tenha um interface de linguagem natural, pois este é considerado o único interface aceitável para uma máquina que visa ter um elevado grau de interactividade com o Homem. Além disso, mais nenhum outro tipo de interface é suficientemente flexível para

[*] Projecto de 5º Ano e Estágio final de Licenciatura em Matemática Aplicada à Tecnologia, Universidade do Porto.

resolver o problema dos símbolos, de modo a tornar o robô de simples utilização [14].

A. Um robô chamado Carl

O projecto CARL está a utilizar a plataforma móvel PIONEER 2-DX da ActivMedia Robotics [17]**.** a qual está dotada, entre outras coisas, de:

- três rodas (duas rodas motrizes e uma direccional);
- 16 sonares repartidos em dois conjuntos de 8 (um conjunto à frente e outro atrás);
- pára-choques com sensores de colisão;
- um computador baseado no processador Pentium 266 MMX da Intel, com 64 MB de memória RAM e com o sistema operativo Linux instalado;
- 1 saída de áudio (para a coluna de som);
- 1 entrada de som (para o microfone);
- 1 ligação para uma bússola;
- uma ligação *robot-to-desktop* para comunicação com um PC, entre outras possíveis ligações a outros acessórios;
- também é acompanhado pelo *software* Saphira. Este permite ler informação proveniente dos sonares, controlar os motores, etc.

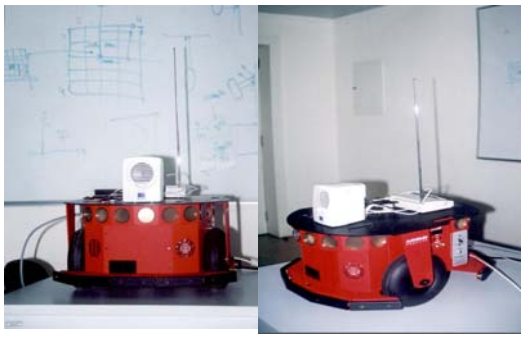

Figura 2 - Imagens do Carl.

III. DESCRIÇÃO GERAL DA INTERFACE

Podemos dividir a interface em quatro blocos principais, tal como se pode ver na Figura 3.

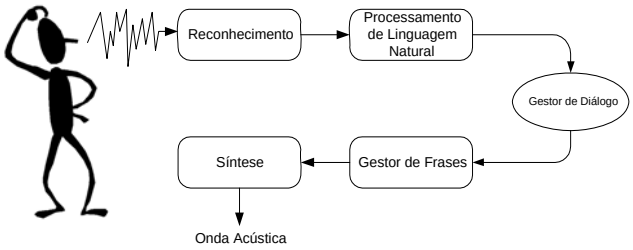

Figura 3 - Blocos da interface de voz

Temos o bloco de reconhecimento que converte o sinal acústico (sinal de voz) numa sequência de fonemas que formam cada palavra e posteriormente a frase. A frase reconhecida apenas nos traz informação disso mesmo - o que foi reconhecido (frase mais provável de ter sido dita pelo interlocutor). Esta informação em si não tem grande significado ao nível da máquina. Como exemplo, podemos ter como frases reconhecidas: *Meio-dia menos um quarto*, *onze horas e quarenta e cinco minutos*, *onze e quarenta e cinco*; todas estas frases, para nós, têm o mesmo significado mas para a máquina, são apenas um conjunto de palavras constituindo no seu conjunto uma frase.

É pois necessário atribuir um significado a estas, para que todas elas sejam compreendidas da mesma forma. Para o efeito existe o bloco de processamento de linguagem natural.

Para que a máquina (robô) possa exprimir através da voz *o que lhe vai na alma*, é necessário um bloco de síntese que transforme um conjunto de palavras (frase) num sinal acústico (voz) perceptível pelo ser humano.

Para gerir todo este processo de reconhecimento e síntese há a necessidade de um quarto bloco – o gestor de diálogo. Este bloco é responsável pelas regras de comunicação para que seja de facto possível dialogar de uma forma amigável com o robô.

Na figura aparece um quinto bloco – gestor de frases. Este bloco pode também ser considerado parte integrante do gestor de diálogo. Serve sobretudo para não tornar monótonos os diálogos, ou seja, evita que numa dada situação o robô diga sistematicamente a mesma frase.

IV. RECONHECIMENTO

O bloco de reconhecimento é o bloco mais crítico e onde se nos depararam a maioria dos problemas inerentes ao interface de voz. As dificuldades surgem quando temos um elevado nível de ruído ambiente: conversas paralelas à frase a ser reconhecida; mudança súbita de contexto por parte do orador; ruído dos motores do robô também pode afectar o processo de reconhecimento.

