

FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO
DEPARTAMENTO DE ENGENHARIA ELECTROTÉCNICA E DE COMPUTADORES



Universidade do Porto
Faculdade de Engenharia
FEUP

TESTE DE INTERFACES DE VOZ

RUI JORGE REIS GOMES

LICENCIADO EM ENGENHARIA INFORMÁTICA E COMPUTAÇÃO
PELA FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

DISSERTAÇÃO SUBMETIDA PARA SATISFAÇÃO PARCIAL DOS REQUISITOS DO GRAU DE
MESTRE EM ENGENHARIA INFORMÁTICA

DISSERTAÇÃO REALIZADA SOB A SUPERVISÃO DE
PROFESSOR DOUTOR JOÃO CARLOS PASCOAL DE FARIA,
DO DEPARTAMENTO DE ENGENHARIA ELECTROTÉCNICA E DE COMPUTADORES
DA FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

PORTO, MAIO DE 2007

*Quanto faças, supremamente faze.
Mais vale, se a memória é quanto temos,
Lembrar muito que pouco.
E se o muito no pouco te é possível,
Mais ampla liberdade de lembrança
Te tornará teu dono.*

(Quanto Faças, Ricardo Reis)

Resumo

Esta dissertação estuda o problema do teste de interfaces de voz. Nela descreve-se o estado da arte na área das tecnologias da fala e das metodologias de teste e avaliação de sistemas conversacionais, são apresentados casos concretos de teste e avaliação deste tipo de sistemas e elaboradas algumas propostas de metodologias e ferramentas adequadas ao teste de serviços de voz. Estas propostas são elaboradas tendo em vista a adopção por parte da empresa em que o autor exerce a sua actividade profissional, com a finalidade de, essencialmente, reduzir os custos do processo de teste e melhorar a sua efectividade.

À medida que as exigências do mercado e a complexidade do *software* aumentam, são cada vez mais necessários métodos e ferramentas adequadas de teste que procurem controlar e garantir a qualidade dos produtos. Neste sentido, o teste de *software* está rapidamente a tornar-se numa área crítica da Engenharia de Software. A resposta para tentar garantir a qualidade dos produtos é efectuar verificações e validações em vários pontos ao longo do processo de desenvolvimento.

Avanços registados nestes últimos anos nos sistemas de reconhecimento de fala e tecnologias de desenvolvimento potenciaram o surgimento de inúmeras oportunidades para as empresas, como as operadoras de telecomunicações, fornecerem novas interfaces para aplicações existentes ou novos tipos de aplicações, com custos mais baixos. Testes completos e consistentes são, assim, extremamente importantes para se obter a qualidade desejada. No entanto, no caso das aplicações de voz, os desafios colocados na especificação e execução dos testes são particularmente significativos, muitos deles ainda sem respostas satisfatórias, e para ultrapassá-los é necessário possuir conhecimentos específicos não só da área de Engenharia de Software mas também das Tecnologias de Fala e Factores Humanos. É exactamente neste tema que se insere a presente tese.

Assim, nos capítulos iniciais do documento são apresentados os resultados do trabalho de pesquisa e levantamento do estado da arte efectuado nas áreas da linguística, teoria da comunicação, neurolinguística, fisiologia da fala, tecnologias da fala, arquitectura dos sistemas conversacionais e usabilidade. Esta pesquisa tem por finalidade adquirir conhecimentos, conceitos e teorias, sem os quais não é de todo imaginável desenvolver uma estratégia de testes para este tipo de aplicações, ou realizar os mesmos.

Nos capítulos finais do documento, são apresentados os diferentes tipos de testes que podem ser realizados a um sistema conversacional, e é reportada a experiência do autor num conjunto de projectos nesta área. É também analisada a situação actual a nível de teste de interfaces de voz na empresa em que o autor exerce a sua actividade profissional e são apresentadas diversas propostas tendo em vista a sua melhoria.

Abstract

This dissertation studies the problem of voice interface testing. The state of the art in speech technologies, testing and evaluation of conversational systems is described, along with concrete cases of testing and evaluation of this type of systems. Finally some proposals are also made for the adoption of adequate methodologies and test tools at the company department where the author works, with the purpose of, essentially, reducing the costs of the test process and improving its effectiveness.

The importance of proper testing methods and tools is increasing with the complexity of nowadays software, and so software testing is rapidly becoming a critical software engineering area. The solution to ensure the aimed quality is to verify and validate at several points along the software development process.

Recent breakthroughs in automatic speech recognition and development technologies for speech applications - for instance - allowed the birth of numerous opportunities for the companies, especially telcos, to provide new interfaces for existing applications or even new type of applications, with lower costs and greater return-of-investment (ROI). So, complete and consistent testing is extremely important to attain the required quality. In voice applications, the test specification and its execution pose particular challenges and some of them are yet to be properly addressed - despite being recognized as an area with great future ahead. This is exactly the theme of the present dissertation. To overcome those challenges, one has to be familiar with knowledge, not only in the software engineering filed, but also in speech technologies, human factors and linguistics.

The opening chapters of the present document present the results of the research work and survey of the state of the art in areas such as linguistics, communication theory, neurolinguistics, speech physiology, speech technologies, spoken dialog systems architectures and usability. The aim of this research is to acquire knowledge, concepts and theories, without which it would not be possible to develop an integrated testing strategy, or even to conduct those same tests.

The last chapters of the document present different possible tests for a spoken dialog system and the author hands-on experience from several projects in the area. The document concludes with the analysis of the testing process implemented at the company department where the author works and several proposals for the adoption of methodologies and test tools at the site.

Índice

ABREVIATURAS E ACRÓNIMOS	IX
GLOSSÁRIO	XI
1 INTRODUÇÃO.....	1
1.1 MOTIVAÇÃO.....	1
1.2 INTERFACES DE VOZ	1
1.3 OBJECTIVO	3
1.4 ESTRUTURA DO DOCUMENTO	3
2 CONCEITOS IMPORTANTES.....	5
2.1 LINGUAGEM E LÍNGUA	5
2.2 TEORIA(S) DA COMUNICAÇÃO	5
2.3 NEUROLINGÜÍSTICA	7
2.4 A VOZ HUMANA.....	9
2.4.1 <i>Fonética e fonologia</i>	10
2.4.2 <i>Representação digital</i>	12
2.5 OUVIDO HUMANO.....	13
3 BREVE HISTÓRIA DAS TECNOLOGIAS DE FALA.....	15
4 SISTEMAS CONVERSACIONAIS	20
4.1 ARQUITECTURAS.....	21
4.2 RECONHECIMENTO AUTOMÁTICO DE FALA	23
4.3 SÍNTESE TEXTO-FALA.....	27
4.4 GESTÃO DO DIÁLOGO	30
4.5 ASPECTOS DE USABILIDADE	32
5 TESTE E AVALIAÇÃO DE INTERFACES DE VOZ: ESTADO DA ARTE.....	35
5.1 TESTES DE DESEMPENHO DO ASR.....	36
5.1.1 <i>Corpus de voz</i>	36
Seleccção de corpora de voz	37
Criação de um corpus de voz	39
Alternativa à utilização de corpora de voz	41
5.1.2 <i>Testes estatísticos</i>	42
Precisão e velocidade do ASR.....	42
Robustez do ASR	43
5.2 TESTES DE QUALIDADE DO TTS	45
5.2.1 <i>Inteligibilidade</i>	46
5.2.2 <i>Naturalidade</i>	46
5.2.3 <i>Qualidade do Front-End Processing</i>	47
5.2.4 <i>Outros testes</i>	47
5.3 TESTES FUNCIONAIS DO GESTOR DO DIÁLOGO.....	48
5.4 TESTES DE CARGA E DESEMPENHO DO SISTEMA	50
5.5 AVALIAÇÃO DE USABILIDADE	51
5.5.1 <i>Prototipagem</i>	51
5.5.2 <i>Entrevistas e questionários</i>	52
5.5.3 <i>Avaliação heurística</i>	52
5.5.4 <i>Wizard of Oz</i>	53
5.5.5 <i>Testes de usabilidade</i>	53
6 EXPERIÊNCIAS REALIZADAS	56
6.1 O CORPORA NO PROJECTO FASiL.....	56
6.1.1 <i>O projecto</i>	56
6.1.2 <i>Processo de criação dos corpora</i>	56
6.1.3 <i>Lições aprendidas</i>	58
6.2 TESTE BASEADO EM CENÁRIOS: O CASO DO VPA FASiL.....	59
6.2.1 <i>O Assistente Pessoal Virtual</i>	59
6.2.2 <i>Revisão da documentação</i>	60
6.2.3 <i>Preparação</i>	62

6.2.4	<i>Casos de teste</i>	63
	Determinar funcionalidades a testar	63
	Definição dos casos de teste	65
6.2.5	<i>Criação das scripts de teste</i>	68
6.2.6	<i>Execução dos testes</i>	71
6.2.7	<i>Análise dos resultados obtidos a partir da execução dos testes</i>	74
6.2.8	<i>Lições aprendidas</i>	75
6.3	AVALIAÇÃO DE USABILIDADE - O CASO DO PROJECTO FASiL	76
6.3.1	<i>A avaliação heurística</i>	77
	Participantes	77
	Desenho e materiais da avaliação	77
	Método	78
	Resultados da avaliação heurística	78
6.3.2	<i>Avaliação do VPA2</i>	79
	Participantes	79
	Preparação	79
	Método	80
	Tarefas	81
	Resultados	81
6.3.3	<i>Lições aprendidas</i>	83
7	PROCESSO DE TESTE DE INTERFACES DE VOZ NUM CONTEXTO EMPRESARIAL:	
	SITUAÇÃO ACTUAL E PROPOSTAS DE MELHORIA	85
7.1	CONTEXTO EMPRESARIAL	85
7.1.1	<i>O departamento</i>	85
7.1.2	<i>Serviços e aplicações</i>	85
7.1.3	<i>O sistema de qualidade e o processo de testes</i>	87
7.2	SITUAÇÃO ACTUAL DO TESTE DE INTERFACES DE VOZ	90
7.3	PROPOSTAS DE MELHORIA DO PROCESSO DE TESTE DE INTERFACES DE VOZ	90
7.3.1	<i>A equipa de testes</i>	90
7.3.2	<i>Teste de ASR e TTS</i>	91
7.3.3	<i>Testes funcionais</i>	91
7.3.4	<i>Base de dados de erros</i>	92
7.3.5	<i>Avaliação de usabilidade</i>	93
8	CONCLUSÕES E TRABALHO FUTURO	95
	REFERÊNCIAS E BIBLIOGRAFIA	97

Índice Figuras

Figura 1 - Modelo de comunicação (Fonte: [Shannon 48]).....	6
Figura 2 - Áreas do cérebro importantes na linguagem (Fonte: [Damasio 99]).....	8
Figura 3 - Visão parcial do aparelho fonador	9
Figura 4 - Ouvido Humano.....	14
Figura 5 - Máquina de von Kempelen	16
Figura 6 - Apresentação do VODER em Nova Iorque, 1939	17
Figura 7 - Principais componentes num sistema conversacional	21
Figura 8 - Arquitectura de um sistema conversacional proposta pelo Galaxy Communicator	22
Figura 9 - Principais componentes de um sistema de reconhecimento de fala	25
Figura 10 - Etapas da geração de saída de fala.....	28
Figura 11 - Etapas da síntese de fala	29
Figura 12 - Arquitectura e testes	35
Figura 13 - Segmentação e transcrição (Fonte: [Maeda 02])	41
Figura 14 - Saída do sclite (Fonte: [Greenberg 00]).....	43
Figura 15 - Taxa de reconhecimento consoante SNR (Fonte: [Möller 02a])	45
Figura 16 - Estrutura do modelo PESQ (Fonte: [Rix01]).....	48
Figura 17 - Diagrama de estados de uma interface de voz simples.....	49
Figura 18 – Estrutura de objectivos no PARADISE (Fonte: [Walker 97]).....	55
Figura 19 - Plataforma WOz	57
Figura 20 - Fluxo de trabalho do TestDirector	89

Índice Tabelas

Tabela 1 - Tabela de conteúdos do corpus	58
Tabela 2 - Ficha do produto.....	60
Tabela 3 - Checklist para revisão da documentação.....	61
Tabela 4 - Tabela das Funcionalidades Seleccionadas.....	65
Tabela 5 - Matriz de Rastreabilidade – Casos de Teste para Envio de Correio Electrónico....	67
Tabela 6 - Matriz de Rastreabilidade – Casos de Teste para Pesquisa de Correio Electrónico	68
Tabela 7 - Script (Sequência) para Enviar Mensagem	70
Tabela 8 - Tabela para registar os resultados dos testes	73
Tabela 9 - Resultados do Envio de Mensagens	75
Tabela 10 - Intuitividade, facilidade de utilização e confiança no sistema (combinada).....	82
Tabela 11 - Intuitividade, facilidade de utilização e confiança no sistema (português).....	82
Tabela 12 - Níveis de satisfação (combinada).....	82
Tabela 13 - Níveis de satisfação (português).....	82
Tabela 14 - Tabela para anotação dos resultados da avaliação heurística.....	93

Abreviaturas e Acrónimos

ASCII	<i>American Standard Code for Information Interchange</i>
ASR	<i>Automatic Speech Recognizer</i>
CMMI	<i>Capability Maturity Model Integration</i>
CTI	<i>Computer Telephony Integration</i>
DARPA	<i>Defense Advanced Research Projects Agency</i>
DRT	<i>Diagnostic Rhyme Test</i>
DTMF	<i>Dual Tone Multi-Frequency</i>
DTW	<i>Dynamic Time Warping</i>
ELRA	<i>European Language Resources Association</i>
grXML	<i>Grammar eXtensible Markup Language</i>
HMM	<i>Hidden Markov Model</i>
IPA	<i>International Phonetic Alphabet</i>
IPQ	<i>Instituto Português da Qualidade</i>
ISO	<i>International Organization for Standardization</i>
IST	<i>Information Society Technologies</i>
ITU	<i>International Telecommunication Union</i>
IVR	<i>Interactive Voice Response</i>
LDC	<i>Linguistic Data Consortium</i>
LPC	<i>Linear Predictive Coding</i>
MIT	<i>Massachusetts Institute of Technology</i>
MOS	<i>Mean Opinion Score</i>
MRCP	<i>Media Resource Control Protocol</i>
NIST	<i>National Institute Of Standards and Technology</i>
ODC	<i>Orthogonal Defect Classification</i>
OVE	<i>Orator Verbis Electris</i>
PAT	<i>Parametric Artificial Talker</i>
PDA	<i>Personal Digital Assistant</i>

PESQ	<i>Perceptual Evaluation of Speech Quality</i>
PIM	<i>Personal Information Manager</i>
PSQM	<i>Perceptual Speech Quality Measurement</i>
ROI	<i>Return-Of-Investment</i>
SLG	<i>Spoken Language Generation</i>
SLU	<i>Spoken Language Understanding</i>
SNR	<i>Signal-to-Noise Ratio</i>
SUS	<i>Semantically Unpredictable Sentence Test</i>
TTS	<i>Text-To-Speech</i>
VoiceXML	<i>Voice eXtensible Markup Language</i>
VPA	<i>Virtual Private Assistant</i>
WA	<i>Word Accuracy</i>
WER	<i>Word Error Rate</i>
WKER	<i>Weighted Keyword Error Rate</i>
WSFT	<i>Weighted Finite State Transducers</i>

Glossário

ASR	<i>Automatic Speech Recognizer</i> : componente da aplicação de voz que faz a conversão de um sinal acústico de fala num conjunto de palavras.
Call-Centers	Centros de atendimento com uma central telefónica, tipicamente, onde são atendidas as chamadas dos clientes e, por vezes, se efectuam as campanhas de <i>tele-marketing</i> .
Corpus	Colecção de ficheiros ou informações relativos a uma disciplina ou tema.
DTMF	<i>Dual Tone Multi-Frequency</i> : tonalidades sonoras associadas a cada tecla para transmitir sinais de um telefone ao comutador telefónico que completa a ligação.
Gestão do Diálogo	Designa as estratégias que determinam o fluxo do diálogo falado entre o utilizador e o sistema, especificando a acção a tomar em função do significado da elocução do utilizador.
HMM	<i>Hidden Markov Model</i> : é um modelo estatístico do processo de produção da fala: dada uma sequência de palavras, pode ser usado para gerar o sinal. Pode também ser usado para calcular a probabilidade de um determinado sinal ter sido gerado a partir de uma determinada frase.
Interface Multimodal	Interface com suporte para diversos estilos de interacção em simultâneo ou não (teclado, rato, comandos voz, entre outros).
Interface Unimodal	Interface com suporte para apenas um estilo de interacção.
IVR	<i>Interactive Voice Response</i> : sistema de resposta interactiva de voz. É um sistema computadorizado que permite ao utilizador, tipicamente pelo telefone, seleccionar uma opção de um menu de voz usando o teclado do telefone. Mais recentemente, passaram a poder integrar reconhecimento e síntese de fala.
Matriz de Rastreabilidade	Matriz que relaciona casos de teste (no presente documento) com os casos de uso de um determinado sistema.
Off-the-shelf components	Componentes existentes antecipadamente ao desenvolvimento de um sistema. Componentes de reutilização desenvolvidos dentro da própria organização ou por entidades externas.
Parsing	Análise sintáctica. Verificar se uma determinada frase está de acordo com as regras gramaticais definidas.

Prompts	Saída de voz da aplicação, ou seja, “aquilo” que a aplicação diz ao utilizador. Também designadas por “anúncios”.
Sistemas Conversacionais	Sistemas com os quais o utilizador interage por voz, num diálogo mais ou menos natural. Funcionam como uma interface de voz entre o utilizador e aplicações como bases de dados, sistemas periciais, entre outros.
SNR	<i>Signal-to-Noise Ratio</i> : A relação sinal-ruído designa a razão entre um a informação com significado e o ruído ambiente.
Survey	Método de levantamento de informação sobre uma determinada área ou grupo de indivíduos.
Test Incident Report	Uma vista detalhada do resultado do teste e do resultado esperado, onde falhou e indicações que possam ajudar na depistagem.
TTS	<i>Text-To-Speech</i> : sistema de conversão texto-fala. Componente da aplicação de voz que faz a conversão de um conjunto de palavras num sinal acústico de fala.
Utterances	Frase/elocução do utilizador para a aplicação de voz. Pode ser considerada como a unidade de discurso do utilizador, sendo delimitadas pelas pausas.
Wizard Of Oz	Neste tipo de experiências, o utilizador opera com um sistema aparentemente funcional, mas na verdade as suas funcionalidades são fornecidas por um operador humano escondido.

1 Introdução

1.1 Motivação

Nos últimos anos, os sistemas de *software* têm evoluído exponencialmente quer em termos de complexidade quer em termos de importância nas nossas vidas. O *software* é omnipresente no dia-a-dia de praticamente todas as pessoas: nos automóveis, nos telemóveis, nos locais de trabalho, nos locais de entretenimento ou até nas próprias habitações. A sua importância é tal que as empresas e as economias dependem em grande medida do sucesso do desenvolvimento e utilização dos sistemas de *software*, sendo necessário tentar desenvolvê-los otimizando recursos (temporais e monetários) e com constante procura de qualidade. Neste sentido, e tal como Ilene Burnstein constata no seu livro *Practical Software Testing*, [Burnstein 03], o teste de *software* está rapidamente a tornar-se numa área crítica da Engenharia de Software, sendo cada vez mais importante conceber e conhecer métodos e ferramentas adequadas de teste que possam de alguma forma controlar e garantir a qualidade dos produtos de *software*.

Os testes de *software* procuram avaliar a qualidade do produto e também, ainda que de forma indirecta, melhorá-la, através da identificação de defeitos e problemas. Apesar de poder evitar custos enormes no final do projecto, o processo de testes tem um custo não negligenciável em todo o desenvolvimento, pelo que é fundamental ter o processo de testes perfeitamente integrado no processo de desenvolvimento e conhecer as técnicas e ferramentas mais adequadas.

No caso das aplicações de voz, os desafios colocados na especificação e execução dos testes são particularmente significativos e escasseiam os trabalhos publicados na área. Para ultrapassar estes desafios, é necessário possuir conhecimentos específicos não só da área de Engenharia de Software mas também das Tecnologias de Fala e Factores Humanos.

Nos últimos cinco anos, nos quais desempenhou funções profissionais no departamento de serviços e aplicações de voz do pólo do Porto da Portugal Telecom Inovação S.A., o autor tornou-se ciente da problemática inerente ao teste deste tipo de aplicações e lacunas existentes ao nível da quantidade de trabalhos publicados. Por ser uma área de reconhecido futuro e na qual surgem agora novas e aliciantes oportunidades proporcionadas pelos recentes avanços tecnológicos, a motivação para desenvolver uma tese de mestrado na qual possa transpor os conhecimentos adquiridos na actividade profissional, por um lado, e que possa contribuir quer para a comunidades científica quer para a indústria onde o autor se insere, por outro, é elevada.

1.2 Interfaces de voz

A fala é das mais importantes formas de interacção humana, desenvolvimento, aquisição e transmissão de conhecimento. A integração dos computadores com a voz é um objectivo há muito perseguido – parece-nos natural interagir com uma máquina utilizando a nossa própria voz sem ter de recorrer a periféricos. No entanto, para além da pouca maturidade das tecnologias de reconhecimento e síntese de voz, o desenvolvimento de aplicações era bastante dispendioso: tinha de ser feito à medida de uma plataforma específica (baixo nível de reutilização de código), requeria conhecimentos muito específicos e a manutenção era complexa.

Os avanços registados nos últimos anos nos sistemas de reconhecimento de fala e tecnologias de desenvolvimento (como o VoiceXML ou o MRCP - *Media Resource Control Protocol*) potenciaram o surgimento de inúmeras oportunidades para as empresas, como as

operadoras de telecomunicações, fornecerem novas interfaces para aplicações existentes ou novos tipos de aplicações, com custos mais baixos. Muitas empresas já disponibilizam informação aos funcionários, fornecedores e clientes através de interfaces automáticas de voz, usando o telefone por exemplo, reduzindo custos com *call-centers* ou complementando os seus sítios *web*.

Comparando o número de pessoas que possuem telefone (fixo ou móvel) com o número de utilizadores de computadores pessoais, podemos perceber o potencial de mercado das aplicações com interfaces de voz para o telefone e a razão pela qual as empresas apostam cada vez mais nestas tecnologias. O telefone também pode ser um meio privilegiado para disponibilizar estas aplicações a pessoas com fobia às tecnologias ou com diferenças de acessibilidade. Esta cada vez maior simbiose entre o telefone e o computador tem o nome de Integração Computador-Telefonia – *Computer Telephony Integration* (ou CTI).

Um dos tipos de aplicações existentes é os designados sistemas IVR (*Interactive Voice Response*). Neste tipo de aplicações, o utilizador pode utilizar os toques das teclas (DTMF - *Dual Tone Multi-Frequency*) para percorrer os menus, mas mais recentemente pode já usar a sua voz para o efeito. Nestas aplicações utilizam-se diálogos que são, por natureza, muito rígidos e sequenciais, sendo utilizados por um grupo bastante grande e heterogéneo de utilizadores (ver por exemplo o serviço “118” da Portugal Telecom).

Outras aplicações permitem aos utilizadores através do telefone aceder a dados e serviços, como por exemplo, um sistema de notícias ou um serviço de bilhetes para eventos. Neste tipo de aplicações, o diálogo já não é tão rígido e os utilizadores produzem frases mais naturais e de mais difícil reconhecimento.

Se avançarmos um pouco mais na escala de complexidade, podemos encontrar aplicações de voz que permitem os utilizadores realizarem tarefas como interagir com um serviço de voz para gerir a sua informação pessoal (como e-mail, contactos e agenda) – ver o exemplo do projecto europeu FASiL (<http://www.fasil.co.uk/>) [Gomes 05].

Entre as razões pelas quais as empresas apostam cada vez mais nas aplicações de voz, e para além da inerente naturalidade deste modo de interacção, em [Markovitz 96], encontramos a procura de aumentos de produtividade, um rápido retorno de investimento, acesso a novos mercados e a diferenciação dos produtos e, por vezes, das próprias empresas.

De uma forma simplista, um sistema conversacional, ou sistema de diálogo falado (do Inglês *Spoken Dialogue System*), é uma interface que permite ao utilizador interagir com as aplicações computacionais, como bases de dados ou sistemas periciais, usando a sua própria voz, através de um diálogo mais ou menos natural. Isto por oposição, por exemplo, aos sistemas de comando-e-controlo (*command and control*) por voz, ou os sistemas de ditado por voz, nos quais não existe um diálogo propriamente dito: o sistema limita-se a executar os comandos ou a transcrever o que o utilizador disse, respectivamente. Os sistemas conversacionais podem, assim, ser vistos como uma aplicação avançada das tecnologias de fala, incluindo áreas como o reconhecimento e síntese de fala, processamento da linguagem e gestão do diálogo.

Nem sempre a voz é a interface mais indicada para um determinado sistema. Por um lado, porque ainda está sujeita a erros de reconhecimento que reduzem o nível de satisfação do utilizador, e, por outro lado, existem tarefas, como por exemplo apresentar listas de dados ou opções, em que a voz é pouco aconselhável. A solução, como mostra o crescente número de aplicações com estas características, poderá passar por combinar a voz com outras modalidades (gráficos, texto, caneta e outras) criando assim interfaces multimodais. A actual tendência de redução do tamanho dos terminais computacionais móveis (PDAs, telemóveis, etc.) torna igualmente atractiva a possibilidade de interacção multimodal, pois as reduzidas dimensões dos aparelhos colocam dificuldades de usabilidade ao nível, por exemplo, do algo tão simples como o teclado.

1.3 Objectivo

As expectativas dos utilizadores em relação às aplicações de voz são muito elevadas e ainda não é tecnicamente possível corresponder a esse nível de expectativas. Pelo menos em interfaces com reconhecimento de fala. Nas aplicações telefónicas existe contacto “directo”, numa relação de um para um, entre a interface de voz e o cliente pelo que o mínimo erro na aplicação ou na qualidade da síntese de fala pode ser imediatamente detectado pelo utilizador e resultar em insatisfação por parte deste. No caso dos *call-centers*, por exemplo, se o utilizador está insatisfeito com a interface automática de voz, desliga a chamada ou opta pelo atendimento por parte de um operador humano e “perdem-se” quer o investimento feito no sistema automático quer a redução de custos que este proporciona. Assim este tipo de aplicações pode ser considerado, de certa forma, crítico para a empresa, pelo que a resposta para procurar garantir a sua qualidade é efectuar verificações e validações em vários pontos ao longo do processo de desenvolvimento. Testes completos e consistentes são, assim, extremamente importantes para se obter a qualidade desejada. Contudo no caso das interfaces de voz, a verificação e validação colocam desafios significativos e são campos de estudo ainda em aberto. É exactamente neste tema que se insere a presente tese de mestrado. É um campo demasiado vasto e complexo. Não é de todo imaginável desenvolver uma estratégia de testes para este tipo de aplicações, ou realizar esses mesmos testes, sem o conhecimento de áreas como as arquitecturas das aplicações, estrutura dos diálogos, funcionamento dos reconhecedores de voz, factores humanos envolvidos ou testes de usabilidade (só para mencionar algumas).

Apesar da necessidade de testes ao longo de todo o processo, como já referido, será dada ênfase ao nível de testes de sistema.

O objectivo principal é, assim, efectuar o levantamento do estado da arte nas áreas relevantes para o teste de aplicações de voz, transmitir os conhecimentos adquiridos ao longo dos últimos cinco anos de actividade profissional, documentar os projectos da área em que o autor esteve envolvido e propor metodologias e ferramentas adequadas tendo em vista a adopção das mesmas na organização onde o autor se insere.

1.4 Estrutura do documento

A presente dissertação encontra-se organizada em sete capítulos. No capítulo 1 apresenta-se uma introdução descrevendo a motivação para o trabalho desenvolvido e a sua contextualização. São também definidos os objectivos e identificados os desafios que se apresentam com a realização do trabalho.

No capítulo 2 são apresentados conceitos importantes para a compreensão do fenómeno fisiológico e neurológico da fala, por um lado, e de base para as áreas tecnológicas focadas no restante documento, por outro. São abordados temas como a neurolinguística, a semiótica ou a teoria da comunicação.

No capítulo 3 é apresentada uma breve história das tecnologias de fala, com especial incidência na síntese e reconhecimento automático da fala.

O capítulo 4 é um capítulo mais técnico sobre sistemas conversacionais: o que são, que arquitecturas são mais comuns, como funcionam os seus principais componentes e que tipo de testes são mais apropriados. É feita ainda uma breve abordagem aos aspectos de usabilidade que devem nortear o desenvolvimento destes sistemas interactivos.

O capítulo 5 detalha o tipo de testes mais importantes no processo de teste e avaliação de interfaces de voz e respectivos componentes.

No capítulo 6 são apresentados casos concretos de teste e avaliação de interfaces de voz nos quais o autor esteve envolvido na sua actividade profissional.

No capítulo seguinte, 7, é analisada a situação actual na empresa ao nível do processo de testes e são elaboradas algumas propostas tendo em vista a sua adopção por parte da

empresa, com a finalidade de reduzir os custos do processo de teste, melhorar a sua efectividade e o *know-how* técnico do departamento.

O oitavo e último capítulo é dedicado às conclusões e perspectivas de trabalho futuro, sendo realizado um resumo do trabalho desenvolvido e apontados possíveis caminhos de continuação.

2 Conceitos importantes

2.1 Linguagem e língua

Por vezes é notória alguma confusão entre conceitos como fala e voz, ou língua e linguagem. No entanto são conceitos importantes para o tema abordado no presente documento e como tal, é necessária uma clarificação dos mesmos. É disso que trata esta secção.

O Dicionário de Língua Portuguesa On-Line da Priberam define linguagem como “expressão do pensamento, por meio de palavra” ou “qualquer meio de exprimir o que se sente ou pensa”. Apesar de ser possível encontrar várias definições, que variam consoante as épocas, escolas de pensamento e/ou o detalhe pretendido, podemos dizer, de uma forma relativamente simplista, que a linguagem é a faculdade de expressão que faz uso de um sistema de signos convencionados que representam o mundo real, ideias ou pensamentos. Podemos considerar linguagens não-verbais, como a gestual; linguagens verbais, que incluem linguagens em código; e linguagens mistas. O estudo da linguagem tem o nome de linguística.

A língua é uma forma particular, uma instância, de linguagem, um sistema de signos vocais, que podem ser transcritos graficamente, comum a um povo/nação (exemplo: Português). Uma língua não é apenas um dicionário e uma gramática que esclarecem o modo como funcionam os "signos convencionados": é inseparável da cultura do local onde é falada, o que significa que é acompanhada de entoações, gestos, olhares e expressões faciais, entre outros tipos de linguagem. Também não é algo estático. Não só evolui com o tempo como também apresenta variações consoante o indivíduo que a usa e a sua intenção discursiva. Estas variações ocorrem a todos os níveis de uma língua: fonético, fonológico, morfológico, sintático e lexical.

A fala é a língua no momento em que está a ser usada por um indivíduo – fazendo uso dos sons, palavras e regras gramaticais dessa língua. Falar é não só o que é dito mas também o que é transmitido: entoação, pausas, ritmo, fluência e gestos (se o emissor está presente fisicamente).

Já a voz, é o som produzido pelo aparelho fonador Humano.

2.2 Teoria(s) da comunicação

Comunicar não é um aspecto accidental da vida humana, mas sim algo essencial e natural. Afinal, o ser Humano é um ser social.

A palavra comunicação deriva do latim *communicare* que significa “tornar comum”, “fazer saber”, “participar”. No âmbito do presente documento, e de forma concisa, comunicar pode ser visto como um processo de troca de informações, partilha de experiências ou sentimentos, entre um emissor e um receptor por meio de um sistema de signos (palavras, gestos, sons, imagens, etc.). No entanto a sua definição não é consensual, sendo o seu estudo um campo vasto que envolve diversas disciplinas. Entre as muitas correntes de investigação existentes, podemos identificar pelo menos duas cuja abordagem é de interesse para a área de interfaces de fala: uma que entende a comunicação sobretudo como fluxo linear de informação (troca de mensagens), e outra que se debruça mais sobre os signos usados na criação e estruturação das mensagens e seu significado.

A primeira corrente, normalmente designada por Teoria da Informação, deriva do conhecido e importante trabalho de Shannon e Weaver, *A Mathematical Theory of Communication* [Shannon 48]. Se hoje transmitimos palavras, imagens, dados, etc., em código digital para todo o mundo (e até para o espaço...) a velocidades elevadíssimas, tal

deve-se a este trabalho. Claude Elwood Shannon é, sem dúvida, um dos fundadores da “Era da Informação”. Formado em matemática e engenharia electrotécnica, desenvolveu um modelo, baseando-se em anos de trabalho e na álgebra booleana, no qual a informação é definida e reduzida à sua forma mais simples – uma situação sim/não, ou seja, um sistema binário de 1s e 0s – introduzindo assim o conceito bit, e identificou as relações críticas entre os vários elementos da comunicação. Trata-se de um modelo linear, que define os conceitos de emissor, receptor, mensagem, canal, ruído, redundância, etc., e que visa a eficiência do fluxo informativo. O importante nesta corrente são as questões da transmissão, destacando-se a quantificação da informação, e não tanto do significado ou estrutura das mensagens. A figura seguinte, retirada do artigo [Shannon 48], ilustra este modelo: uma fonte passa a informação a um transmissor que a coloca num canal (mais ou menos sujeito a ruído) que a leva a um receptor que a passa a um destinatário.

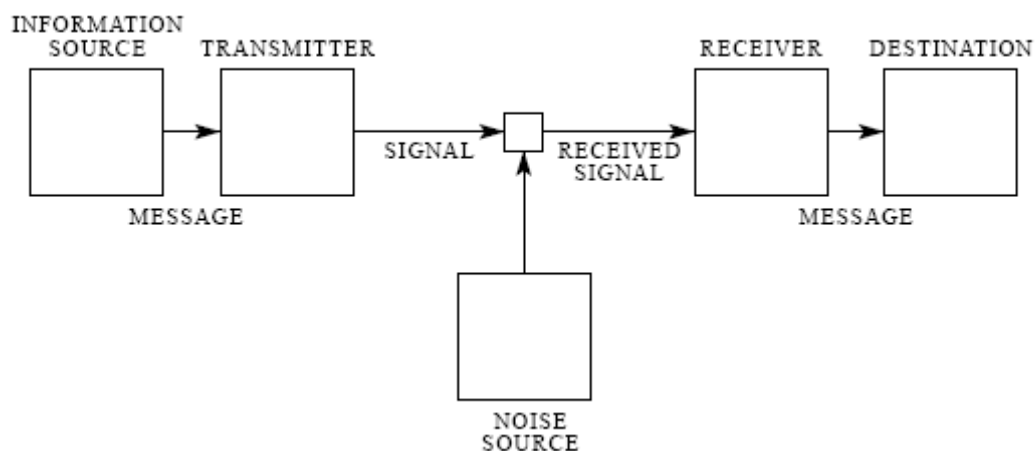


Figura 1 - Modelo de comunicação (Fonte: [Shannon 48])

Uma analogia possível é ver o cérebro como a fonte de informação, o aparelho fonador como transmissor, a voz como sinal, o ouvido do parceiro de comunicação como receptor e o seu cérebro como destinatário final. A mesma analogia é válida se um dos interlocutores é um sistema interactivo de fala.

- Quando o utilizador é o emissor, então o seu cérebro é a fonte, o seu aparelho fonador é o transmissor, o reconhecedor automático de fala é o receptor e, por exemplo, o módulo de gestão de diálogo é o destinatário final;
- Quando o utilizador está no papel de receptor, o módulo de gestão do diálogo é a fonte, o sintetizador de fala é o emissor, o ouvido do utilizador é o receptor e o seu cérebro é o destinatário final.

A segunda corrente mencionada anteriormente é a semiótica. A semiótica estuda os signos: a sua natureza, tipos e funções. De forma simplista, um signo é qualquer elemento (imagem, som) que seja utilizado para exprimir um conceito (realidade física ou psicológica); nesta relação, o primeiro funciona como significante em relação à segunda, que é o significado (ou referente). No modelo semiótico é dada ênfase aos significados e à formação das mensagens a transmitir. A comunicação não é vista de forma linear como um fluxo do emissor até ao receptor, mas antes com um sistema de signos. O estudo da comunicação passa, assim, pelo estudo dos signos utilizados, dos códigos em vigor, das culturas em que os signos se criam (só para mencionar os aspectos mais importantes) sendo este estudo realizado a três níveis:

- sintáctico;

- semântico;
- pragmático (estuda, por exemplo, o valor dos signos para os utilizadores).

Estas correntes têm implicações no desenvolvimento, integração e avaliação de sistemas interactivos de fala. Quem projecta, desenvolve e testa este tipo de sistemas deve ter presente e compreender os conceitos envolvidos na comunicação, tendo em vista a eficiência da interacção de fala e aceitação do sistema por parte dos utilizadores.

O acto de comunicar implica a existência dos elementos nucleares do acto comunicativo: o emissor, o receptor e a mensagem. É, no entanto, normal considerar mais três elementos: o código, o canal e o contexto. Para haver comunicação é necessária a intervenção de, pelo menos, dois parceiros, um que emita, outro que receba; algo tem que ser transmitido pelo emissor ao receptor; para que o emissor e o receptor comuniquem é necessário que esteja disponível um canal através do qual se realize a comunicação; a informação a transmitir tem que estar "codificada" (as ideias convertidas em signos) de forma perceptível quer para o emissor, quer para o receptor; e, finalmente, todo o acto comunicativo realiza-se num determinado contexto e é determinado por esse contexto. Uma mensagem tem muito mais hipóteses de ser compreendida se disser respeito, simultaneamente, ao campo de experiência do emissor e do receptor. A mensagem é a ideia transmitida e o código é a língua, que tem uma sintaxe, uma semântica e uma pragmática. A combinação da mensagem com o código utilizado é o sinal. O canal é o meio onde se propaga o sinal, como por exemplo o ar. O contexto pode conter um grande número de componentes que interferem na comunicação, indo de factores subjectivos ou psicológicos – como a motivação ou o conhecimento que os interlocutores possuem do domínio –, até factores físicos – como o ruído. Designa-se por ruído tudo aquilo que afecta a transmissão de informação, incluindo uma voz baixa, uma articulação deficiente ou o barulho ambiental.