A escolha do reconhecedor a usar teve em conta as contribuições de ruído; a dimensão do vocabulário usado; a dependência/independência do orador [3]. O requisito principal para o reconhecedor é o de que permita ao utilizador a utilização de linguagem o mais próxima possível da que habitualmente utiliza para interagir com seres humanos. Para isso tem de permitir a utilização de discurso contínuo, ser independente do utilizador sem necessidade de treino, lidar com características da voz natural como as hesitações e responder em tempo real [3].

Um reconhecedor que vai ao encontro destas características é o *grapHvite* [11], desenvolvido pela Entropic (empresa actualmente parte da Microsoft). O *grapHvite* fornece várias ferramentas para o desenvolvimento de aplicações de reconhecimento de voz das quais destacamos a sua *Hidden (Markov Model) ToolKit Application Interface* (HTK API), denominada de HAPI [10] que contém todo um conjunto de rotinas em *C* para o processo de reconhecimento.

Além das rotinas da HAPI, um reconhecedor necessita basicamente de três fontes externas de dados:

- *Lattice*
- Dicionário
- Ficheiro de configuração

A *Lattice* irá conter todas as hipóteses de frases e palavras possíveis de serem reconhecidas. O dicionário contém todas as palavras usadas na aplicação, juntamente com a sua transcrição fonética. O ficheiro de configuração contém a localização de todos os ficheiros externos necessários, bem como todos os parâmetros de inicialização por defeito da HAPI.

O algoritmo de reconhecimento encontra-se na Figura 4.

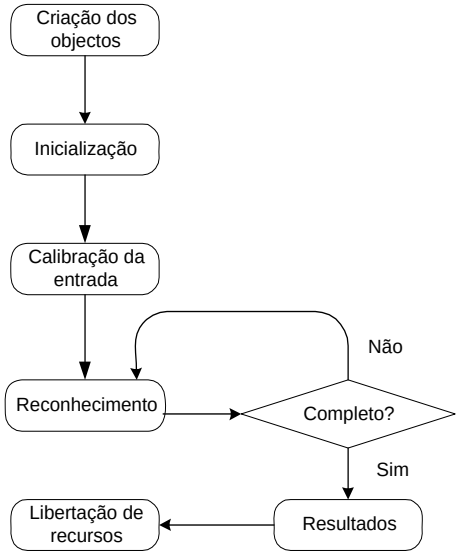

Figura 4 – Algoritmo de reconhecimento

Como resultado o reconhecedor pode devolver para além da frase reconhecida um nível de confiança na frase reconhecida. Na implementação efectuada, esse nível foi utilizado para rejeitar frases pouco prováveis.

V. PROCESSAMENTO DE LINGUAGEM NATURAL

O resultado do reconhecimento em si não traz informação útil para o robô. Há a necessidade de dar um novo formato a essa informação.

É feita uma análise semântica [1] à informação proveniente do reconhecedor, baseando-nos numa gramática específica APS [5] (mais detalhes sobre APS no Apêndice 1). Existe desta forma uma grande dependência entre a gramática da *lattice* e a usada para análise semântica.

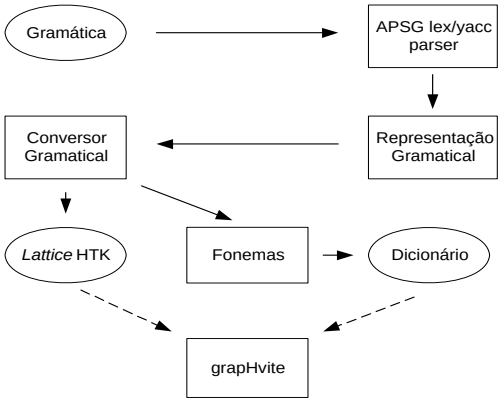

Figura 5 - Conversor gramatical

Para minimizar potenciais fontes de erro provenientes de alterações nas gramáticas, foi desenvolvido um conversor gramatical que, através de apenas uma gramática, gera a *lattice* e o dicionário para o reconhecimento, ficando desta forma a informação coerente nos três sítios, ou seja, todo o processo de compreensão da linguagem é controlada por apenas uma gramática. O processo de conversão é apresentado na Figura 5.

A análise semântica foi feita em conformidade com a representação HRCL (Human Robot Communication Language) apresentada em [14]. (ver Apêndice 2).

Um exemplo de reconhecimento e análise semântica é apresentado de seguida:

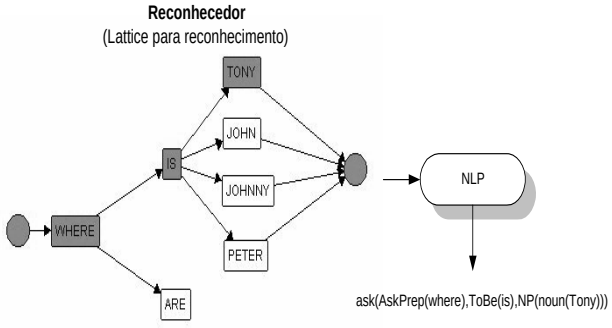

Figura 6 – Análise semântica

O interlocutor diz a frase “Where is Tony”; esta informação ao passar pelo módulo NLP, fica com o seguinte aspecto:

```
ask(AskPrep(where),ToBe(is),NP(noun(Tony)))
```

onde é indicado ao robô que o interlocutor está a perguntar algo “ask” e que a frase é constituída por uma preposição “where”, por um verbo To *Be* “is” e por fim um sintagma nominal (NP) com um nome “Tony”.