2.3 Neurolinguística

O organismo Humano é um sistema em interacção constante com o meio. Deste recebe entradas que se traduzem por estímulos. Estes estímulos são processados e o comportamento é o conjunto de respostas. Neste processo estão envolvidos três estruturas principais: os mecanismos de recepção (os sentidos), os mecanismos de conexão (o sistema nervoso) e os mecanismos de reacção (os músculos ou as glândulas, por exemplo).

Juntamente com a espinal-medula, o cérebro faz parte do sistema nervoso central. É no cérebro que reside a memória, a aprendizagem, o pensamento ou a linguagem. É constituído por cerca de 12 biliões de neurónios e cada neurónio pode chegar a ter 5000 sinapses, o que lhe possibilita processar simultaneamente um número incalculável de informação. A sua complexidade é extraordinária, e apenas um pequena percentagem do seu funcionamento é actualmente compreendido. Geralmente dividem-se os órgãos encefálicos em três estruturas: cérebro posterior, cérebro médio e cérebro anterior, do qual faz parte o córtex cerebral que cobre as outras zonas do cérebro. O córtex cerebral está dividido em hemisfério esquerdo e hemisfério direito. Cada hemisfério, por sua vez, apresenta quatro lobos: frontal, parietal, temporal e occipital. Desde 1861, em que Paul Broca detectou a ligação entre a lesão numa área do cérebro e a perda da fala num homem, que se tem tentado identificar as relações entre as áreas cerebrais e as funções principais e comportamentos humanos: área motora, área visual, área auditiva – só para mencionar as mais relevantes para o tema do presente documento.

No modelo clássico da linguagem, é fundamentalmente o hemisfério esquerdo do cérebro que controla a compreensão e produção de linguagem. Neste modelo, no hemisfério esquerdo existe uma área frontal expressiva para planeamento e execução da fala, designada por área de Broca (Pierre Paul Broca, 1861), e uma área posterior, receptiva, para análise e

identificação dos estímulos linguísticos sensoriais, designada por área de Wernicke (Carl Wernicke, 1874). Uma lesão total da área de Broca traduz-se na incapacidade de articular palavras (afasia de Broca), ao passo que uma lesão total na área de Wernicke provoca a incapacidade de compreender o que é dito (agnosia auditiva) e emissão de palavras desprovidas de sentido (afasia de Wernicke). As áreas de Broca e Wernicke estão ligadas por um feixe de fibras denominado fascículo arqueado. Lesões no fascículo arqueado provocam a incapacidade de repetir palavras [Pereira 03]. Apesar do hemisfério esquerdo ser predominante, no hemisfério direito, no local onde do lado esquerdo estão as áreas de Broca e Wernicke, encontram-se as áreas que controlam, respectivamente, a compreensão e produção da prosódia.

As áreas de Broca e Wernicke são imprescindíveis, mas, não suficientes, pois uma coisa é ser capaz de articular e reconhecer palavras e outra é atribuir conteúdo semântico a essas palavras. O modelo clássico é, assim, insuficiente para compreendermos os mecanismos cerebrais da linguagem [Binder 97]. Estudos feitos por neurologistas como o português António Damásio e a sua esposa Hanna mostraram que lesões no córtex podem levar à perda de capacidades específicas na linguagem: algumas lesões no lobo temporal podem levar à incapacidade em pronunciar substantivos, outras ao nível dos lobos frontal e parietal acarretam problemas em lidar com os verbos, por exemplo [Damásio 92] e [Damásio 93].

As impressões auditivas são recebidas na parte superior do lobo temporal. Na área auditiva primária são recebidos os sons elementares, detectando características como volume e altura. A área psicoauditiva interpreta-os, identifica-os, analisa a informação, reconhecendo um som completo, como palavras que são depois enviadas para a área da linguagem. Uma lesão na área auditiva primária provoca surdez, enquanto que uma lesão na área psicoauditiva provoca a impossibilidade de atribuir o significado a um som, apesar de o ouvir.

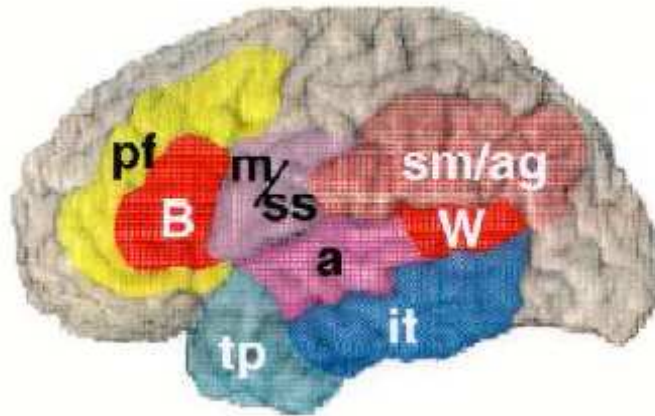


Figura 2 - Áreas do cérebro importantes na linguagem (Fonte: [Damasio 99])¹

A linguagem não é uma função independente, estando relacionada e dependente de outros tipos de funções, em que interferem a informação sensorial visual e auditiva, a atenção, a realização de operações comparativas e outras operações sobre a informação obtida, a selecção de uma resposta baseada nessas operações e a construção da resposta.

A ciência que estuda os mecanismos cerebrais que estão subjacentes à compreensão, produção e conhecimento da linguagem é a neurolinguística.

¹ Legenda da figura: B=área de Broca; W=área de Wernicke; a=área auditiva; tp=polo temporal; it=região inferotemporal; m/ss=área motora e somatostésica; pf=áreas pré-frontais; sm=área motora suplementar;

2.4 A voz Humana

O ser Humano não possui nenhum aparelho específico destinado exclusivamente à produção do som. A questão que se coloca é, então, como se processa a produção da voz. O ar inspirado passa pelas cordas vocais em posição aberta, enchendo os pulmões. Na expiração é que ocorre a fonação, quando o ar passa pelas cordas vocais em posição fechada em direcção ao tracto vocal (faringe, cavidade oral e cavidade nasal). A forma do tracto vocal é modelada pelo véu palatino, a língua, lábios e maxilares – os chamados articuladores – criando assim câmaras de ressonância.

Ao conjunto formado pelo tracto vocal, cordas vocais e pulmões dá-se, normalmente o nome de aparelho fonador, ilustrado na figura abaixo, à qual se segue uma breve descrição.

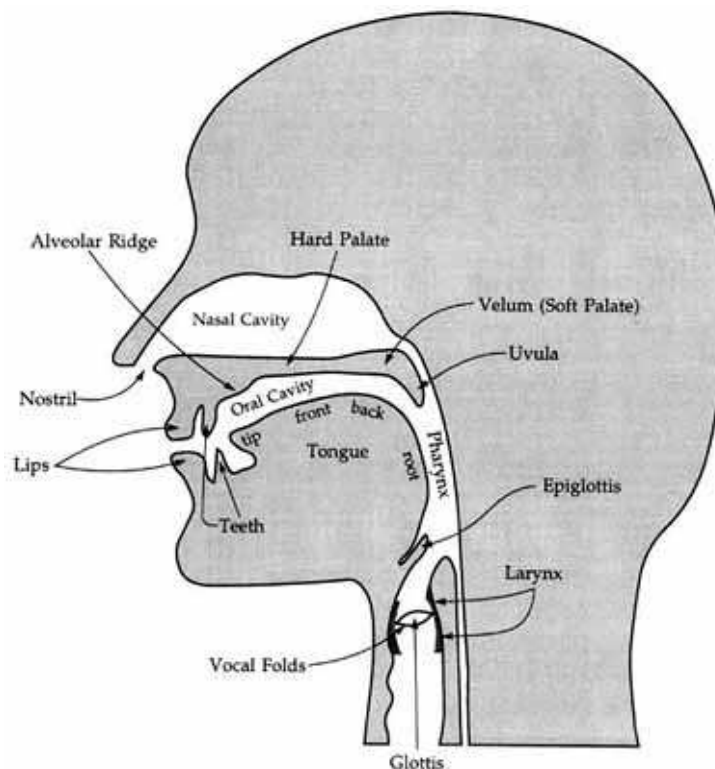


Figura 3 - Visão parcial do aparelho fonador

Os pulmões, os brônquios e a traqueia são os órgãos respiratórios que permitem a corrente de ar. A maioria dos sons que conhecemos é produzida na fase de expiração, servindo a inspiração como momento de pausa. Ao expirar, os pulmões libertam ar que passa pelos brônquios para entrar na traqueia e chegar à laringe (*Larynx*, na figura).

Na laringe o ar encontra a glote (*Glottis*), mais ao menos ao nível da maçã-de-adão, que é a abertura nas mais conhecidas cordas vocais (*Vocal Folds*). Semelhantes a duas pregas musculares, as cordas vocais podem estar fechadas ou abertas: se estiverem abertas, o ar passa sem obstáculo real, dando origem a um som surdo; se estiverem fechadas, o ar força a passagem fazendo as pregas musculares vibrar, o que dá origem a um som. A tonalidade é a frequência de excitação das cordas vocais. Para sons em que as cordas vocais não são excitadas, não existe tonalidade. O valor da tonalidade é diferente para homens, mulheres e crianças. Podemos falar também nos sons não-vozeados, nos quais não houve vibração do ar, e os sons vozeados.

Depois de sair da laringe, o ar entra na faringe (*Pharynx*) onde encontra uma encruzilhada: primeiro a entrada para a boca e depois para as fossas nasais. No meio está o véu palatino (*Velum*) que permite que o ar passe livremente pelas duas cavidades, originando

um som nasal; ou que impede a passagem pela cavidade nasal, obrigando o ar a passar apenas pela cavidade bucal – resultando num som oral.

Por fim, o ar está na cavidade bucal (a boca) que funciona como uma caixa de ressonância onde, usando os maxilares, as bochechas e, especialmente, a língua e os lábios, é possível modular-se uma infinidade de sons.

Recorrendo a conhecimentos de acústica é possível prever com razoável aproximação as frequências e amplitudes relativas das ressonâncias produzidas pelas diferentes caixas de ressonância modeladas na cavidade bucal. Estas ressonâncias aparecem sobrepostas no espectrograma à frequência fundamental (normalmente designada F0) de vibração produzida pelas pregas vocais, sob forma de bandas horizontais mais escuras (picos) e têm o nome de formantes. Para um indivíduo do sexo masculino, a frequência fundamental situa-se tipicamente na casa dos 80 a 200Hz e, no caso feminino, entre os 150 e os 350Hz.

As formantes são frequentemente utilizadas para detectar os fonemas, considerando-se para tal suficientes as 3 primeiras formantes. Este conceito (fonema) parte do princípio de que o som vocal elementar que realmente interessa na comunicação linguística é um pequeno número de propriedades articulatórias e acústicas (os traços) e não todo o conjunto de emissão fónica. Esses traços permitem distinguir os sons elementares dos outros. Cada fonema é assim um conjunto de certos traços distintivos, e representa a unidade mínima de som capaz de produzir modificações de sentido de palavras por mera comutação.

2.4.1 Fonética e fonologia

A fonologia estuda os sistemas sonoros das línguas. De entre o conjunto de sons que o aparelho vocal pode produzir, e que é o campo de estudo da fonética, apenas um subconjunto relativamente pequeno é usado em cada língua de forma distinta. O domínio desse conhecimento é necessário ao desenvolvimento dos sistemas de reconhecimento e síntese de fala. A fonologia do português é composta por cerca de 33 a 38 fonemas, variando conforme o local. Os fonemas da língua portuguesa classificam-se em vogais, semivogais e consoantes:

- Vogais: são fonemas pronunciados sem obstáculo à passagem de ar, chegando livremente ao exterior pela boca, ou pela boca e fossas nasais;
- Semivogais: são os fonemas que se juntam a uma vogal, formando com esta uma só sílaba. Exemplos: couro, baile – as vogais “u” e “i” transformam-se em semivogais porque estão junto de outra vogal numa mesma sílaba. Quando duas vogais se encontram na sílaba e uma se transforma em semivogal, temos um ditongo;
- Consoantes: são fonemas produzidos mediante a resistência que os órgãos bucais (língua, dentes, lábios) opõem à passagem de ar.

Baseando-nos na gramática portuguesa, as vogais costumam ser classificadas da seguinte forma:

- quanto à zona de articulação
 - o anteriores (ou palatais)
 - o médias
 - o posteriores (ou velares)
- quanto ao timbre
 - o abertas
 - o fechadas
 - o reduzidas
- quanto ao papel das cavidades bucal e nasal

- ☐ orais
 - ☐ nasais
- quanto à intensidade
 - ☐ átonas
 - ☐ tónicas

Já os ditongos podem ser classificados com:

- crescentes/decrescentes
- orais/nasais
- abertos/fechados

Finalmente, e sempre tendo por base a gramática portuguesa, para as consoantes a classificação normalmente adoptada é a seguinte:

- quanto ao modo de articulação
 - ☐ oclusivas (quando o ar encontra um obstáculo total)
 - ☐ constrictivas (quando o ar encontra um obstáculo parcial)
 - fricativas – caracterizada pela passagem do ar através de uma estreita fenda na via bucal, criando turbulências no ar
 - laterais – caracterizadas pela passagem do ar pelos dois lados da cavidade bucal
 - vibrantes – caracterizadas pelo movimento vibratório rápido da língua ou véu palatino, criando uma série de brevíssimas oclusões
- quanto ao ponto de articulação
 - ☐ bilabiais
 - ☐ labiodentais
 - ☐ linguodentais
 - ☐ alveolares
 - ☐ palatais
 - ☐ velares
- quanto ao papel das cordas vocais
 - ☐ surdas
 - ☐ sonoras
- quanto ao papel das cavidades bucal e nasal
 - ☐ orais
 - ☐ nasais

Na maior parte das línguas, o texto escrito (grafemas) não corresponde à forma como é pronunciado, pelo que é necessária uma representação simbólica para descrever a pronúncia correcta. Cada língua tem um alfabeto fonético, com um diferente conjunto de fonemas, e diferentes combinações desses fonemas. O alfabeto fonético de uma língua pode ser definido como o conjunto mínimo de símbolos necessário para descrever qualquer palavra possível dessa língua. Esta representação, no entanto, não é perfeita porque a fala é um sinal contínuo e a notação fonética é discreta. Além disso durante a fala, os movimentos articulatórios de um determinado fonema dependem dos fonemas precedentes e dos seguintes (fenómeno conhecido por coarticulação), pelo que podem ocorrer ligeiras variações na pronúncia desse fonema. As diversas emoções e características do orador são também impossíveis de serem descritas através de fonemas e por isso introduz-se a noção de sons dos fonemas.

Os fonemas são categorias linguísticas abstractas que nos permitem agrupar subconjuntos de sons da fala (fones). Embora não haja dois fones iguais, todos os fones

agrupados numa determinada categoria de fonemas são suficientemente semelhantes para terem o mesmo significado. Dá-se o nome de difone a um par adjacente de fones e são bastante úteis na síntese de fala porque combinando difones pré-gravados soa bastante mais natural que combinando apenas simples fones, uma vez que a pronúncia de cada fone varia conforme os fones que lhe são adjacentes.

Têm sido desenvolvidos esforços no sentido de criar alfabetos fonéticos independentes da língua. Um dos mais conhecidos é o Alfabeto Fonético Internacional (IPA – *International Phonetic Alphabet*). Há duas maneiras de inserir os caracteres do alfabeto em textos:

- para representar os fonemas, usamos a transcrição fonológica, com os caracteres entre barras;
- quando queremos representar os sons dos fonemas, usamos a transcrição fonética, com os caracteres entre colchetes.

Inicialmente, dada a sua complexidade e o uso de símbolos gregos, a sua manipulação computacional era mais difícil, mas hoje em dia a sua aceitação é generalizada.

2.4.2 Representação digital

A voz humana é um sinal analógico e, como tal, possui características que variam continuamente durante o tempo. Entre as características acústicas mais importantes da fala temos a frequência, a amplitude, estrutura harmónica e ressonância.

Todos os sons são perturbações das moléculas do ar que podem ser representadas graficamente por ondas sinusoidais. Fisiologicamente, isto corresponde à vibração do tímpano (ver secção seguinte do presente documento). O número de vibrações por segundo é a frequência e o volume do som é a amplitude. A maior parte dos sons, incluindo a voz, possui uma frequência predominante (frequência fundamental) e um conjunto de frequências secundárias sobrepostas. São estas frequências secundárias que nos permitem distinguir as vozes de pessoas diferentes. Algumas destas frequências secundárias, chamadas formantes, são importantes também para distinguir os diferentes fonemas e são produzidas por ressonância. Como vimos anteriormente, a cavidade bucal funciona como uma câmara de ressonância: usando os maxilares, as bochechas e, especialmente, a língua e os lábios, podem modular-se uma infinidade de sons.

A complexidade das formas de onda dos fonemas e a constante alternância de fonemas torna extremamente difícil a sua análise a partir da representação sinusoidal. Por isso usam-se os espectrogramas, nos quais os padrões harmónicos e ruído sobressaem. No eixo vertical temos as frequências do sinal e no eixo horizontal vemos como essas frequências variam no tempo conforme a fala decorre. Frequências de maior amplitude aparecem mais escuras: frequência fundamental e formantes. A localização das formantes reflecte o tamanho, posição e forma do tracto vocal no momento em que o fonema é dito. Esta localização, a distância entre as formantes e a forma como elas variam conforme a fala é produzida são dados críticos para o reconhecimento automático da fala. A amplitude e duração do ruído também fornece pistas para a natureza dos fonemas mas devem ser complementadas com a informação das formantes mais próximas.

Para que seja possível a manipulação da voz em um ambiente computacional, é necessário que, de alguma forma, a mesma seja capturada e transformada em um sinal digital. No entanto a quantidade de informação num sinal de voz é enorme e, para o reconhecimento automático funcionar a uma velocidade aceitável, essa quantidade tem de ser reduzida. Felizmente, alguma informação num sinal de voz é redundante, outra é irrelevante e outra ainda tem de ser removida para o correcto funcionamento do sistema de reconhecimento. O desafio consiste em eliminar este tipo de informação do sinal sem perder ou distorcer as partes fundamentais. Uma das técnicas utilizadas é o uso de filtros com a função de seleccionar a

gama de frequências que interessam. Praticamente toda a informação contida na fala situa-se entre os 200Hz e os 8kHz

Ao processo de conversão do sinal analógico para formato digital dá-se nome de digitalização. A digitalização é feita através da amostragem (*sampling*), que é a captura e medição da amplitude do sinal de áudio em intervalos regulares de tempo – pelo teorema de Nyquist, a frequência de amostragem deve ser pelo menos o dobro da frequência mais alta com interesse para o reconhecimento. Após a amostragem é necessário quantificar e codificar a informação. As amostras analógicas podem tomar uma infinidade de valores, sendo necessário quantificar estes valores para que possam ser representados por uma quantidade finita de bits, para obter um sinal digital. Esta conversão é feita por um circuito chamado conversor analógico-digital A/D. Cada amostra é transformada numa quantidade predefinida de *n bits*. Após a quantificação é necessário codificar as sequências de bits, até porque a capacidade dos meios de transmissão (dada pela Lei de Shannon) e/ou armazenamento dos dados é finita, e precisamos encontrar um meio-termo entre duas necessidades antagónicas: diminuir a quantidade de bits necessária para a representação da informação e preservar sem distorção a informação crítica.