Para a parte de síntese, também foram usadas ferramentas de processamento de linguagem natural para geração aleatória de frases. A API usada para o processamento de linguagem natural foi desenvolvida na Universidade de Aalborg, Dinamarca, pela equipa de Tom

Brøndsted, tendo sido criadas por nós novas rotinas adequadas à aplicação em causa.

VI. SÍNTESE

O módulo de síntese de voz transforma uma mensagem proveniente do gestor de diálogo numa onda acústica. Além da habitual conversão de texto para voz é competência deste módulo transmitir informação paralinguistica, como a emoção. Pretende-se que consoante o tipo de mensagem a sintetizar (informação, pedido, mensagem urgente, ..) sejam utilizados parâmetros do sintetizador (como o tom, intensidade e velocidade) adequados. É nossa convicção que mesmo uma abordagem rudimentar será apreciada pelos utilizadores.

A. Requisitos

Os principais requisitos para o módulo de síntese são: a voz sintética tem de ser inteligível e o mais natural possível; deve ser possível manipular a velocidade, o volume e o tom (*pitch*); o sistema deve possuir boa capacidade de normalização de texto, para lidar com acrónimos, números, etc.; existir uma boa interface para programação (API); poder ser usada em Linux e, se possível, em Windows.

B. Sistema utilizado

Para síntese de voz estão disponíveis várias ferramentas. Foram testados dois sintetizadores: o Festival [2] e o *ViaVoice™ Outloud da IBM.*

Após alguns testes (ver [7]) foi escolhido o sintetizador oferecido pela IBM: o *ViaVoice™ Outloud* [3]. Nos testes este sintetizador revelou-se o mais rápido, de qualidade superior e exigindo recursos computacionais (memória e espaço em disco) muito inferiores ao Festival.

Este sintetizador permite alteração de inúmeros parâmetros da voz sintetizada, conseguindo assim atribuir emotividade e identidade à voz do robô.

A síntese não foi alvo de grande pesquisa da nossa parte, tendo sido implementado apenas um conjunto de frases pré-programadas para vários contextos e com um pouco de emoção. O tom, velocidade e intensidade utilizado para transmitir mensagens urgentes, como bateria fraca, é diferente do utilizado para simples informações.

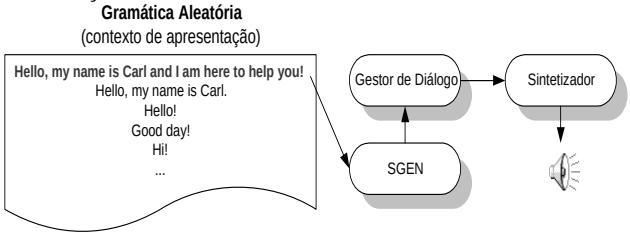

Figura 7 – Escolha de uma frase para síntese

Para melhorar a comunicação as mensagens geradas pelo robô devem variar. Esta variação pode ser conseguida através da utilização de um processo aleatório de escolha dos vários componentes das mensagens. Na implementação efectuada foi utilizado um gerador de mensagens aleatórias com base numa gramática baseado na ferramenta SGEN (Sentence GENerator) do CPK NLP [6]. Foi definida uma gramática para síntese, contendo várias sub-gramáticas para diferentes contextos de conversação. Destas sub-gramáticas é então escolhida uma frase aleatoriamente como o ilustrado na Figura 7.

VII. INTERLIGAÇÃO

A interface foi desenvolvida em *UNIX*, estando cada bloco atrás descrito, implementado como um processo único que comunica com os outros através de troca de mensagens ([7] cap. 5). O bloco de Processamento de Linguagem Natural, não se encontra estanque num processo, mas faz parte dos módulos de reconhecimento e gestão de diálogo. O diagrama de blocos da interface final pode ser visualizado na Figura 8.