2.5 Ouvido Humano

O ouvido é constituído por um conjunto de órgãos cuja finalidade é a percepção dos sons e a manutenção do equilíbrio. É composto por três partes: o ouvido externo, o ouvido médio e o ouvido interno.

O ouvido externo tem dois componentes: o pavilhão auricular, também conhecido como orelha, e o canal auditivo externo. A função principal do pavilhão auricular é captar sons, direccionando-os para o canal auditivo. Outra função é a de filtrar o som, processo este que ajuda a localizar a origem dos sons que chegam ao indivíduo. Além disso, no caso dos humanos, este processo selecciona sons na faixa de frequência da voz humana facilitando o entendimento. Já o canal auditivo externo tem a função de transmitir os sons captados pela orelha para o tímpano.

O ouvido médio é uma cavidade estreita, situada interiormente à membrana do tímpano, que contém 3 ossículos articulados (Martelo, Bigorna e Estribo) que ligam o tímpano à parte interna do ouvido. O tímpano transmite a vibração do ar captada pela orelha para os ossículos que por sua vez vibram transmitindo as suas vibrações ao ouvido interno.

O ouvido interno é composto, essencialmente, pelo caracol (cóclea) e pelo aparato vestibular. A cóclea é um canal que tem um formato espiralado e é oco sendo preenchido por fluidos. A energia mecânica das vibrações que chegam à cóclea provenientes do ouvido interno é transformada em impulsos nervosos pelo o órgão de Corti. Estas correntes estimulam as terminações do nervo auditivo e assim chegam ao córtex temporal através das vias e centros nervosos.

A frequência e a intensidade são tratadas diferenciadamente na cóclea. Assim, cada tom, segundo a sua frequência, é captado por um local diferente do órgão de Corti.

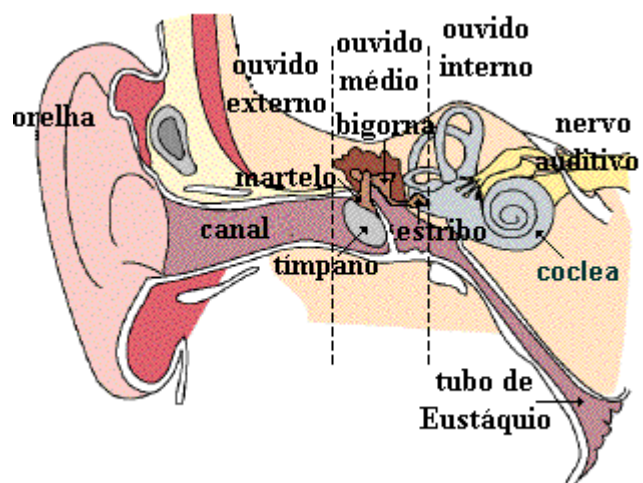


Figura 4 - Ouvido Humano

3 Breve história das tecnologias de fala

A produção de voz humana de forma artificial tem sido um objectivo perseguido desde tempos bastante remotos. Por exemplo, já desde a Grécia Antiga que se realizavam experiências tendo em vista a geração de vozes artificiais: na maior parte dos casos eram simples conjuntos de tubos, outras vezes engenhos bastante complicados. Mais no campo do mito do que realidade, são conhecidos os exemplos de Gerbert d'Aurillac – mais tarde, Papa Silvestre II – (950-1003), Albertus Magnus (1198-1280) e Roger Bacon (1214-1294).

Gerbert d'Aurillac, além de clérigo era também um cientista, tendo estudado matemática em Barcelona e nas, então, islâmicas Córdoba e Sevilha. Diz a lenda que também terá estudado feitiçaria e teria, supostamente, criado uma cabeça em bronze que respondia a questões com as palavras “sim” ou “não”.

Albertus Magnus, também conhecido como Santo Alberto o Grande, tornou-se famoso pelo seu conhecimento universal e pela sua demanda por uma coexistência pacífica entre a ciência e a religião. Ele era um dos alquimistas com a fama de ter conseguido descobrir a Pedra Filosofal. Foi também acusado de feitiçaria, tendo alegadamente criado um autômato capaz de falar.

Roger Bacon foi um estudioso, filósofo e professor, que mais tarde se tornou clérigo. A sua fama advém essencialmente do seu brilhante intelecto, tendo antecipado invenções como os óculos, telescópios, ou máquinas voadoras. Tinha também a fama de possuir conhecimentos ocultos e que teria criado uma cabeça de bronze que podia responder a qualquer pergunta.

Para encontrar exemplos concretos e documentados de esforços para produzir voz humana artificial temos de avançar no tempo e situarmo-nos nos finais do século XVIII. Em 1779, o dinamarquês Christian Kratzenstein explicou as diferenças fisiológicas das cinco vogais e criou um aparelho mecânico semelhante ao trato vocal humano capaz de as produzir artificialmente.

Por essa altura, Wolfgang von Kempelen, ao serviço da imperatriz da Áustria, teria já começado (desde 1769 ao que se julga) as suas próprias tentativas de criar uma máquina que falasse, com fins terapêuticos. O seu trabalho foi apresentado em 1791. A sua máquina de fala acústico-mecânica (*Acoustic-Mechanical Speech Machine*) foi a primeira a produzir não só alguns sons próprios da fala, como também palavras completas e até frases curtas! A constituição da máquina procurava reproduzir o aparelho fonador Humano, desde os pulmões, passando pelas cordas vocais, tracto vocal com lábios e até uma espécie de nariz tinha. Para se produzirem os sons tinha que se operar com todos estes elementos. Esta máquina encontra-se preservada no *Deutsche Museum* em Munique. A figura seguinte, retirada dos apontamentos do próprio von Kempelen, ilustra esta máquina:

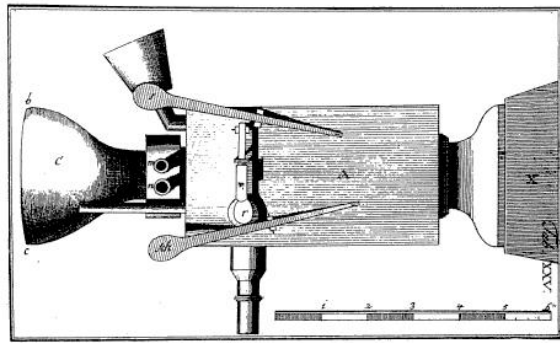


Figura 5 - Máquina de von Kempelen

Os trabalhos de von Kempelen vieram também mostrar a importância do tracto vocal ao nível da produção dos sons da fala. Antes dele, acreditava-se que toda a actividade da fala estava centrada nas cordas vocais.

Apesar do brilhantismo e importância desta invenção, na época não foi dada a devida atenção pois anos antes von Kempelen tinha espantado o mundo quando inventou um autómato capaz de jogar xadrez (o “Turco”) cujo “mecanismo” era, na verdade, conforme se descobriu muito posteriormente, uma pessoa no seu interior... Este autómato era considerado a ilusão mais famosa da história, chegando a fazer *tournées* pela Europa fazendo exhibições perante multidões...

No século 19, foram criadas mais algumas máquinas similares à de von Kempelen, mas sem nenhum avanço fundamental a registar. Um dos exemplos mais conhecidos foi o da máquina criada por Charles Wheatstone, um inventor inglês bastante profícuo, sendo-lhe atribuídas invenções como a concertina ou o estereoscópio. Aperfeiçoando os desenhos de von Kempelen, criou uma máquina que produzia uma maior gama de sons que a máquina original. Digno de registo também foi a máquina apresentada por Joseph Faber na mesma altura, que, como particularidades, incluía um modelo da língua e era controlada por pedais e por um teclado. Diz-se que esta máquina, chamada *Euphonia*, não só produzia fala como também cantava o hino inglês.

A partir do início do século 20, com os avanços então registados, começou a ser possível pensar em sintetizar os sons da fala a partir de meios eléctricos, por oposição aos meios mecânicos usados até então. Em 1920, por exemplo, foi apresentado pela Elmwood Button Co. o *Rádio Rex*, que era um brinquedo constituído por uma casota e um cão dentro dela. Assim que o utilizador chamava (literalmente) pelo cão este saltava cá para fora. Era uma aplicação arcaica de reconhecimento de fala mas é um marco que representa a primeira aplicação desta tecnologia.

Em 1930, o cientista húngaro Tihamer Nemes tentou registar uma patente para o desenvolvimento de um sistema automático de transcrição de fala. O pedido foi recusado por ter sido considerado irrealista...

No início dos anos 30, nos laboratórios Bell, Homer Dudley desenvolve o *VOCODER* (ou *Voice Coder*). O *VOCODER* tinha por objectivo reduzir a largura de banda necessária para a transmissão de fala, para que um maior número de chamadas telefónicas pudesse ser transmitido numa determinada linha. A redução na quantidade de informação necessária era conseguida através da codificação da voz em séries de números que representavam as variações dos parâmetros acústicos da portadora (frequência fundamental) enquanto o sujeito falava. Ou seja, só era necessário guardar/transmitir as modificações nas frequências. Para “reconstruir” a fala, bastava simplesmente inverter o processo e a fala assim produzida era inteligível (mas longe de perfeita...). O *VOCODER*, ou mais precisamente, o início da utilização dos meios eléctricos, marca também o início da investigação na área do

reconhecimento de fala. Nessa altura os projectos de investigação eram financiados pelo governo dos Estados Unidos da América (mais concretamente pelo Departamento de Defesa, através do projecto DARPA - *Defense Advanced Research Projects Agency*). A ideia inicial era usar este tipo de tecnologia para espionagem e descodificação de mensagens. Os primeiros resultados foram aparecendo ao longo da década de 40. No entanto, só a partir dos anos oitenta é que as tecnologias de reconhecimento de fala surgiram comercialmente.

Em 1936, precisamente no dia 24 de Julho, a *Telephone Company* do Reino Unido (hoje BT – *British Telecom*) inaugurou o seu relógio falante (*Speaking Clock*, também conhecido por TIM), que consistia num conjunto de discos de vidro, motores e válvulas que ocupavam uma sala inteira. O seu funcionamento consistia, basicamente, em concatenar de forma correcta as palavras pré-gravadas. Logo no primeiro ano de funcionamento registou mais de 13 milhões de chamadas e ainda hoje recebe mais de 135 milhões de chamadas por ano.

Em 1937, Homer Dudley aperfeiçoou o *VOCODER*, criando o *VODER* (de *Voice Demonstrator*) que foi apresentado na exposição mundial de Nova Iorque, em 1939. A figura seguinte ilustra a apresentação do *VODER*.

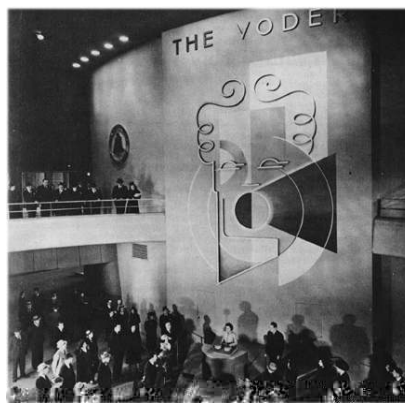


Figura 6 - Apresentação do *VODER* em Nova Iorque, 1939

Este foi o primeiro aparelho deste género a ser considerado um sintetizador de fala e a atrair grande atenção por parte do público. No entanto a sua operação era difícil e requeria muito tempo de treino. De qualquer maneira, estava demonstrada a possibilidade real de se criar voz artificial. A partir daqui, os sintetizadores mecânicos foram sendo gradualmente substituídos por sintetizadores eléctricos, embora ainda se faça pesquisa na área dos mecânicos.

Foi necessário esperar cerca de uma década até que algo inovador aparecesse. Em 1951, Franklin Cooper e a sua equipa desenvolveram um sintetizador de reprodução de padrões (*Pattern Playback*). Ao fim ao cabo, o que fazia era transformar espectrogramas previamente obtidos em sons (não era só voz...).

Em 1952, um novo marco nas tecnologias de voz: nos laboratórios Bell era lançado o *AUDREY* (de *Automatic Digit Recognizer*). O *AUDREY* tinha por objectivo fazer o reconhecimento automático de dígitos isolados ditos por um único sujeito e foi construído com componentes analógicos. O seu modo de funcionamento consistia em analisar o espectro do som produzido pelo orador, “procurando” as ressonâncias típicas dos dígitos, usando para tal um conjunto de filtros. Apesar de necessitar de afinações extensas para funcionar, tinha uma taxa de erro inferior a 2%. Em 1958 a ideia foi adaptada ao reconhecimento de vogais e consoantes.

Em 1953, Walter Lawrence apresentou o primeiro sintetizador de formantes, também conhecido por sintetizador de ressonâncias, com o nome de *PAT* (*Parametric Artificial Talker*). O *PAT* era constituído por três filtros dispostos paralelamente que simulavam as

propriedades de ressonância do trato vocal (ressonadores). Tinha como entrada um sinal sonoro ou besouro e depois controlavam-se as frequências da frequência fundamental e das três primeiras formantes para se obter a fala (ver secção 2.4 do presente documento). Mais ou menos na mesma altura, Gunnar Fant introduz o OVE I (*Orator Verbis Electris*), que tinha os ressonadores não em paralelo, mas em cascata, representando secções do trato vocal.

Em 1962, a IBM apresentou na exposição mundial de Seattle a *ShoeBox*, que não era mais que um calculadora com interface de voz (dígitos e comandos).

Nos meados da década de 60, com o aparecimento nos laboratórios dos conversores AD/DA (*Analog-Digital/Digital-Analog*) e dos primeiros computadores, foi possível introduzir novos e determinantes avanços nas tecnologias de fala. Em 1968, o primeiro sistema de síntese de texto para fala para Inglês foi desenvolvido no Japão por Noriko Umeda. Este sistema era baseado num modelo articulatório e incluía um módulo de análise sintáctica. A fala produzida era bastante inteligível, embora monótona e com uma qualidade ainda distante da apresentada pelos sistemas actuais. Os constantes avanços nos computadores tornaram possível desenvolver sistemas de síntese de fala com cada vez mais qualidade e cada vez mais práticos. Também nesse ano, 1968, estreou o célebre filme de ficção científica “2001 – Odisseia no Espaço” no qual aparecia o HAL 9000 (ou simplesmente HAL). HAL era um computador onisciente, capaz de manter um diálogo perfeito com interlocutores humanos. O HAL tornou-se a referência pela qual, no mundo real, os sistemas de inteligência artificial e de reconhecimento e síntese de fala passaram a ser avaliados. Teve também o condão de despertar o imaginário das pessoas para as tecnologias de fala e colocou as suas expectativas em patamares que ainda hoje, quase 40 anos depois, estamos muito longe de atingir... De facto, as capacidades do HAL excediam muito aquilo que se costuma considerar actualmente como reconhecimento e síntese de fala: ele era capaz de a “compreender”, o que implica muito mais que converter fala em texto, e era capaz de gerar fala a partir de conceitos, o que também ultrapassa a mera conversão texto-fala.

Nesta altura, começaram também as primeiras experiências com a tecnologia LPC (*Linear Predictive Coding*). Inicialmente a sua qualidade era bastante baixa, mas, com sucessivos melhoramentos, ainda hoje é utilizada com bons resultados. No início dos anos 70, surgem métodos e algoritmos, como o algoritmo Viterbi, os modelos de Markov ocultos (HMM – *Hidden Markov Models*), FFTs (*Fast Fourier Transform*) e *Dynamic Time Warping* (DTW), que representaram avanços extraordinários nas tecnologias de fala, mas que só começam a ser aplicados mais regularmente na década de 80 – os quais, juntamente com as redes neuronais, formam a base do actual estado da arte das tecnologias de voz. Também no início desta década foi desenvolvido o *VIP 100*, considerado o primeiro produto de reconhecimento de fala, que demonstrou a viabilidade do reconhecimento de fala dependente do utilizador, pequeno vocabulário e discurso discreto.

Em 1978 a Texas Instruments introduziu o seu famoso brinquedo *Speak & Spell*, que constituiu um dos primeiros aparelhos destinado ao grande público a incluir um *chip* que podia sintetizar a fala.

Em 1979 o M.I.T. (*Massachusetts Institute of Technology*) apresentou o sistema de síntese texto-fala *MITalk*. Dois anos mais tarde, Dennis Klatt introduz o seu *Klattalk*. A tecnologia presente nestes dois sistemas ainda forma a base de muitos sistemas comerciais actuais. Da década de 70, datam também as primeiras experiências de sistemas de reconhecimento contínuo de fala, como o projecto *SUR - Speech Understanding Research* da DARPA, de 1971, com um financiamento de 15 milhões de dólares para 5 anos. Dos sistemas desenvolvidos no âmbito desse projecto, apenas um cumpria os objectivos propostos: o *Harpy* da *Carnegie Mellon University*, que reconhecia 1011 palavras com uma taxa de erro de 5%. No entanto, todos os sistemas desenvolvidos tiveram enorme impacto na área.

Na década de 80, com as técnicas já mencionadas, dá-se o amadurecimento do reconhecimento de fala contínuo, começando também o reconhecimento independente do

orador a dar os primeiros passos. Em 1986 é lançado o primeiro sistema de reconhecimento de fala comercial de vocabulário muito grande, independente do orador. É também nesta década que o reconhecimento de fala começa a ser testado em telefonia.

A década de 90 fica marcada por avanços nas áreas da robustez ao ruído, adaptação ao orador, refinamentos no modelo HMM, naturalidade e inteligibilidade da síntese de fala, entre outras. A maturidade e a importância que as tecnologias de fala atingiram fica bem patente no número de produtos comerciais lançados, na criação de diversas empresas na área, fusões entre elas ou aquisições.

Actualmente, as tecnologias de voz ficam marcadas por avanços na área da verificação do orador (*Speaker Verification*), mobilidade – em qualquer lugar, em qualquer altura –, multiplicidade de línguas, integração em automóveis, PDA's, brinquedos e electrodomésticos.

No actual estado da arte, a inteligibilidade e naturalidade dos sistemas de síntese de texto para fala são limitados não por factores tecnológicos, mas sim pelas limitações do nosso conhecimento acerca de produção e percepção da fala. Quanto aos sistemas de reconhecimento, apesar da relativa maturidade, ainda temos desafios tecnológicos por resolver.

4 Sistemas conversacionais

Os sistemas conversacionais (do Inglês *Spoken Dialogue Systems*) são sistemas com os quais o utilizador interage por voz, num diálogo mais ou menos natural. Funcionam como uma interface de voz entre o utilizador e aplicações como bases de dados, sistemas periciais, entre outros. A sua complexidade varia muito em função da complexidade do diálogo (pergunta-resposta ou algo mais livre), da iniciativa do diálogo (de um dos interlocutores ou mista) ou do tamanho do vocabulário do domínio de aplicação, só para mencionar alguns factores.

Um sistema conversacional tem de lidar, no mínimo, com as seguintes tarefas:

- Desambiguar;
- Confirmar;
- Tratar erros;
- Preencher informação em falta;
- Gerir o contexto;
- Questionar bases de dados e/ou gerar respostas.

Existem diferenças claras entre sistemas conversacionais e os mais comuns sistemas de ditado e sistemas de comando-e-controlo por voz. Nos sistemas de ditado, o sistema transcreve pura e simplesmente aquilo que o utilizador diz, sem interpretar. Nos sistemas de comando-e-controlo por voz, o sistema reconhece o comando (seleccionado de uma lista restrita) e executa-o, não havendo diálogo para além disso. Num sistema conversacional há interpretação do discurso do utilizador e existe um diálogo que, como já referido, pode ser mais ou menos complexo. Os utilizadores de sistemas de voz muitas vezes não sabem que informação podem obter do sistema nem como a obter e, como tal, é importante que os sistemas consigam dialogar em vez de responder apenas a perguntas do utilizador. Os sistemas conversacionais podem, assim, ser vistos como aplicações de voz avançadas.

Embora não seja consensual, por uma questão de simplificação, consideramos que os sistemas conversacionais podem ser classificados em 3 grandes grupos consoante os modelos utilizados para controlar o diálogo com o utilizador:

- baseados em máquinas de estados finitos;
- baseados em *frames*;
- baseados em agentes.

Os dois primeiros são mais estáticos, ao passo que o último representa um modelo de gestão do diálogo mais dinâmico.

Nos sistemas baseados em máquinas de estados finitos a interface é um fluxo de diálogo entre o sistema e o utilizador, no qual as transições de estado para estado são desencadeadas através de frases do sistema (*prompts*) e respostas (*utterances*) do utilizador. O sistema mantém o controlo do diálogo, reconhece (ou rejeita) palavras e frases específicas e desencadeia as acções apropriadas. Este fluxo do diálogo, também designado por fluxo da chamada no caso das aplicações telefónicas, é o mais utilizado pois, entre outras razões, é conceptualmente simples, é atractivo em termos de processamento computacional e existe a possibilidade de desenhar as *prompts* cuidadosamente de forma a condicionar o utilizador, facilitando assim o reconhecimento. Contudo, é claro que quando se pretendem interfaces e

diálogos mais complexos, a estratégia de definir todos os estados e transições é bastante complexa dada a explosão de estados possíveis.

Nos sistemas baseados em *frames*, são colocadas questões ao utilizador de forma a ir preenchendo (*slot filling*) um *template* de forma a executar uma tarefa, tal como apresentar um horário, por exemplo. Neste tipo de sistema o fluxo do diálogo não está predeterminado, dependendo dos pedidos do utilizador e da informação que o sistema necessita. O utilizador pode fornecer logo toda a informação necessária e o sistema “age” como um sistema baseado numa máquina de estados finitos, ou, se não for esse o caso, o sistema vai pedindo ao utilizador a informação em falta. Com este tipo de sistema já é possível implementar diálogos mais complexos. Este foi o tipo de sistema usado no projecto MUST, do Eurescom (ver [Gomes 02]).

Os sistemas baseados em agentes são desenhados para permitir uma comunicação complexa entre o utilizador e o sistema de forma a resolver uma determinada tarefa. A naturalidade do diálogo aqui é muito superior ao possibilitado pelos tipos de sistemas referidos anteriormente, procurando-se uma certa cooperação entre o utilizador e o sistema. O modelo de diálogo mantém, tipicamente, um histórico do contexto anterior e, em função disso, evolui dinamicamente em sequências lógicas. A complexidade de implementação destes tipos de sistemas é também maior, obviamente.

4.1 Arquitecturas

Um sistema conversacional é composto por diversos componentes, como o reconhecedor de fala, o sintetizador de fala, o gestor do diálogo ou uma base de dados. A construção de um sistema deste tipo implica ainda um grande esforço não só na integração de todos estes componentes, mas também no desenho, configuração e teste do comportamento observável do mesmo. Isto é particularmente verdadeiro para sistemas de iniciativa mista. A figura seguinte ilustra de forma genérica os principais componentes num sistema conversacional e as suas inter-relações:

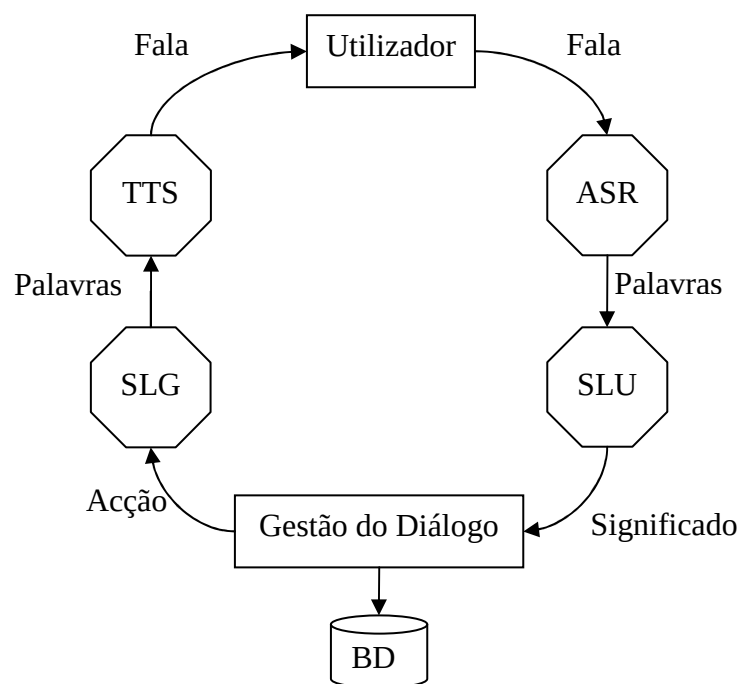


Figura 7 - Principais componentes num sistema conversacional

Quando o utilizador fala, o sinal acústico é convertido num conjunto de palavras pelo reconhecedor automático de fala (ASR - *Automatic Speech Recognizer*), que poderão posteriormente ser submetidas a um maior tratamento linguístico para extracção do significado relevante para a aplicação em causa (SLU - *Spoken Language Understanding*), através de gramáticas (por exemplo, *Context Free Grammars*) e/ou, mais recentemente, métodos estatísticos [Wang 05]. A “gestão do diálogo” (do Inglês *dialogue management*) designa as estratégias que determinam o fluxo do diálogo falado entre o utilizador e o sistema, especificando a acção a tomar em função do significado da elocução do utilizador. A acção, no caso dos sistemas conversacionais, é uma resposta falada e o processamento do pedido do utilizador (acesso a base de dados, por exemplo). A resposta é convertida em representação textual (SLG - *Spoken Language Generation*), sendo habitual considerar duas abordagens para tal: a) baseada em *templates*, através da concatenação, ou b) estocástica, baseada em conhecimentos linguísticos e métodos estatísticos. Finalmente, as palavras que constituem a representação textual são convertidas no sinal acústico de fala pelo sistema de conversão texto-fala (CFT) - ou sintetizador texto-fala (TTS - *Text-To-Speech*) -, completando assim o ciclo. É comum os componentes SLS e SLG estarem integrados no próprio gestor de diálogo (*Dialog Manager*), estando apenas representados no diagrama acima pela sua importância para melhor compreensão dos passos envolvidos e por serem eles próprios problemas difíceis de resolver, sendo activos campos de estudo.

Para o efeito de teste e avaliação, consideraremos, assim, como verdadeiramente importantes o módulo de reconhecimento de fala, o sintetizador texto-fala e o módulo de gestão do diálogo.

A arquitectura global mais comum de um sistema conversacional é a arquitectura Cliente-Servidor. Como componentes temos um cliente simples (interface de voz) e um servidor relativamente complexo. Ao nível do servidor podemos ter várias arquitecturas, mas as mais comuns são baseadas no *Galaxy Communicator* do projecto DARPA (<http://communicator.sourceforge.net/>). O *Galaxy Communicator* é uma arquitectura *open-source* para a construção de sistemas conversacionais que podem ser facilmente estendidos com a inclusão de outras modalidades, sendo, por isso, adequada também a sistemas multimodais. A sua concepção permite integrar de forma totalmente transparente componentes comerciais já existentes (como o ASR e o TTS) com Gestores de Diálogo específicos e/ou outros módulos que se pretendam desenvolver. A figura seguinte, retirada da página web do *Galaxy Communicator*, ilustra a arquitectura:

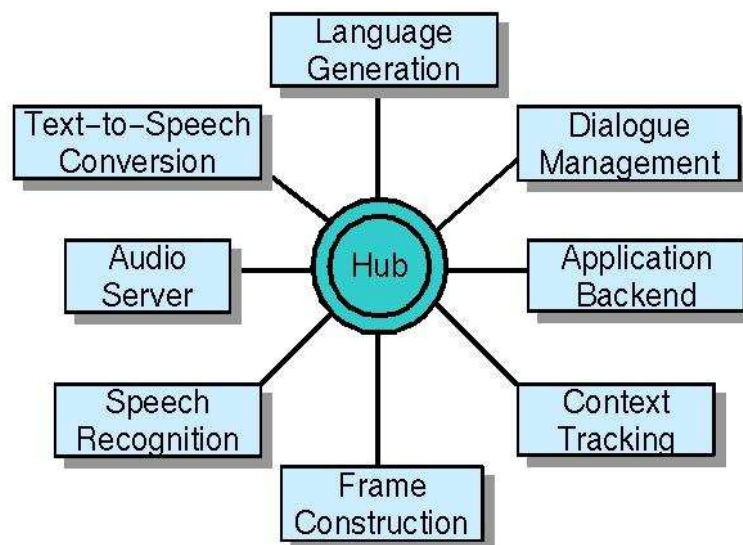


Figura 8 - Arquitectura de um sistema conversacional proposta pelo Galaxy Communicator

Observando o “interior” do Servidor, a sua arquitectura apresenta um estilo “Cubo e Raios” (*Hub-and-Spoke*). Temos um componente central (o cubo) que gere a transmissão das mensagens para os componentes internos do servidor (que se encontram nas extremidades dos “raios”) e implementam a lógica do serviço. Estes componentes internos funcionam eles próprios como clientes e servidores uns dos outros. A comunicação entre os componentes do servidor é assegurada pela plataforma de comunicação do *Galaxy communicator*. Cada componente envia as suas mensagens (pedidos ou respostas) para o *Hub* central sem se preocupar com quem é o destinatário ou a sua localização física, pois o *Hub* trata de as reencaminhar para o destino correcto. Tudo isto permite desenvolver componentes em diversas linguagens (Java, C e C++) e em sistemas operativos distintos (Windows e Linux). Estão assim garantidos os objectivos principais como a modularidade, distribuição, integração entre os módulos e a flexibilidade. É possível, por exemplo, adicionar ou substituir componentes (*off-the-shelf* ou desenvolvidos à medida) sempre que necessário sem ter de alterar nada.

Devido à existência de múltiplos modos de entrada e múltiplos modos de saída, a arquitectura de um sistema multimodal tende a ser mais complicada que a de um convencional sistema unimodal. Obriga também a algum tipo de gestão e interpretação dessa mesma entrada, até porque a informação de cada um dos canais pode ser complementar ou então concorrente. O mesmo se aplica à saída.

O desenvolvimento de aplicações multimodais coloca ainda mais ênfase na possibilidade de adicionar ou substituir componentes de forma flexível, sem ter de alterar toda a aplicação, de forma a, por um lado, acelerar o desenvolvimento e, por outro, permitir a sua evolução. Esta evolução poderá passar, por exemplo, pela adição de uma nova modalidade (voz, reconhecimento de gestos, ou qualquer outra) de entrada/saída. A importância deste aspecto evolutivo da arquitectura é enfatizado se tivermos em conta que estamos a falar de um campo de investigação actual, quer a nível de conceitos (gestão do diálogo, interfaces com o utilizador), quer ao nível tecnológico (reconhecimento e síntese de voz, plataformas de desenvolvimento).

Apesar da arquitectura Cliente-Servidor ainda ser a mais comum, mais recentemente as arquitecturas baseadas em agentes têm vindo a ganhar espaço nos sistemas conversacionais. Por exemplo a *Open Agent Architecture* (<http://www.ai.sri.com/~oaa/>) é uma arquitectura genérica baseada em agentes que tem sido utilizada em vários projectos de sistemas conversacionais ou multimodais (por exemplo [Moran 97]). Um dos mais recentes exemplos é a arquitectura *open-source Jaspis* [Turunen 04], que é simultaneamente uma *framework* que fornece componentes para o desenvolvimento de aplicações. Enquanto que na maior parte das arquitecturas de sistemas conversacionais baseadas em agentes, estes tratam de tarefas principais (como resolução de um determinado problema), na arquitectura *Jaspis* os agentes tratam de tarefas atómicas, como tratar de confirmações explícitas, ou apresentar a informação num determinado formato. Existem ainda agentes que podem desempenhar tarefas iguais, mas em circunstâncias diferentes e de forma diferente. A *Jaspis* é, assim, particularmente interessante do ponto de vista de aplicações distribuídas, flexíveis e adaptativas.

4.2 Reconhecimento automático de fala

O reconhecimento de fala refere-se à conversão de um sinal acústico de fala num conjunto de palavras, que poderão posteriormente ser submetidas a um maior tratamento linguístico. No entanto, e ao contrário da simplicidade da definição apresentada, o reconhecimento de fala é um problema extremamente difícil, principalmente devido às muitas fontes de variabilidade associadas ao sinal [Makhoul 95]. Para começar, e como já foi referido em secções anteriores, a acústica dos fonemas está dependente do seu contexto – variabilidade fonética (exemplo:

“casa do” e “casado”). Outro tipo de variabilidade é a chamada variabilidade acústica, que pode resultar de ruído ambiente ou de características técnicas/posição do microfone. Obviamente que as diferenças entre os indivíduos (sotaque, tom de voz, entre outros) são também elas fonte de variabilidade (inter-orador), e até um mesmo indivíduo pode induzir variabilidade (intra-orador) através de mudanças de estado físico ou emocional. Outras razões para esta complexidade podem ser compreendidas se tivermos em conta a espantosa quantidade de informação espectral do sinal de voz, o ruído ou até a carga semântica das palavras/expressões utilizadas.

Podemos considerar três tipos de tecnologias de reconhecimento de fala: comparação de *templates* (*template matching*), reconhecimento acustico-fonético e processamento estocástico [Markowitz 95]. A primeira é uma forma de reconhecimento de padrões em que cada palavra é armazenada como um *template* diferente. A fala do utilizador é organizada em *templates* que são depois comparados com os armazenados. Como a mesma palavra pode ser pronunciada a diferentes velocidades, entoações ou condições ambiente têm de ser realizadas correcções – aplicando-se técnicas como o *dynamic time warping* (DTW). Este tipo de reconhecimento funciona bastante bem para aplicações de vocabulário reduzido e nos quais as palavras são foneticamente claramente distintas, razão pela qual, embora “ultrapassada”, ainda hoje é utilizada em aplicações como a identificação do orador (*speaker verification*).

A segunda tecnologia de reconhecimento, acustico-fonético, opera ao nível dos fonemas, e não das palavras como na comparação de *templates*. Assim sendo, em vez de termos de armazenar todas as palavras que for necessário reconhecer, basta, teoricamente, armazenar todos os fonemas de uma determinada língua (recordar da secção 2.4.1 que a fonologia do português, por exemplo, é composta por cerca de 33 a 38 fonemas). No entanto, as semelhanças entre os diversos fonemas conjuntamente com a variabilidade desses mesmos fonemas (devido à coarticulação, entre outras razões) obrigam a utilização de métodos estatísticos para calcular as inúmeras hipóteses que daí resultam. Infelizmente, na altura em que esta tecnologia de reconhecimento apareceu, o equipamento necessário para implementá-la de forma satisfatória em termos de velocidade e precisão era bastante caro, pelo que se foi perdendo o interesse no seu desenvolvimento.

Já o reconhecimento estocástico representa o actual estado da arte, mais concretamente as abordagens híbridas - que combinam os modelos de Markov ocultos (HMM) com as redes neuronais. Tal como na comparação de *templates*, na abordagem estocástica também é necessário armazenar modelos do que é possível reconhecer, mas neste caso não há comparação directa entre a fala e o que está armazenado, sendo, em vez disso, feitas complexas análises estatísticas recorrendo aos HMMs. O reconhecimento de fala é, essencialmente, uma tarefa de classificação (de palavras, sílabas ou fonemas) e sendo as redes neuronais excelentes sistemas de classificação, possibilitando ainda um processamento paralelo, a sua utilização neste contexto torna-se particularmente atractiva. A implementação das redes neuronais não obriga à especificação completa do problema a resolver, em vez disso, as redes neuronais “aprendem” padrões através da exposição a grandes quantidades de exemplos desses padrões. Este processo é designado por treino. Podem ser treinadas para reconhecer palavras, sendo-lhes fornecida a identidade das palavras que correspondem às entradas a reconhecer: depois de processar as entradas, o algoritmo de treino compara as decisões tomadas com as respostas correctas e “corrige” eventuais erros através da alteração das ligações entre os nodos que formam a rede.

A figura seguinte ilustra, de uma forma muito simplista, os principais componentes de um moderno sistema de reconhecimento de fala:

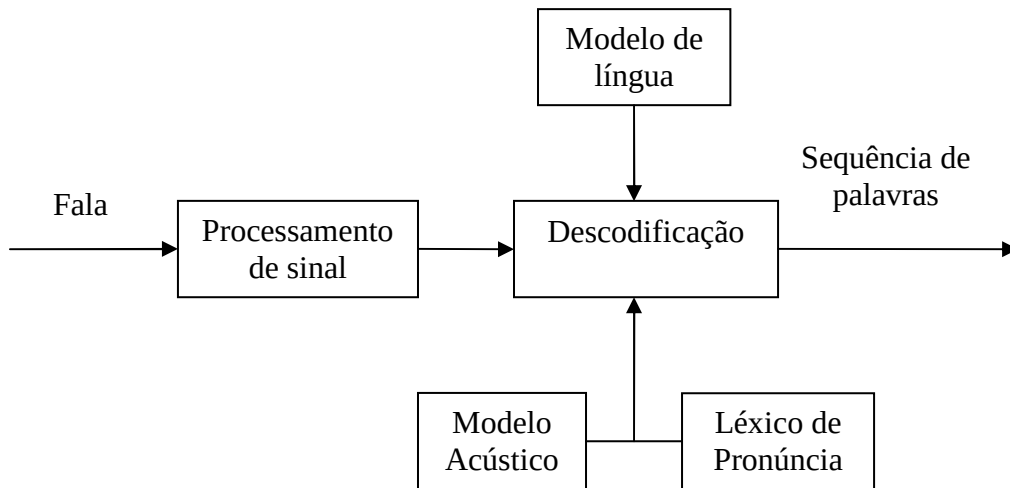


Figura 9 - Principais componentes de um sistema de reconhecimento de fala

Como se pode observar, o reconhecimento de fala tem duas fases principais: o processamento de sinal e a descodificação.

No processamento de sinal, o som é digitalizado e depois transformado a uma determinada cadência (através de análise espectral, usando, por exemplo, *Linear Predictive Coding*) num conjunto de medições ou características – um processo conhecido por *feature extraction*. Estas medições são depois usadas no bloco de descodificação para pesquisar a palavra que com maior probabilidade foi dita, fazendo uso de regras estabelecidas pelos modelos de língua, modelo acústico e pelo léxico.