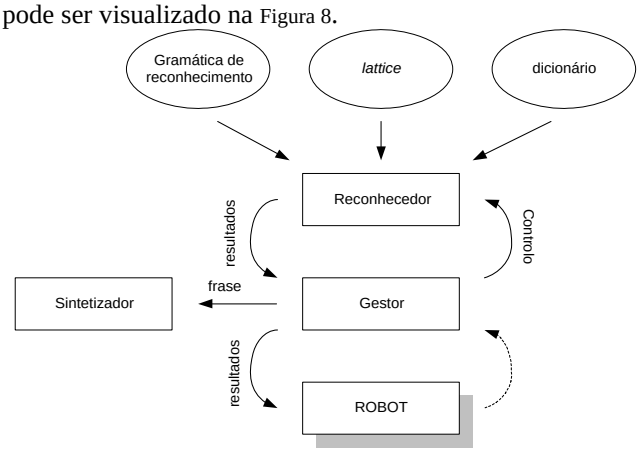

Figura 8 - Interface de voz

VIII. AVALIAÇÃO DO DESEMPENHO DO RECONHECEDOR

Existe um programa para avaliar o desempenho do reconhecedor. O método é o que normalmente é utilizado nas avaliações de sistemas de reconhecimento pelo *National Institute of Standards and Technology* (NIST) americano [19], pois só assim é que se pode comparar resultados com outros projectos/publicações. Este compara as frases ditas pelo orador com as frases reconhecidas. Para melhor perceber o funcionamento deste vai ser dado um exemplo:

```
o orador diz: "Teixeira is in Aveiro"
o robô reconhece: " Is Teixeira in Aveiro now"
```

O programa vai tentar, a partir da frase pronunciada, chegar à frase reconhecida e escolhe a maneira que seja menos penalizada. Para isso pode realizar 3 operações com as seguintes penalizações:

- substituir palavras (S), penalização: 10;
- inserir palavras (I), penalização 7;
- remover palavras (R), penalização 7.

Por exemplo, duas formas para atingir a frase reconhecida a partir da pronunciada seriam:

primeiro caso:

Original	Teixeira	Is	in	Aveiro		
Reconhe-cida	Is	Teixeira	in	Aveiro	Now	
Operação	I	S			I	
Penaliz.	10+	10+	0+	0+	7	=27

segundo caso:

Original		Teixeira	is	in	Aveiro		
Reconhe-cida	Is	Teixeira		in	Aveiro	Now	
Operação	I		R			I	
Penaliz.	7+	0+	7	0+	0+	7	=21

Entre estes dois casos o *software* escolheria o segundo caso. Para avaliar o melhor o programa utiliza um algoritmo de programação dinâmica.

IX. RESULTADOS

Foram feitos alguns testes ao interface, que nos permitiram tirar algumas conclusões acerca do trabalho realizado.

A. Primeiro teste de desempenho e efeito da distância ao microfone

Os testes basearam-se em pronunciar quatro vezes as frases abrangidas pela gramática e foram feitos para uma distância curta ao microfone e para uma distância de aproximadamente 2 metros do microfone. O microfone usado para reconhecimento, foi o microfone da Labtec®, LVA-7280 [20], que é um *array* de microfones direccional com um DSP que efectua cancelamento do ruído de fundo.

Os resultados apresentam-se de seguida em forma de gráfico. Apresenta-se a taxa de palavras correctamente reconhecidas WRR (do Inglês *Word Recognition Rate*), percentagem de inserções WIR (de *Word Insertion Rate*) e nível de confiança devolvido pelo reconhecedor para a frase CONF (de *Confidence).*

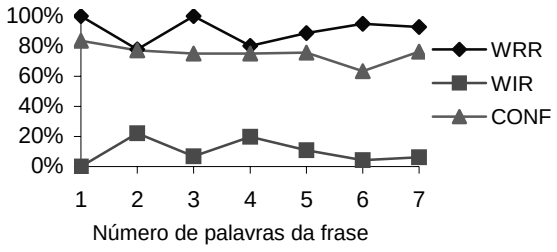

Figura 9 - Reconhecimento perto do microfone

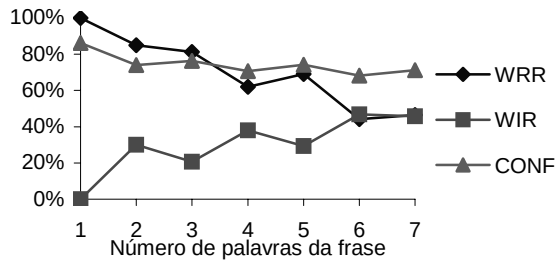

Figura 10 - Reconhecimento a 2 metros do microfone

Através da análise dos gráficos, verificamos que para distâncias curtas foram obtidos resultados aceitáveis. A taxa de reconhecimento de palavras (WRR) tem algumas oscilações mas tende a variar em torno de um valor constante, já para distâncias maiores esta tende a decrescer com o aumento do número de palavras.

B. Teste das gramáticas desenvolvidas

A gramática utilizada nos testes apresentados anteriormente foi posteriormente melhorada [13]. O teste às gramáticas visa quantificar a percentagem de frases gramaticalmente incorrectas aceites pela gramática original e pela gramática melhorada.

O teste consistiu em gerar uma frase aleatória da gramática. O orador decide se a frase é correcta do ponto de vista gramatical. O processo de geração aleatória e classificação de frases continuou até terem sido geradas 200 frases válidas.

Os resultados obtidos no teste das gramáticas são claros. Como se pode verificar na Figura 11, mais de 90% das frases permitidas pela gramática melhorada são gramaticalmente correctas, enquanto que menos de 60% das frases possíveis com a gramática original estão correctas do ponto de vista gramatical.

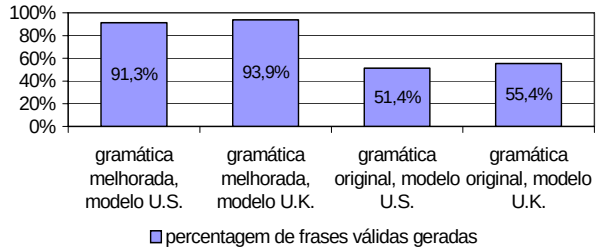

Figura 11 - Desempenho das gramáticas.

C. Teste aos modelos HMM

O teste aos HMM's [12] pretendeu medir o desempenho do reconhecedor, para um determinado orador e com diferentes modelos para inglês (U.K. e U.S.A). Deve ficar claro que o resultado deste teste não pode ser utilizado para saber qual o melhor modelo. Apenas se poderá dizer que o orador escolhido fala inglês mais parecido com o praticado no Reino Unido ou com o praticado nos Estados Unidos da América. Este teste é útil para determinar qual

o desempenho do reconhecedor com os modelos actuais e também para ilustrar a importância do treino dos modelos. Neste teste optou-se pela utilização de um microfone sem fios (Shure TCHS) para evitar a degradação dos resultados com a distância do interlocutor humano ao robô.

Para este teste, o gerador aleatório de frases foi utilizado, com o objectivo de obter sugestões de frases que garantidamente a gramática do reconhecedor está preparada para aceitar. Cada frase correcta foi proferida perante o robô. Isto foi feito para 200 frases correctas.

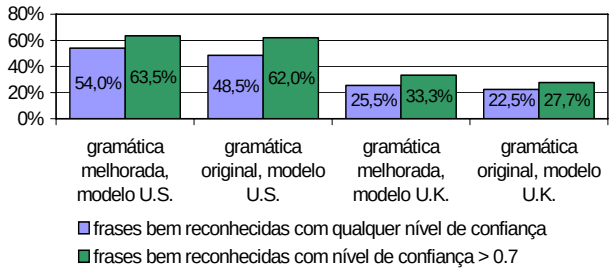

Figura 12 - Frases correctamente reconhecidas.

A percentagem de frases correctamente reconhecidas é apresentada na Figura 12. Como se pode ver, utilizando o mesmo conjunto de palavras, os modelos de voz para os Estados Unidos da América (modelo U.S.) obtiveram sensivelmente o dobro da percentagem de frases bem reconhecidas que os modelos de voz para o Reino Unido (modelo U.K.). Assim é de esperar que, ao treinar modelos para o inglês que se pratica em Portugal (ou pelo menos para as pessoas envolvidas neste projecto), a percentagem de frases correctas seja superior.

É igualmente importante analisar o número de palavras bem reconhecidas. Por vezes, mesmo que uma palavra seja mal reconhecida, esta não influencia o sentido da frase, ou o erro poderá não influenciar a conversa. Por exemplo:

```
o orador pergunta: "Where is the toilet ?"
e o robô reconhece: "Where is a toilet ?"
```

A frase não foi bem identificada, mas a ideia é a mesma.

O número de palavras bem reconhecidas no decorrer do teste encontra-se no gráfico da Figura 13.

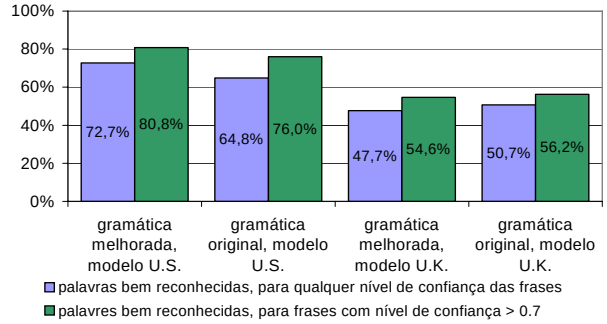

Figura 13 – Palavras bem reconhecidas (WRR)

Como seria de esperar, os modelo U.S. tiveram um melhor desempenho que os U.K e não houve diferenças relacionadas com as gramáticas. No entanto é de referir que a percentagem de palavras bem reconhecidas com modelos U.K. é cerca de 50%.