Para efectuar a descodificação são então necessários dois componentes (fontes de conhecimento) principais: um com a probabilidade da sequência de parâmetros extraída para cada uma das palavras possíveis e outro que indique qual a probabilidade de uma dada palavra se seguir às anteriores já reconhecidas. No cálculo da primeira probabilidade referida estão tipicamente envolvidos modelos para todos os fones possíveis da língua (modelo acústico) e um léxico de pronúncia que associa a cada palavra o seu desdobramento em sequência de fones. No cálculo da segunda (modelo de língua) usam-se regras gramaticais ou modelos estatísticos - como N-gramas (*N-gram*) indicando a probabilidade de uma palavra dadas as N anteriores (normalmente 2 ou 3) -, que são os modelos de língua.

Resumindo, a descodificação combina a informação (probabilidades) do modelo acústico, do modelo de língua e do léxico para, percorrendo o espaço de soluções, determinar a sequência de palavras mais provável usando, por exemplo, o algoritmo *Viterbi*. Finalmente, ao resultado da pesquisa pode-se aplicar uma estratégia para melhorar a robustez do reconhecimento, como por exemplo o uso de índices de confiança (*confidence scores*). Os sistemas de reconhecimento usam estes valores para decidir se devem “acreditar” na frase reconhecida ou não. No entanto, a experiência mostra que nem sempre existe uma relação directa entre um índice de confiança alto e um reconhecimento bem sucedido e vice-versa (para mais detalhes, ver, por exemplo, [Skantze 03]).

A sua estrutura e o algoritmo de treino utilizado para determinar os parâmetros de reconhecimento tornam os HMMs apropriados para aplicações de grande vocabulário (*large-vocabulary*), independente do orador (*speaker-independent*) e fala contínua (*continuous speech*), ou seja, o estado-da-arte ao nível das aplicações de fala.

Os modelos ocultos de Markov fazem essencialmente o mapeamento entre o sinal de fala amostrado e uma sequência de unidades fonéticas – modelo acústico. Trata-se de um sistema hierárquico, sendo os HMMs também usados para mapear a sequência fonética numa sequência de palavras.

No plano conceptual, um HMM é um modelo estatístico do processo de produção da fala: dada uma sequência de palavras, pode ser usado para gerar o sinal (ver secção de síntese de fala do presente documento). Pode também ser usado para calcular a probabilidade de um determinado sinal ter sido gerado a partir de uma determinada frase.

No reconhecimento de fala, um HMM é construído para cada frase permitida pela gramática, daí também a complexidade colocada por sistemas de grande vocabulário, por exemplo. Quando é necessário reconhecer a fala, o sistema calcula a probabilidade dela ter sido gerada por cada HMM. O resultado do reconhecimento é a frase que corresponde ao HMM que com maior probabilidade gera essa fala. O desempenho dos HMMs depende em grande medida da qualidade dos dados utilizados para definir as diferentes probabilidades internas. Estes dados podem ser obtidos a partir de utilizadores individuais, grupos de oradores ou bases de dados preexistentes (corpora linguístico). É, assim, evidente a importância de dados de treino ou corpora linguístico (ver o exemplo em [Gomes 05]).

Devido à natureza temporal do sinal de fala, é importante ter um mecanismo que lide com essa dimensão. Num HMM este mecanismo é constituído pelas cadeias de Markov. Usando a cadeia de Markov, uma frase pode ser decomposta numa sequência de estados fonéticos. Se esta sequência for conhecida, é quase trivial efectuar o reconhecimento. No entanto, quando é feito o reconhecimento, apenas a forma de onda do sinal de fala é “visível”: a sequência fonética está “oculta”. É esta sequência que se procura reconhecer só que, pela própria natureza duplamente estocástica dos HMMs, é impossível determinar a sequência de estados que gerou um determinada fala (saída de símbolos), pelo que se procura a que tem maior probabilidade [Makhoul 95]. Teoricamente seria necessário examinar todos os estados possíveis e calcular as suas probabilidades – por exemplo, uma sequência de 8 dígitos tem 100 milhões de frases possíveis... Na prática, partindo do pressuposto *Markoviano* de que cada estado depende apenas do anterior, algoritmos eficientes, como o *Viterbi*, são usados para reduzir e percorrer o espaço de soluções. Para além disso o sistema de reconhecimento faz uso da estrutura hierárquica dos HMMs: HMMs das frases são a concatenação de HMMs de palavras e estes, por sua vez, são a concatenação de HMMs de unidades fonéticas... Durante o reconhecimento, o esforço computacional pode ser partilhado entre as várias hipóteses – por exemplo, e de uma forma muito simplista, a mesma computação é necessária para as primeiras duas palavras de “hoje está sol” e “hoje está chover”. Na maior parte dos casos, esta partilha não é suficiente para se conseguir um reconhecimento em tempo real e outras heurísticas têm de ser usadas como o corte (poda) de hipóteses pouco prováveis (designado por *prunning*). Uma das grandes dificuldades é exactamente balancear a velocidade com a precisão do reconhecimento. Têm sido nos últimos anos feitos avanços em técnicas como os Transdutores de Estado Finito Ponderados (do Inglês WSFT - *Weighted Finite State Transducers*) para representação do espaço de pesquisa e optimização do processo de pesquisa, convertendo os HMMs para este formalismo [Trancoso 04].

A saída do reconhecedor pode ser uma lista das N melhores hipóteses (*N-Best list*), ficando depois a cargo de outros módulos seleccionar a melhor em função do contexto, por exemplo, ou podem ser palavras-chave – neste caso o reconhecedor procura por palavras específicas em vez da frase toda.

Os sistemas de reconhecimento podem ser caracterizados através de diversos aspectos, de entre os quais se podem destacar:

- tipo de discurso – podem reconhecer apenas palavras isoladas ou discurso contínuo. O reconhecimento de palavras isoladas difere do de fala contínua, pois neste último o locutor não precisa fazer pausas entre as palavras. Deste modo, a determinação das fronteiras entre as palavras e do número destas no discurso deve ser feita pelo sistema de reconhecimento;

- estilo do discurso – podem suportar discurso espontâneo ou não. A fala quando gerada de modo espontâneo é mais relaxada, contém mais coarticulações, hesitações ou interjeições, e portanto é muito mais difícil de reconhecer do que quando gerada através de leitura (ver por exemplo [Furui 03]). Outra dificuldade adicional advém do facto de que quando os utilizadores ficam familiarizados com o sistema, querem acelerar o diálogo, interrompendo o sistema (*barge-in*). O *barge-in* dificulta o reconhecimento pois, por um lado o reconhecedor tem de estar sempre “alerta” (captando qualquer ruído) e, por outro, o sinal de entrada (do utilizador) pode vir “misturado” com a *prompt* de saída (do sistema);
- treino – pode ser necessário treinar o sistema com os “utilizadores tipo” ou pode ser independente do utilizador (*speaker independent*);
- tamanho do vocabulário suportado – é mais difícil o reconhecimento para vocabulários maiores e/ou que tenham muitas palavras parecidas;
- tipo do modelo de língua – desde modelos de estados finitos até gramáticas sensíveis ao contexto;
- perplexidade – número médio de palavras que pode suceder a uma determinada palavra reconhecida. Por exemplo, o reconhecimento de dígitos apresenta uma baixa perplexidade;
- robustez ao ruído ambiente;
- tipo de microfone.

Hoje em dia já estão disponíveis, pelo menos na comunidade científica, sistemas de reconhecimento de vocabulário muito grande (*very-large-vocabulary*), com mais de 20.000 palavras, grande perplexidade, independentes do orador (*speaker-independent*) e de reconhecimento contínuo. Estamos no limiar do aparecimento de serviços e aplicações com um nível de reconhecimento que permite a sua comercialização: podemos já encontrar serviços telefónicos (como o 118 ou Portal de Voz) em muitos países, ou aplicações de produtividade (*MS Office*). Em laboratório, os protótipos são mais variados e avançados, como a aplicação de informação turística no projecto MUST [Gomes 02] ou assistentes pessoais virtuais no projecto FASiL [Gomes 05]. No entanto, e apesar dos progressos, ainda estamos longe de sistemas conversacionais bons, e muito longe do nível do HAL (ver capítulo 3). Fora dos ambientes laboratoriais, os erros ocorrem com uma frequência, por vezes, não negligenciável... Os aspectos que precisam de melhorar e são alvo de estudo incluem a robustez (e as estratégias associadas) – ver por exemplo [Sa 03], a portabilidade, o suporte para o discurso espontâneo (com pausas, hesitações, interjeições), a prosódia ou as palavras fora do vocabulário – só para mencionar os mais prementes.

A qualidade do ASR pode ser avaliada, tradicionalmente, em termos de precisão – taxa de erro (*Word Error Rate* – WER) tipicamente –, robustez e velocidade. No capítulo 5 do presente documento este tema será abordado em maior profundidade.

4.3 Síntese texto-fala

A saída de fala de um sistema conversacional pode ser constituída apenas por fala sintetizada, apenas por fala pré-gravada ou por uma solução híbrida em que temos segmentos pré-gravados intercalados com segmentos de fala sintetizada, como por exemplo “O número de telefone solicitado é <número_sintetizado>”. Quando temos grande quantidade de informação imprevisível ou variável, ou a informação da saída é seleccionada a partir de um conjunto

dados, a única solução é a sintetização. As etapas de geração de fala para uma determinada aplicação, podem ser representadas, de forma genérica, como no diagrama seguinte:

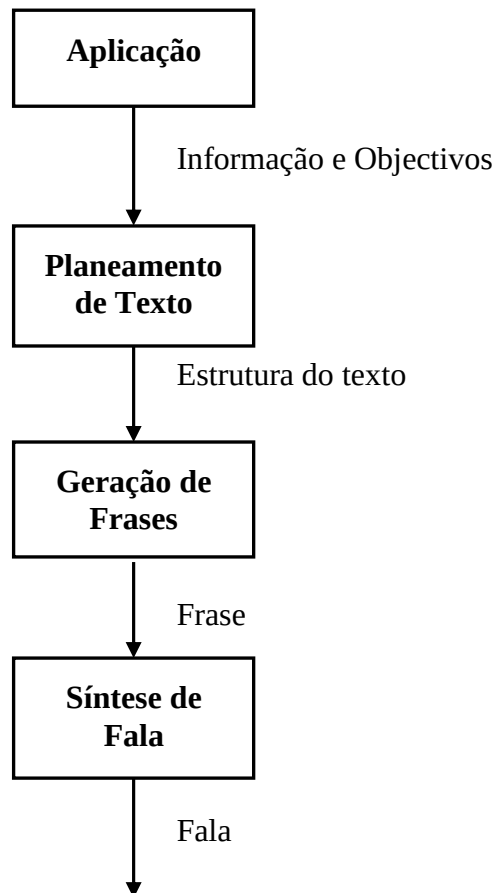


Figura 10 - Etapas da geração de saída de fala

Nesta secção será abordado mais pormenorizadamente a última etapa de geração de saída de fala, a síntese de fala (ver figura anterior).

Nos primórdios da síntese de fala, a investigação estava mais focada na reprodução dos mecanismos de produção de fala humanos (ver secção 2.4). Embora este modelo ainda seja usado em situações específicas, os avanços registados no poder computacional alargaram os campos de investigação, passando a incluir, por exemplo, conversão de texto para fala (TTS – *Text-To-Speech*). Esta modelação é feita, normalmente, por um conjunto de regras (*rule-based approach*) derivadas de áreas como a fonética, a acústica, entre outras, pelo que é designado como síntese de fala por regras. Em contraste com a abordagem da modelação por regras temos a abordagem baseada em corpus linguístico (*corpus-based approach*). Nesta abordagem dados de fala perfeitamente definidos, como por exemplo, gravações de diálogos ou apenas elocuções independentes num determinado domínio (ver o exemplo em [Gomes 05]), são anotados a vários níveis (sintáctico, semântico, fonético), sendo usados depois como base estatística para a modelação. As características espectrais e prosódicas da fala são analisadas em função das anotações e, com base nos resultados, é gerado um modelo computacional que é posteriormente treinado usando o corpus.

O funcionamento de um sistema TTS pode ser dividido em duas fases principais. A primeira fase consiste na análise do texto, na qual o texto de entrada (ver última etapa na Figura 10) é transcrito para uma representação fonética, e a segunda fase é a geração da forma de onda, onde a saída acústica é produzida a partir da representação fonética obtida na fase anterior. A figura seguinte ilustra este processo:

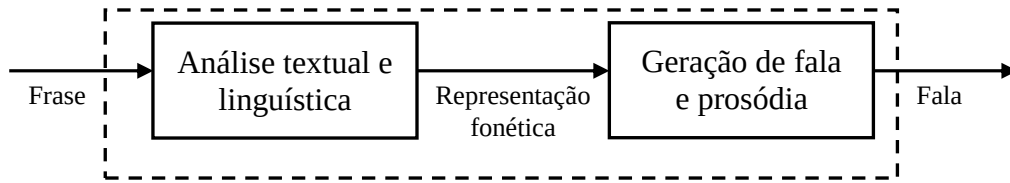


Figura 11 - Etapas da síntese de fala

Na figura acima, podemos observar que a entrada para a fase de geração da fala e prosódia é, basicamente, uma cadeia de caracteres fonéticos resultante da conversão dos grafemas para fonemas realizada na fase de análise. Normalmente essa cadeia é complementada por uma série de marcas prosódicas representando, por exemplo, acentos e pausas. A cadeia em si pode ser constituída pelos fonemas ou pelos fones. Como já foi referido na secção 2.4.1 do presente documento, durante a fala, os movimentos articulatórios de um determinado fonema dependem dos fonemas precedentes e dos seguintes (coarticulação), pelo que podem ocorrer ligeiras variações na pronunciação desse fonema. Assim sendo, no caso das cadeias de fonemas, pode ser necessário aplicar algumas regras de transformação para se obter a transcrição para fones. Sem elas, a saída do sintetizador poderia ser inteligível, mas não possuiria uma boa naturalidade. Quanto às marcas prosódicas, estas são transformadas num conjunto de parâmetros para controlar o sintetizador, como por exemplo, a frequência fundamental (F0), a duração de cada segmento de fala, a intensidade e o timbre. Para que estes parâmetros controlem o sintetizador é necessário implementar modelos de entoação e duração.

A derradeira etapa para a saída de fala é a sintetização da forma de onda de acordo com os passos anteriores. A geração de sinal de fala (sintetização) pode ser conseguida através de diversos métodos que podem ser classificados em 3 tipos principais (embora existam outros):

- sintetizadores articulatórios;
- sintetizadores de formantes;
- sintetizadores concatenativos (por concatenação).

Os primeiros são modelos físicos baseados em descrições detalhadas da anatomia e acústica do aparelho fonador humano (ver secção 2.4). Os parâmetros típicos são a posição e a cinemática dos articuladores. A partir destes, o som que seria emitido na boca é calculado. Teoricamente, esta seria a forma mais satisfatória de gerar fala com alta qualidade. No entanto devido à sua complexidade e existência de problemas teóricos e práticos por resolver, o seu potencial ainda não foi atingido, o que, por sua vez, condiciona a sua difusão comercial.

Os sintetizadores de formantes representam uma abordagem acustico-fonética baseada no modelo fonte/filtro, controlado por um conjunto de regras fonéticas. A geração de fala não é realizada através de resolução de equações, como na categoria anterior, mas sim através de modelação das principais características acústicas do sinal de fala. Tipicamente, são necessárias as 3 primeiras formantes para produzir fala inteligível, ou até 5 para a produção de fala com alta qualidade. Este tipo de sintetizadores já está disponível comercialmente há anos e com uma boa inteligibilidade, embora a naturalidade por vezes não seja a melhor (um som mais robótico que Humano é característico deste tipo de sintetização). Uma das principais vantagens deste tipo de sintetizadores é o facto de ocupar pouco espaço, pelo que é frequentemente usado em sistemas embebidos.

Na síntese por concatenação, segmentos fonéticos pré-gravados são concatenados e é efectuado algum processamento de sinal, como o uso do LPC (*Linear Predictive Code*), para “suavizar” a transição entre as unidades fonéticas usadas – difones ou unidades de comprimento variável como sílabas e fones - e para obter a prosódia pretendida. Tal como o

anterior, este tipo de sintetizadores já está disponível comercialmente há anos e com uma boa qualidade, quer ao nível da naturalidade, quer da inteligibilidade. Como desvantagens principais temos o facto de estar geralmente limitado a uma voz, ocupar muito espaço, requerendo a existência de corpora de grandes dimensões (bases de dados de amostras de voz) e, finalmente, a dificuldade associada aos algoritmos de selecção dos segmentos/unidades a concatenar.

As mais recentes tecnologias de síntese de fala envolvem métodos e algoritmos complexos, como os modelos de Markov ocultos (HMM), desde a década de 80, para a síntese de fala, ou as redes neuronais, desde a década de 90, que modelam, por exemplo, o tracto vocal ou a prosódia.

Apesar de um certa maturidade, a síntese de fala ainda tem um caminho a percorrer no sentido de obter uma maior inteligibilidade e, principalmente, naturalidade. Existem vários problemas por resolver no pré-processamento do texto, tais como as abreviações, acrónimos ou numerais. A obtenção de uma análise de prosódia e pronúncia correctas a partir de texto ainda é difícil, pois o texto escrito não contém emoções de forma explícita ou a indicação da finalidade do discurso, estado mental do orador, entre outros factores. No campo da geração da forma de onda, as vozes femininas ou infantis são difíceis de sintetizar dado que possuem frequências fundamentais mais elevadas, dificultando, assim, a determinação das formantes. Por outro lado, como é óbvio, a geração de certas línguas é muito mais difícil que outras e os recursos disponíveis/mercado potencial também não é o mesmo para todas as línguas.

Ao nível da avaliação de qualidade, a tarefa também não é fácil dada a multidimensionalidade do próprio termo. Os critérios de qualidade normalmente considerados para a avaliação do sintetizador de fala são a inteligibilidade e a naturalidade. Por inteligibilidade entende-se a sintetização correcta dos fonemas e pode ser medida em função de diversas unidades (fonemas, sílabas, palavras, frases). Já a naturalidade é mais difícil de definir, mas podemos dizer que está ligada aos elementos prosódicos (como a velocidade, o timbre ou o nível) e sintetização correcta dos alofones (variantes fonéticas dos fones, ou seja, diferentes realizações de um mesmo fonema). A maior parte dos sistemas de síntese de fala consegue ter uma boa inteligibilidade, mas a vertente da naturalidade tende a apresentar menor qualidade. Por vezes, também se inclui nos parâmetros de teste a capacidade do TTS em processar abreviaturas ou sequências numéricas, por exemplo, vulgarmente designado por *Front-End Processing*. À semelhança do teste e avaliação do ASR na secção anterior, também aqui será remetido para o capítulo 5 do presente documento o teste e avaliação do TTS.

4.4 Gestão do diálogo

As decisões acerca da estratégia de diálogo têm um grande impacto na forma como o utilizador interage com o sistema, no *look-and-feel* do sistema e, em última análise, na aceitação desse mesmo sistema.