Esta medida (WRR) torna-se insuficiente, pois não é penalizada quando faltam palavras, quando estas são mal reconhecidas e quando são reconhecidas palavras que não foram ditas. Existe uma medida, designada nos quadros 1 e 2 por "Percent accuracy", que penaliza a percentagem de palavras bem reconhecidas do seguinte modo:

$$\text{Percent accuracy} = \left(\frac{T - S - I - R}{T} \times 100 \right)\%$$

onde T é o número total de palavras, S o número de palavras substituídas, I o número de palavras inseridas e R o número de palavras removidas.

O gráfico da Figura 14 mostra os resultados obtidos para frases com nível de confiança acima de 0.7. Em virtude de o robô só aceitar frases nestas condições, só é interessante analisar este caso. Neste gráfico a percentagem de "palavras correctas" é calculada com a medida "Percent accuracy". Ao somar para uma coluna, as diferentes percentagens (palavras correctas, substituídas, inseridas e removidas), o total pode não dar exactamente 100% devido a erros cometidos pela aproximação.

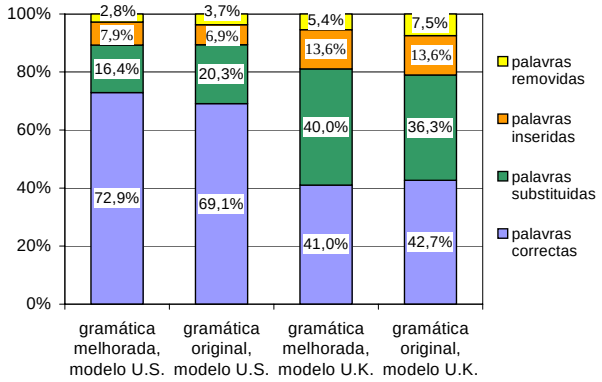

Figura 14 - Análise das palavras

Naturalmente a percentagem de palavras correctas com esta nova medida é inferior à revelada pela Figura 13.

Verifica-se que a percentagem de palavras correctas é superior quando o modelo utilizado é o U.S.. Este resultado seria de esperar, pois os resultados apresentados anteriormente apontam para um melhor desempenho deste modelo com o orador que efectuou o teste.

O número de palavras substituídas reflecte o número de palavras mal reconhecidas, isto é, o orador pronúncia uma palavra e o robô reconhece outra. É nesta medida que existe a maior diferença entre os modelos. Este resultado é natural pois o modelo U.S. ajusta-se melhor a este orador que o modelo U.K.. Ao observar os gráficos, repara-se que a percentagem de palavras substituídas é, no modelo U.K., quase igual à percentagem de palavras correctas. Este dado revela um desempenho fraco deste modelo com este orador. Para o modelo U.S. substituíram-se cerca de três vezes menos palavras que aquelas identificadas correctamente.

A inserção de palavras podem ter duas causas: (1) restrições impostas pela gramática; (2) ruído.

Pode-se perceber melhor como é que restrições impostas pela gramática podem conduzir à inserção de palavras através de um exemplo:

```
o orador diz: "This is the computer lab"
o robô reconhece: "Doty is in the computer lab"
```

A única palavra mal reconhecida foi "This". Devido a ter sido identificada a palavra "Doty" e a gramática não permitir a frase "Doty is the computer lab", foi acrescentada a palavra "in". Obviamente isto só acontece se a frase reconhecida, mesmo com a inserção, tiver um nível de confiança superior a qualquer outra alternativa. As palavras removidas podem provir, à semelhança das inseridas, de restrições impostas pela gramática.

D. Primeira demonstração

Como primeira demonstração do trabalho o robô foi auxiliado através de comandos por voz na tarefa de navegação pelo edifício do Instituto de Engenharia Electrónica e Telemática de Aveiro (IEETA) com resultados encorajadores. O filme desta primeira aventura do Carl foi apresentado no *ECAI Workshop on Service Robotics* [16], que decorreu em Berlim, na *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'2000)* [15], no Japão, e durante a Semana da Ciência e Tecnologia de 2000.

X. CONCLUSÕES

Através dos resultados obtidos, verificamos que há ainda muito trabalho a desenvolver, especialmente quando o orador se encontra afastado do microfone. Os resultados indicam que:

- O *array* de microfones ainda não consegue resolver o problema de captação da voz com o interlocutor humano algo afastado do robô;
- O desenvolvimento de reconhecedores para a HRCL, linguagem em desenvolvimento para a interacção com o robô, é possível e facilitada pela utilização de ferramentas como o CPK NLP toolkit;
- Com o microfone sem fios, foi já possível que o robô desse o primeiro passeio recebendo ordens de um operador humano;
- Os modelos HMM ainda não são os mais adequados, sendo necessário o seu retreino para as condições de ruído e forma de falar dos utilizadores.

A. Trabalho Futuro

No futuro, que se espera próximo, pretende-se que sejam abordados, entre outros, os seguintes assuntos:

- a melhoria do desempenho do sistema de reconhecimento. Uma abordagem poderá passar por colocar um módulo de pré-processamento antes do reconhecedor. Para melhores resultados, os modelos do reconhecedor devem ser retreinados para os diferentes meios (salas, corredores) que o robô irá encontrar, sendo activados quando este os estiver a percorrer. Foram incluídas em [7] outras sugestões para um possível melhoramento do reconhecedor;
- avaliar novas hipóteses para o reconhecimento, como o sistema de reconhecimento *IBM ViaVoice*;
- evolução do vocabulário, frases reconhecidas e geradas pelo robô;
- melhoria da arquitectura e da integração com os subsistemas de navegação e aprendizagem do CARL;
- o sintetizador deve ser estudado profundamente de forma a ser possível dar uma identidade própria à voz do robô;
- avaliação exaustiva do desempenho do sistema. O sistema deverá ser testado em tarefas reais, como a de servir de guia a um visitante do IEETA;
- conversão do processo de reconhecimento e síntese para a utilização da língua portuguesa .

AGRADECIMENTOS

Este trabalho foi financiado pela FCT através do projecto CARL PRAXIS/P/EEI/ 12121/1998. Agradece-se a todos os que colaboraram no projecto CARL durante o ano lectivo de 1999/2000, em particular, ao Eng. Ricardo Ruivaco. Agradece-se, também, ao Departamento de Matemática Aplicada da Faculdade de Ciências da Universidade do Porto na pessoa da Doutora Ana Paula Rocha ter tornado possível o estágio do Mário Rodrigues no IEETA.

APÊNDICE 1. FORMALISMO APSG

As gramáticas APSG (Augmented Phrase Structure Grammar) [3] são gramáticas usadas para geração de *frames* semânticas através de frases definidas na gramática. O princípio básico de funcionamento da gramática é caracterizar cada palavra com as suas características de relevo, importantes à aplicação. Por exemplo a palavra *this* poderia ser categorizada como

```
{lex=this,cat=pron,type=demonstrative},
```

o que indica que esta palavra é um pronome *(pron)* do tipo *(type)* demonstrativo. Podemos atribuir as categorias que quisermos a cada palavra, por exemplo onde está *type*, poderia estar *tipo*.

O formato APSG inclui quatro tipos de regras: *axiom, structure building rules (B-rules), lexical rules (L-rules) e semantic rules (M-rules)*. Para a criação de uma gramática, começa-se por definir um *axiom*, que em termos linguísticos corresponde à frase *(sentence)*:

[{cat=s}].

Posteriormente é necessário definir as *B-rules* que são nada mais nada menos do que a estrutura das frases. Se tivermos uma frase constituída pelo sintagma nominal (noun phrase, NP), por um verbo, por exemplo *To Be*, e por outro verbo no gerúndio, teríamos a regra:

```
{cat=s, stype=frase1}
{
    {cat=np},
    {cat=verb,vtype=tobe},
    {cat=verb,vtype=gerund},
}.
```

onde *np* seria:

```
{cat=np,stype=person}
{
    {cat=noun,semtype=person}
}.
```

Com esta regra qualquer uma das frases seguintes seria aceite:

```
Tony is eating.
They are riding.
```

As *L-rules* são as regras mais simples e consistem basicamente na categorização das palavras usadas na gramática. Por exemplo as palavras *Tony* e *is* podem ser categorizadas como:

```
{lex=Tony,cat=noun,type=proper,semtype=person}.
{lex=is,cat=verb,vtype=tobe,nb=singular}.
```

Por palavras: *Tony* é um nome próprio de uma pessoa e a palavra *is* é um verbo do tipo *to be* no singular.

As *M-rules*, são regras semânticas, não necessárias à gramática, que permitem dar uma forma ou significado semântico às frases. Estas regras consistem em duas partes, separadas por '/' a primeira contendo o formato semântico a dar à informação, a segunda descrevendo a árvore sintáctica à qual se deve aplicar o formato em questão. A parte da condição pode conter *links* para outras *M-rules* definidas na gramática; estes *links* consistem em *labels* antecedidos por '#'. Ainda para o exemplo anterior, poderíamos ter:

```
(frase
        (#NP,ToBe($V),Gerund($G))
)
/
{cat=s}
[
        {cat=np}#NP,
        {cat=verb,lex=$V},
        {cat=verb,lex=$G}
].
```

o *link #NP* iria referir-se à estrutura:

```
(NP
    (noun($N))
)
/
{cat=np}
[
    {cat=noun,lex=$N}
].
```

Se a frase em questão fosse *Tony is sleeping*, o resultado semântico seria:

```
(frase(NP(noun(Tony)),ToBe(is),Gerund(sleeping)))
```

APÊNDICE 2. HUMAN-ROBOT COMMUNICATION LANGUAGE

Como já foi referido, o diálogo utilizando linguagem natural é a única forma prática de um utilizador não especialista especificar e ensinar uma tarefa a um robô. Para implementar essa comunicação, o robô terá de ser capaz de gerar e interpretar mensagens (frases). Internamente essas mensagens têm de ser representadas de uma maneira mais formal. A comunidade científica dos sistemas multi-agente desenvolveu ao longo dos anos diversas linguagens para comunicação entre agentes. Provavelmente a mais conhecida é a ACL, acrónimo de *Agent Communication Language* [8].