O termo “gestão do diálogo” (do Inglês *dialogue management*) designa as estratégias que determinam o fluxo da interacção, ou seja, do diálogo falado entre o utilizador e o sistema. Isto envolve, tipicamente, determinar se já foi solicitada ao utilizador a informação necessária, comunicar com aplicações externas (base de dados ou sistemas periciais, por exemplo) e comunicar informação ao utilizador.

Este fluxo pode-se tornar extremamente complexo à medida que a quantidade de itens de informação a ser discutida aumenta. Estes itens de informação podem ser definidos, genericamente, como a informação que deve ser solicitada ao utilizador, ou que este deve fornecer, para concluir uma determinada tarefa. Por exemplo, no caso de um formulário de informação pessoal, os itens de informação podem ser o “nome”, a “morada”, o “telefone”,

entre muitos outros. Ou, no caso, de um serviço de viagens, podem ser a “origem”, o “destino”, a “data”, por exemplo.

Nos diálogos entre seres Humanos podemos encontrar muitas intervenções (*turns*) que não são dedicadas à transmissão de itens de informação mas sim à manutenção do diálogo. Muitas intervenções podem também ser dedicadas à confirmação da compreensão dos itens de informação. Quando se trata de diálogos entre um ser Humano e um sistema conversacional, pode ser necessário ainda um outro tipo de intervenções (*turns*), raro nos diálogos entre seres humanos, nas quais o sistema tenta expor as suas “capacidades” e o domínio de comandos disponíveis. Por exemplo, a determinado ponto do diálogo, o sistema pode ter necessidade de “dizer” algo acerca da tarefa que se está a tentar executar: vamos imaginar que estamos perante o serviço 118 e o utilizador está a tentar obter um número de telemóvel, nesse caso o sistema poderia dizer “posso apenas fornecer números de telefone fixos. Diga se pretende clientes particulares ou residenciais”, ou algo do género. A iniciativa também poderia ter partido do utilizador, caso este tivesse dito, por exemplo, “que tipo de números telefónicos posso obter?”.

A gestão do diálogo é, assim, a estratégia que determina o tipo da próxima intervenção no fluxo do diálogo. Em termos de controlo do diálogo, podemos ter três tipos: iniciativa por parte do utilizador, iniciativa por parte do sistema e iniciativa mista.

Na iniciativa por parte do utilizador, o tipo de intervenção e o item de informação a tratar são sempre determinados pelo utilizador. No entanto, tal não significa que o utilizador tem liberdade total: por exemplo, certos “campos” a preencher podem ter um tipo predeterminado ou o utilizador pode apenas emitir um determinado conjunto de comandos.

Na iniciativa por parte do sistema, é este que determina o tipo de intervenção e o item de informação a ser tratado na próxima intervenção. Um exemplo deste tipo de sistema são os IVRs clássicos em que o utilizador é questionado pelo sistema para escolher uma opção a cada intervenção. Em cada ponto do diálogo existe um conjunto fixo de respostas válidas que o utilizador deve seleccionar. Quase sempre existe uma opção de recurso (transferir a chamada para um operador Humano, por exemplo) à qual o utilizador pode recorrer caso o diálogo não esteja a correr bem, mas não deixa de pertencer ao conjunto de respostas válidas predeterminadas pelo sistema.

Na iniciativa mista, tal como o nome indica, tanto o utilizador como o sistema podem propor o tipo da próxima intervenção no diálogo, bem como o item de informação a ser tratado. Este tipo de iniciativa é muito mais natural pois é o mais semelhante ao encontrado nos diálogos entre Humanos. No entanto, dada a dificuldade de implementação, a maior parte dos sistemas usa um estilo de iniciativa por parte do sistema.

Os aspectos básicos da gestão de diálogo numa interacção multimodal não apresentam diferenças significativas em relação aos de uma interacção unimodal (de fala, neste caso). Obviamente, que num plano de mais alto nível, a estratégia tem de ter em conta os outros modos de interacção disponíveis.

Em termos de avaliação de desempenho, tipicamente, medem-se o tempo e a taxa de sucesso na execução de uma determinada tarefa bem definida e, de forma indirecta o tempo de aprendizagem, como por exemplo, “obter o número de telefone fixo de uma pessoa, tendo o nome e a morada”.

Um tipo de teste que se pode aplicar ao módulo de gestão do diálogo é “desligá-lo” dos módulos de ASR e TTS e interagir com ele através de comandos de texto, sendo a saída também através de texto. Este esquema permite a automatização do teste abrindo assim a porta a um maior rigor no mesmo, a sua repetição com maior facilidade, uma melhor análise de resultados, entre outras. No capítulo 5 do presente documento este tema será aprofundado.

4.5 Aspectos de usabilidade

A fala é das mais importantes formas de interacção humana, desenvolvimento, aquisição e transmissão de conhecimento. No entanto existem menos serviços de voz do que seria de esperar a partir de tal afirmação. A razão poderá estar não só nas dificuldades técnicas e problemas de reconhecimento, mas também no facto de “o mais natural” não significar necessariamente “o mais eficiente”... afinal de contas ouvimos recorrentemente que “uma imagem vale mil palavras”. A fala pode ser fácil de produzir e entender por nós Humanos mas, pela sua própria natureza, não apresenta persistência. Do domínio da psicologia é sabido que somos melhores a reconhecer que a recordar. Assim, por exemplo, é mais “fácil” percorrer menus com muitos itens apresentados graficamente do que a situação em que as inúmeras opções são apresentadas por voz e termos de as “recordar”. E graficamente o acesso aos itens pode ser feito aleatoriamente, enquanto nos menus por voz é sequencial. As representações gráficas e escritas apresentam ainda uma outra vantagem: podem ser mais demoradas a produzir, mas o seu processamento por meios digitais é muito mais fácil e rápido. Parece, assim, evidente a importância de avaliar o real benefício da utilização da voz numa interface.

Aspectos como a arquitectura das interfaces ou os modelos de fluxo do diálogo são essenciais ao desenvolvimento de uma boa interface de voz, mas não basta ser boa... os utilizadores têm de a aceitar. É fundamental desenhar a interface de voz com base nas melhores práticas de usabilidade no domínio da voz e, para tal, é necessário conhecer essas práticas e regras ([Anderson 01] e [Weinschenk 00]). É necessário compreender correctamente o problema, os utilizadores, o seu modelo mental e as suas necessidades, bem como definir claramente objectivos de usabilidade e de experiência dos utilizadores. A lista seguinte apresenta exemplos das melhores práticas de usabilidade neste domínio:

- procurar obter informação específica de forma rápida, quer se tratem de utilizadores experientes ou inexperientes;
- fornecer ajuda aos utilizadores quando for pedido;
- “conduzir” o utilizador de forma a este usar palavras suportadas pelas gramáticas, de forma a tentar minimizar o número de erros ocorridos. Numa interface gráfica, os utilizadores clicam em botões com nomes específicos (“abrir”, “fechar”, etc.), ao passo que numa interface de voz podem dizer, virtualmente, qualquer coisa, pelo que é importante “levar” os utilizadores a dizerem aquilo que as gramáticas contêm;
- pedir confirmação dos utilizadores quando uma transacção estiver prestes a ser executada, pois um erro de reconhecimento pode conduzir a transacções não solicitadas;
- tentar assegurar que o utilizador consiga executar as tarefas de forma rápida;
- evitar que o utilizador se possa sentir ofendido de alguma forma;
- permitir que o utilizador possa interromper o sistema em qualquer altura (*barge-in*), de forma a aumentar a sensação de controlo;
- utilizar mensagens de erro específicas. Dados os erros de reconhecimento, é importante que o sistema indique ao utilizador o que dizer em cada contexto;
- utilizar ícones de áudio (*earcons*), pois fornecem pistas ao utilizador de que o sistema “ouviu” o que o utilizador disse e está a efectuar algum tipo de processamento. Sem estas pistas, o utilizador poder-se-á sentir impaciente ou frustrado e/ou repetir comandos sem necessidade pois não sabe o que o sistema está a fazer;

- indicar claramente as mudanças de estado do diálogo, para auxiliar a navegação do utilizador pelos estados da interface de voz;
- identificar o sistema na mensagem de boas vindas, de forma a que os utilizadores se apercebam logo desde o início que estão a falar com um computador e não um ser Humano. Isto tem por objectivo diminuir o nível de expectativas do utilizador e, espera-se, diminuir a taxa de erro de reconhecimento;
- evitar o uso da primeira pessoa no discurso do sistema, pois poderá levar a que os utilizadores usem frases mais “verbosas”, dificultando a tarefa de reconhecimento dos comandos principais;
- evitar o uso de frases ou menus longos, pois não existe *feedback* visual numa interface de voz pelo que a memória de curta duração dos utilizadores rapidamente atinge o limite neste tipo de interfaces. É prática corrente usar apenas três ou quatro opções nos menus de voz. Além disso, é mais comum os utilizadores lembrarem-se das opções nos extremos e esquecerem as que estão no meio da lista, pelo que se devem colocar as mais importantes nos extremos;
- evitar termos técnicos de forma a assegurar que o maior número possível de utilizadores possa compreender as *prompts* do sistema;
- usar temporizações (*time-out*) quando se espera *input* do utilizador. Assim, se ao fim de algum tempo o utilizador ainda não respondeu, o sistema deve voltar a colocar a questão;
- evitar pausas muito longas, pois pode deixar o utilizador frustrado por não saber se o sistema está a processar alguma coisa, ou não está activo, ou qualquer outra situação;
- quando for necessário repetir várias vezes a mesma pergunta ao utilizador, ir reduzindo gradualmente a verbosidade da pergunta até só restarem os comandos principais. Esta técnica é conhecida como *prompt tapering*.
- fazer uso do DTMF (*Dual Tone Multi-Frequency*) se possível, pois em ambientes mais ruidosos e/ou quando ocorrem muitos problemas de reconhecimento, os utilizadores podem sempre recorrer ao teclado do telefone para navegar nos menus;
- em menus com mais de duas opções, descrever o resultado primeiro e de seguida o comando esperado;
- indicar claramente o fim do menu ao utilizador. Por exemplo “Para criar uma novo pedido carregue 1, para listar as pizzas carregue 2 (...). Por favor, indique agora a sua opção.”;
- nos menus, usar os resultados como *feedback* para o utilizador. Por exemplo “Para criar uma novo pedido carregue 1, para listar as pizzas carregue 2 (...). Por favor, indique agora a sua opção.”, supondo que o utilizador carrega no “2”, então o feedback seria “Escolheu 2. A lista de pizzas (...).”;
- se possível, utilizar menus de iniciativa mista, nos quais o utilizador pode fornecer a informação na ordem que desejar. Esta técnica tem por objectivo principal transmitir uma maior sensação de controlo ao utilizador;
- usar técnicas de aprendizagem, se assim se pode chamar, para adaptar o sistema ao utilizador. Por exemplo, pode-se contabilizar o número de erros de reconhecimento ou

pedidos de ajuda para um determinado utilizador e assim “regular” a verbosidade das *prompts* do sistema.

A concepção de interfaces passa por um processo evolutivo, onde os utilizadores devem ser envolvidos em todo o ciclo de desenvolvimento. A evolução da interface deve ser feita de acordo com os contributos dos utilizadores. O principal objectivo desse trabalho é sempre dar resposta às necessidades dos utilizadores com o menor esforço possível destes na aprendizagem e na utilização.

O trabalho de avaliação de uma interface requer a utilização de técnicas de monitorização da utilização do sistema por utilizadores reais, observação e definição de métricas para avaliação quantitativa e qualitativa. Importa conhecer quais os paradigmas, técnicas e métricas de avaliação mais adequadas às interfaces de voz. Apesar dos recentes avanços nas tecnologias empregues nos sistemas conversacionais, os aspectos de usabilidade relacionados são ainda um campo de estudo importante. À medida que produtos nesta área vão chegando ao mercado, o reconhecimento das lacunas ainda existentes e importância da compreensão dos factores de usabilidade nestes sistemas tem ganho relevância. Convém no entanto ter presente que a usabilidade não é, nem nunca será, condição suficiente para o sucesso de um produto ou serviço.

Numa perspectiva global, os estudos de interacção Homem-Máquina têm evoluído no sentido de se concentrarem mais num conjunto abrangente de questões relacionadas com o divertimento, estética e experiência do utilizador, em detrimento de preocupações “clássicas” de usabilidade, como a eficiência e a eficácia.

Este tema será abordado em maior detalhe no capítulo 5 do presente documento.

5 Teste e avaliação de interfaces de voz: estado da arte

Os principais tipos de testes de sistema neste tipo de aplicações são:

- Testes funcionais (*functional testing*);
- Testes de carga e desempenho do sistema (*load and performance testing*);
- Testes de desempenho do reconhecedor de fala e a qualidade da saída;
- Testes de usabilidade.

A figura seguinte representa uma arquitectura simplificada de um sistema conversacional e ajuda-nos a situar os tipos de testes mencionados e os componentes aos quais estes se aplicam. As secções do documento que se seguem detalham cada um destes tipos de testes e permitem uma maior compreensão da figura.

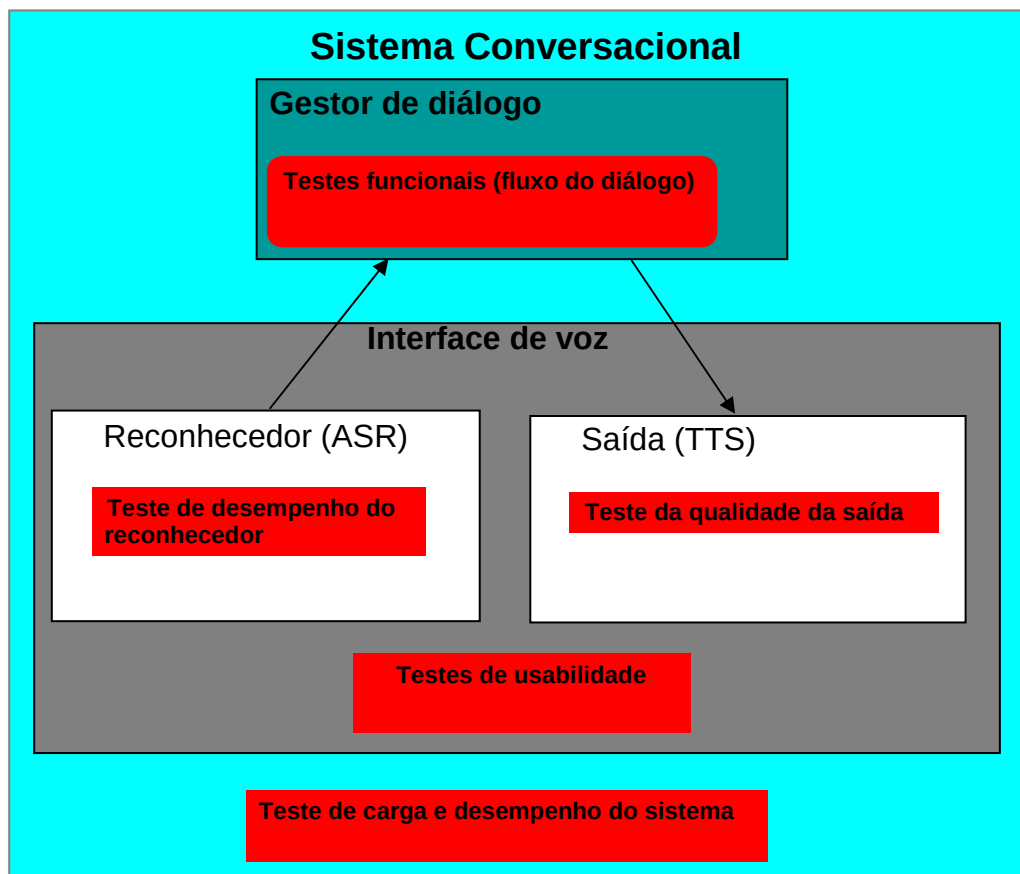


Figura 12 - Arquitectura e testes

Apesar de existir alguma tendência na comunidade de “desenvolvedores” de aplicações de voz para testar a lógica da aplicação (testes funcionais) em simultâneo com o desempenho do reconhecedor, tal não é o mais indicado. São testes que devem ser efectuados de forma independente para que os resultados tenham validade e se possam isolar mais convenientemente eventuais problemas.

5.1 Testes de desempenho do ASR

Um dos principais problemas das aplicações de voz é que o reconhecimento não é perfeito e os erros ocorrem com uma frequência, por vezes, não negligenciável. Testar a *performance* do reconhecimento, particularmente para gramáticas de grande dimensão, pode ser muito complicado e demorado, podendo inclusivamente ser difícil determinar se houve erro ou não no reconhecimento.

É comum as empresas que comercializam sistemas de reconhecimento de fala anunciarem taxas de reconhecimento na ordem dos 90%. É extremamente conveniente e interessante apresentar um número que possa ser facilmente comparado com a concorrência ou que imediatamente indique a qualidade do produto... mas a verdade é que, enquanto integradores ou "desenvolvedores" de sistemas conversacionais, este número não nos diz tudo. Sem a indicação da dimensão e estrutura do vocabulário utilizado nos testes, quantos utilizadores, como foi realizado o treino, quais as condições ambiente (relação sinal-ruído, por exemplo) ou que tipo de microfone foi utilizado (só para citar os aspectos mais comuns) não é possível compreender o significado do número apresentado ou como é que ele se compara com os outros produtos disponíveis no mercado. Portanto, mais importante do que seleccionar um sistema deste tipo baseado apenas num número, é ter a noção que cada sistema de reconhecimento de fala é desenvolvido para aplicações específicas e possui características muito particulares que importa conhecer e analisar:

- tipo de discurso – reconhecer apenas palavras isoladas ou discurso contínuo;
- estilo do discurso – suporte para discurso espontâneo ou não;
- dependência dos utilizadores – pode ser necessário treinar o sistema com os “utilizadores tipo” ou pode ser independente do utilizador (*speaker independent*);
- tamanho do vocabulário suportado;
- robustez ao ruído ambiente;
- tipo de microfone.

e deve ser escolhido de acordo com a finalidade a que se destina.

Após a selecção de um sistema de reconhecimento, é importante avaliar a sua qualidade e aqui consideram-se, normalmente, três métricas: a precisão, a velocidade e a robustez ao ruído. Estas métricas e as metodologias a elas associadas são os temas das subsecções seguintes.

5.1.1 Corpus de voz

A ferramenta básica para a avaliação da qualidade de um sistema de reconhecimento automático de fala é uma base de dados de referência, designada por *corpus*. Um *corpus* de voz contém amostras de voz que reflectem os utilizadores tipo, o vocabulário, a perplexidade, o ambiente e outras condições de utilização. De certa forma, um *corpus* encapsula diferentes perfis de utilização (*usage profiles*). As amostras de voz são complementadas pelas respectivas transcrições e, eventualmente, anotações a vários níveis. É obvio que a qualidade desta base de dados é de fulcral importância para a validade dos resultados. Existem organismos internacionais, como a *European Language Resources Association* (ELRA) (<http://www.elra.info>), o *National Institute Of Standards and Technology* (NIST) (<http://www.nist.gov>), ou o *Linguistic Data Consortium* (LDC) (<http://www.ldc.upenn.edu>), entre outros, que disponibilizam, validam e promovem este tipo de recursos de voz para as mais variadas condições de utilização, domínio de aplicação ou vocabulários. A importância de um *corpus* fica bem patente nas mais variadas aplicações do mesmo:

- no campo da pesquisa em linguística
 - estudos sociolinguísticos
 - *parsing*
 - estudos semânticos
 - prosódia
 - aquisição de linguagem
- no campo das tecnologias de fala
 - síntese e reconhecimento
 - detecção do orador
 - tradução automática
 - detecção automática de tópicos

Da metodologia de teste deve fazer parte uma especificação rigorosa para a construção ou selecção (se já existente) dos corpora a utilizar.