Para representar as mensagens trocadas entre o robô e o utilizador/professor humano inspiramo-nos na ACL. Em desenvolvimento encontra-se a linguagem HRCL, a nossa *Human-Robot Communication Language* [14]. Para a representação de conhecimento nas mensagens, utiliza-se Prolog, dado que esta linguagem é também a utilizada na implementação dos módulos de decisão e de representação de conhecimento. A ontologia está a ser definida tomando em consideração o tipo de mundo que antecipamos o robô "verá" (dados os sensores instalados) e as capacidades que se pretendem desenvolver. Para simplificar, apenas se considera, por agora, um sistema com apenas um único robô e um único utilizador humano. Algumas "mensagens" para início e fim de diálogos e sub-diálogos foram consideradas. A ideia é de que o robô seja capaz de guardar informação acerca do contexto com a ajuda de uma hierarquia de diálogos e sub-diálogos. Na Tabela I apresenta-se a parte mais importante da linguagem HRCL.

ask	ready	register
ask_if	next	dialogue
tell	rest	dialogue_accept
deny	discard	dialogue_reject
insert	sorry	dialogue_end
delete	standby	dialogue_end_accept
achieve	error	dialogue_end_reject

Tabela I

HUMAN-ROBOT COMMUNICATION LANGUAGE (HRCL)

REFERÊNCIAS

[1] J. Allen, *Natural Language Understanding*. Benjamin/Cummings Publishing Company, 1995.

[2] Alan W. Black, P. Taylor e R. Caley, *The Festival Speech Synthesis Systems (Edition 1.4, for Festival Version 1.4.0*), 1999.

[3] T. BrØndsted *et al.*, *A Platform for developing Intelligent MultimediaApplications*, Tec Rep R-98-1004. Institute of Electronic Systems, Aalborg University, 1998.

[4] T. BrØndsted, "The Natural Language Processing Modules in REWARD and IntelliMedia 2000+", Em *LAMBDA 25*, Copenhagen Business School, pp. 91-108, 1999.

[5] T. BrØndsted, "Implementing a Task Specific Grammar for Recognition and Parsing using the CPK NLP Suite for Spoken Language Understanding".

[6] T. Brøndsted, "The CPK NLP Suite for Spoken Language Understanding", em *Proc. EuroSpeech*, 1999.

[7] N. Ferreira e N. Ferreira, *Interface Humano Máquina do CARL*, Relatório Final de Projecto 5º ano, Dep. Electrónica e Telecomunicações, Universidade de Aveiro, Julho de 2000

[8] M.R. Genesereth e S.P. Ketchpel, "Software Agents", *Communications of the ACM*, p 48-53, 1994.

[9] *IBM ViaVoice™ Outloud API Reference*, Setembro 1999.

[10] J. Odell, *The HAPI Book (for HAPI Version 1.4) – A description of the HTK Application Programming Interface*, Entropic Ltd, 1999.

[11] K. Power, R. Morton, C. Matheson e D. Ollason, *The grapHvite Book (for grapHvite 1.4 - reference release)*, Abril 1997.

[12] L. Rabiner e B.-H. Juang, *Fundamentals of Speech Recognition*, 1993.

[13] M. Rodrigues, *Colaboração no desenvolvimento de um interface de voz*, Relatório de Estágio, Dep. Matemática Aplicada, Faculdade de Ciências da Universidade do Porto, Setembro, 2000.

[14] L. Seabra Lopes A. Teixeira, "Teaching Behavioral and Task Knowledge to Robots through Spoken Dialogues", em *My Dinner with R2D2: Natural Dialogues with Practical Robotic Devices*, AAAI-2000 Spring Symposium Series, Stanford, Março 2000.

[15] L. Seabra Lopes e A. Teixeira, "Human-Robot Interaction through Spoken Language Dialogue", em *IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS)*, Takamatsu, Japão, 2000.

[16] L. Seabra Lopes, K. L. Doty, F. Vaz e J. A. Fonseca. "Service Robotics and the Issue of Integrated Intelligence", em *Proc. ECAI Workshop on Service Robotics - Applications and Safety Issues in an Emerging Market*, Berlim, pp. 35-44, 2000.

[17] http://www.activrobots.com/robots/p2dx.html

[18] http://www.ieeta.pt/CARL

[19] http://www.nist.gov/speech/tests/index.html

[20] http://www.labtec.com/