Um dos problemas existentes com alguns corpora para o campo dos sistemas conversacionais é não conterem discurso espontâneo: em muitos casos os dados foram lidos a partir de textos preparados antecipadamente ou elicitados de alguma forma. Por vezes a solução passa por criar um *corpus* à medida para substituir ou complementar os corpora que se possam adquirir. A criação de um *corpus* à medida é uma tarefa extremamente morosa, pois obriga a estudos prévios do tipo de potenciais utilizadores da aplicação a desenvolver, a definição de uma amostra representativa desses utilizadores, especificação do tipo de dados a recolher, definição e execução de uma metodologia de recolha e definição da estrutura final do corpus (ver exemplo de criação de um *corpus* em português em [Branco 05]). No entanto, quando bem construído, dar-nos-á uma medida mais fidedigna de como o sistema de reconhecimento se comportará na aplicação final. Isto é de particular importância nas aplicações de voz disponibilizadas por telefone (*Computer Telephony Integration*), pois este introduz um factor de complexidade bastante grande no ambiente de operação.

É importante notar que o tipo, quantidade e qualidade dos dados usados para avaliar os sistemas de reconhecimento tem impacto directo nos factores de desempenho que podem ser examinados e na significância estatística das conclusões. Corpora maiores fornecem uma melhor representação da diversidade e variabilidade linguísticas. Contudo existe um compromisso a estabelecer entre a quantidade de dados e a morosidade/complexidade de recolha destes.

Dada a importância destas bases de dados de voz, serão dedicadas as secções seguintes à sua selecção e criação.

Seleção de corpora de voz

A selecção do corpora a utilizar para avaliar a qualidade do ASR deve ser rigorosa e, à semelhança da selecção do próprio ASR, deve ser guiada pelos requisitos da aplicação a desenvolver, características dos utilizadores tipo e em função do domínio em causa. Segue-se uma lista das características a ter em conta na especificação dos requisitos dos corpora:

- domínio de aplicação;
- características demográficas dos oradores;
- tipo de discurso – palavras isoladas ou discurso contínuo;
- estilo do discurso – discurso espontâneo ou não;
- tamanho do vocabulário;
- ambiente de gravação (ruído ambiente);

- tipo de microfone usado na gravação;
- dimensão.

Por exemplo, suponhamos que se pretende desenvolver uma aplicação de gestão de correio electrónico por voz para ser acedido em situações de mobilidade (telemóvel). O corpora a seleccionar para testar o ASR da aplicação deve estar nesse domínio, incluir discurso espontâneo, ter sido gravado em ambientes com ruído (preferencialmente com o telemóvel), e as características demográficas dos oradores devem corresponder o mais possível às características dos potenciais utilizadores da aplicação a desenvolver - só para mencionar alguns dos aspectos a ter em linha de conta.

Após termos definido os requisitos do corpora a seleccionar, devemos efectuar um *survey* pelo corpora existente para determinar o que melhor se adequa e qual o seu preço, por exemplo. Este tipo de *survey* pode ser realizado mais ou menos *ad-hoc* ou baseado em *templates* ou questionários disponíveis, adaptando-os a cada caso específico.

Apesar de manifestamente insuficientes, existem vários exemplos de corpora de voz em português. Para que possamos observar as características de um *corpus*, analisaremos o SpeechDat como exemplo. O SpeechDat foi um projecto financiado pela Comissão Europeia para os campos da produção, normalização, avaliação e disseminação de recursos de língua falada (ver <http://www.speechdat.org/SpeechDat.html>). Tal como é referido em [Trancoso 98], a recolha deste *corpus* foi dividida em duas fases. Na primeira fase foram recolhidas 1000 chamadas telefónicas para 8 línguas europeias. Na segunda fase foram recolhidas 4000 chamadas, gravadas nas redes fixas e móveis, para todas as línguas oficiais do União Europeia (e alguns dialectos principais). A base de dados, de acordo com a informação disponibilizada pela ELRA, contém tanto frases isolados como discurso contínuo. Os sinais de fala são armazenados em formato 8-bit 8kHz a-Law. Cada frase é gravada em ficheiros diferentes. Cada ficheiro de áudio tem um ficheiro ASCII correspondente com informação sobre a sessão, data e hora, guião utilizado e transcrição ortográfica. A informação que cada indivíduo forneceu incluía:

- 3 palavras esperadas em serviços telefónicos;
- 6 sequências de dígitos: telefone e cartão de crédito, por exemplo;
- 2 números;
- 1 quantia monetária: em moeda local, na altura;
- 5 expressões de datas: formato numérico e alfanumérico;
- 5 palavras relacionadas com locais geográficos e nomes de empresas;
- 2 respostas sim/não espontâneas.

A base de dados em português do SpeechDat contém gravações de 4027 indivíduos (1861 masculinos e 2166 femininos) na rede fixa. A recolha e anotação foi realizada pela *Portugal Telecom* e encontra-se repartida em 11 CDs, podendo ser adquirida através da ELRA. A distribuição demográfica obtida foi a seguinte: 241 indivíduos abaixo dos 16 anos, 1404 entre 16 e 30, 1532 entre 31 e 45, 711 entre 46 e 60 e 139 indivíduos acima dos 60.

Também existem corpora de outros modos de interacção, como corpora textuais (como o bem conhecido CETEMPúblico - Corpus de Extractos de Textos Electrónicos MCT/Público – que contém mais de 180 milhões de palavras em Português europeu), corpora gestuais, e corpora multimodais.

Um exemplo incontornável de um corpus de voz, embora em Inglês, é o *Switchboard*. O *Switchboard* representa a primeira grande recolha de discurso conversacional espontâneo por telefone. Foi recolhido em 1990 pela Texas Instruments [Godfrey 92].

Criação de um corpus de voz

A criação de um *corpus* é uma tarefa extremamente complexa e morosa, envolvendo, tipicamente, equipas multidisciplinares. Dependendo dos requisitos, poderá inclusivamente levar 1 ou 2 (ou até mais) anos a criar um *corpus*. Poderiam ser escritos capítulos inteiros para cada uma das etapas necessárias para a criação de um *corpus*. Como tal não é viável, nem necessário para o tema do presente documento, apenas serão abordadas de forma sucinta as principais etapas.

A primeira etapa consiste em especificar os requisitos do *corpus* à semelhança do processo de selecção (ver secção anterior). Embora um *corpus* possa ser utilizado quer para treino quer para a avaliação do desempenho de reconhecedores, a dimensão do mesmo para um treino adequado é muito superior ao necessário para a avaliação. A título de exemplo, em 2002 um dos maiores fornecedores de reconhecedores de fala indicava que para se obter uma boa precisão no reconhecimento, o *corpus* telefónico de treino deveria ter os seguintes atributos (macroscópicos) principais:

- mínimo de 1000 oradores, embora mais de 4000 seria preferível;
- mais de 50000 frases, sendo preferível mais de 200000;
- uma chamada por cada orador;
- 50% masculinos, 50% femininos;
- cobrir toda a variedade de idades, com diferentes percentagens;
- várias zonas do país (diferentes pronúncias);
- telefone fixo e móvel;
- diferentes níveis de ruído, sendo que pelo menos 80% dos dados deveriam ter pouco ruído.

Estas características já dão para fazer uma ideia do custo deste processo... A etapa seguinte consiste no desenho, por assim dizer:

- definir a dimensão da equipa e atribuir responsabilidades aos diversos elementos. É fundamental ter bem definidos os papéis dos diversos elementos, como por exemplo, quem fica responsável por criar os convites/acompanhar os oradores, quem fica responsável pelo equipamento de gravação, qual a constituição da equipa que fará a transcrição das amostras de voz, quem fica responsável por fazer o armazenamento dos dados, quem fica responsável por publicar o corpus – só para enumerar algumas das responsabilidades a definir;
- é necessário definir o formato dos dados a armazenar, quer ao nível da voz quer das transcrições e anotações. Ao nível do áudio é comum para corpora telefónico os ficheiros estarem codificados em 8kHz 8 bit μ -Law (ou a-Law). É ainda necessário definir um esquema de nomes que identifique claramente cada ficheiro, contendo, por exemplo, elementos como a data e a identificação da sessão e do orador, bem como outros elementos que se considerem pertinentes;
- definir os tipos de anotação (ver [Leech 98] e [Klein 98]). De forma a possibilitar a reutilização dos dados para efeitos de treino e avaliação dos sistemas de reconhecimento, estudos da estrutura dos diálogos ou para construção de gramáticas, por exemplo, os corpora devem ser transcritos (transcrição ortográfica) e anotados a diversos níveis (morfo-sintáctico, sintáctico, prosódico e/ou pragmático [Bird 01]).

Estas anotações constituem um esquema de codificação (*coding scheme*). A forma como os esquemas são desenhados depende da tarefa e do fenómeno linguístico que se pretende focar. Limitações técnicas ou dificuldades que surjam na criação do corpus podem igualmente condicionar o esquema adoptado. Também por estas razões, a reutilização dos corpora é dificultada. Várias iniciativas como o DRI (*Discourse Resource Initiative*), a iniciativa EAGLES (*Expert Advisory Group on Language Engineering Standards*) ou o projecto MATE (*Multilevel Annotation, Tools Engineering*), procuraram normalizar os esquemas de codificação nos corpora de voz. A tarefa de transcrição/anotação é das que mais recursos consome na criação de um corpus;

- definir uma configuração técnica (*setup*) para efectuar a gravação e escolher ferramentas a utilizar na gravação, transcrição e anotação (ver por exemplo [Bird 02], [Maeda 02], [Gomes 05], [Gloria 05]);
- definir a dimensão e as características da amostra de oradores e tratar dos aspectos logísticos das suas participações.

A próxima etapa é a gravação e aqui há vários factores a ter em conta:

- aspectos legais [Amith 2000]
 - o ética
 - o política comercial
 - o legislação em vigor (gravações de voz e recolha e armazenamento de dados pessoais, entre outros aspectos)
- sessões de gravação. Nesta fase, o aspecto central são as sessões de gravação para as quais se pode recorrer a diversos métodos, consoante os requisitos do corpus. Por exemplo, podemos ter um estúdio de gravação, no qual os oradores podem ler um guião ou responder a perguntas usando um microfone para uma gravação directa. Podemos ter sessões telefónicas, em que o áudio é gravado recorrendo, por exemplo a um IVR, com separação de canais. Uma estratégia normalmente adoptada são as chamadas experiências Feiticeiro de Oz (*Wizard of Oz*) – WOz [Dahlbäck 93]. Neste tipo de experiências, o utilizador opera com um sistema aparentemente funcional, mas na verdade as suas funcionalidades são fornecidas por um operador humano escondido. A técnica WOz mostrou ser uma abordagem particularmente apropriada para a recolha de discurso espontâneo. O senão desta técnica é que introduz um custo adicional na medida em que é necessário preparar protocolos e treino completo para os feiticeiros, para que estes possam simular o sistema de forma consistente em termos de conteúdo, estilo e tempo de resposta.
- segmentação de áudio
 - o identificar intervenções (*turns*) e segmentar
 - o colocar tempos (*timestamps*)
- transcrição/anotação de acordo com normas/requisitos do corpus ([Leech 98], [Klein 98], [Branco 05])

A imagem seguinte, retirada de [Maeda 02] permite ter uma ideia do processo de segmentação e transcrição:

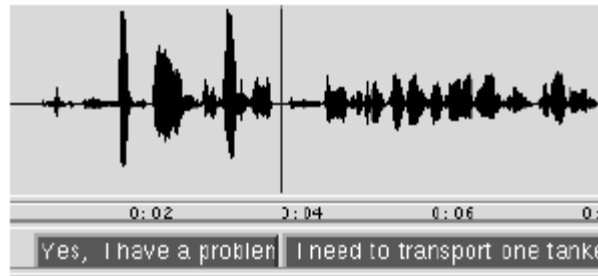


Figura 13 - Segmentação e transcrição (Fonte: [Maeda 02])

Após a etapa anterior de gravação, é necessário fazer um controlo de qualidade final, que inclui

- detectar erros de transcrição;
- detectar erros nos tempos;
- contagem de palavras;
- verificar ficheiros;
- verificar estrutura de armazenamento.

Finalmente faz-se a publicação do *corpus*, que inclui o armazenamento final em CDs ou DVDs, por exemplo, e, eventualmente, submeter a um organismo como a ELRA para que fique disponível para aquisição.

Na secção 6.1 será apresentada a experiência do autor na criação do corpora linguístico.

Alternativa à utilização de corpora de voz

Em alternativa à utilização destas bases de dados, é também possível recorrer a utilizadores para fazerem uso da sua própria voz directamente. Neste caso, e à semelhança da criação de um *corpus*, é necessário todo um trabalho prévio de preparação, como a definição de uma amostra representativa dos utilizadores, textos para que estes leiam ou cenários para que falem livremente, podendo-se, mais uma vez, recorrer à técnica WOZ. No entanto, é necessário definir qual o número certo de utilizadores para que os resultados tenham validade estatística e a logística das sessões de gravação não é, de forma alguma, trivial – como já se viu. Mas, por vezes, recorrer a utilizadores poderá ser a única alternativa se não existir corpora num determinado domínio.

As taxas de erro dos reconhecedores de fala apresentados pelos vendedores são obtidas em condições ditas ideais: as bases de dados são gravadas com equipamento de alta qualidade, com condições ambientais perfeitamente controladas em termos de ruído, para domínios de aplicação e vozes para os quais os modelos foram treinados. Assim sendo, e para ter uma ideia do desempenho do ASR em situação real de utilização da aplicação é importante fazer uma avaliação cuidada. A metodologia a adoptar deve reflectir isso mesmo. Para o desempenho óptimo do ASR é recomendável seguir as indicações do fabricante ao nível do *setup* do equipamento, treino, selecção do microfone, entre outras.

5.1.2 Testes estatísticos

Precisão e velocidade do ASR

Quanto à precisão do ASR consideram-se, tipicamente, três tipos de erros: inserção, remoção e substituição de palavras. Podemos incluir outros erros como falsa aceitação (palavra não está na gramática e devia ser rejeitada), falsa rejeição (falha de reconhecimento de palavra válida) ou falso alarme (reconhece uma palavra chave onde ela não existe). A métrica *de facto* utilizada na avaliação da precisão de um ASR é a taxa de erro (*Word Error Rate* – WER). O WER num determinado *corpus* representa a distância entre a transcrição correcta e as hipóteses de reconhecimento e é definida pela seguinte equação

$$WER = \frac{I + D + S}{N} \times 100$$

em que N é o número de palavras na transcrição correcta, S é número de erros de substituição, D o número de erros de remoção e I é o número de erros de inserção. A taxa de reconhecimento, também conhecida por *Word Accuracy* (WA), pode ser definida com (100-WER).

A métrica WER tem sido bastante útil no desenvolvimento e melhoramento dos algoritmos de reconhecimento, principalmente porque os grupos de investigação tendem a usar os mesmos corpora, podendo assim comparar os resultados obtidos entre si. No entanto, o desempenho dos reconhecedores assim aferida depende em grande medida das características dos corpora de teste, como o tamanho do vocabulário, a perplexidade ou as condições ambiente. Ou seja, as comparações directas dos reconhecedores pode apenas ser feita quando testados nos mesmos corpora.

Apesar da importância do WER, nesta métrica todas as palavras de uma frase são tratadas da mesma maneira, quer sejam palavras-chave (*keywords*) ou irrelevantes (*fillers*). Mas quando falamos em “compreensão” da fala é obvio que determinadas palavras são mais importantes para a extracção da informação relevante em domínios abertos e, mais recentemente, fala-se em taxa de erro ponderada (*Weighted Keyword Error Rate* – WKER) [Nanjo 05], em que cada palavra tem um peso diferente consoante a sua influência.

Na maior parte das situações reais, mais do que obter um número (WER ou WKER), é importante fazer uma análise detalhada dos padrões de erros. Um número anormal de erros indica-nos um claro problema do ASR ou nas condições de teste. Um elevado número de remoções e/ou substituições pode-nos indicar que o desempenho do ASR se degrada com a velocidade da fala e que, nesse caso, não será indicado para tarefas que exijam essa velocidade.

Procurando ultrapassar a dependência do WER com os corpora usados nos testes, tem ganho alguma relevância a comparação do desempenho dos reconhecedores com seres humanos (ver por exemplo [Cox 98] e [Shinozaki 03]). Neste caso, tanto ao reconhecedor automático como ao ser humano, é pedido que escolham as palavras mais prováveis de um dicionário dado um determinado sinal de voz. O desempenho é comparado usando os mesmos critérios. Através da análise dos resultados é possível determinar a causa dos erros do reconhecedor automático, como por exemplo, falta de robustez face a diferentes pronúncias, modelos inadequados, entre outros... algumas palavras podem ser impossíveis de reconhecer até para os seres humanos. Com esta abordagem procura-se obter uma métrica mais relacionada com o desempenho humano do que com o WER, “calibrar” os corpora e obter resultados que sejam menos dependentes das bases de dados de voz.

A avaliação da precisão do ASR é feita, basicamente, através da comparação da saída do ASR (texto hipotético) com o texto de referência que é a transcrição correcta da fala, usando para tal ferramentas como o *sclite* (que é parte do NIST SCTK - *National Institute of*

Standards and Technology Scoring Toolkit). Este processo tem o nome de alinhamento. Após o alinhamento é necessário efectuar os cálculos para se obterem as métricas pretendidas para cada utilizador e cada par hipótese-referência, processo conhecido como *scoring*. A figura seguinte retirada de [Greenber 00] ilustra a saída do *sclite*.

ID	REF WD	HYP WD	WS	RB	HB	RP	HP	PS	RB	HB
2	ONE	ONE	C	5.70	5.70	W	W	C	5.71	5.70
0						AH	AH	C	5.80	5.81
4						N	N	C	5.88	5.87
0	UGHT	****	D	5.90		AO	AO	C	5.92	5.90
B						T	D	S	6.06	6.07
0	EITHER	AUDITOR	S	6.10	5.90	IY	IH	S	6.10	6.12
0						DH	T	S	6.30	6.27
1						ER	ER	C	6.34	6.35
1	TO	TO	C	6.43	6.45	T	T	C	6.44	6.45
B						AX	AX	C	6.52	6.52

Figura 14 - Saída do *sclite* (Fonte: [Greenberg 00])

A coluna ID identifica a sequência completa de palavras submetidas ao reconhecedor. A coluna REF WD são as palavras de referência; a coluna HYP WD é a saída do ASR; WS – de *word score* – é o resultado da palavra, onde C=Correcto e S,D,I têm os significados apontados anteriormente na explicação da fórmula do WER; as outras colunas têm a ver com o tempo gasto, os fonemas, entre outros elementos irrelevantes para o momento. Com esta informação podemos obter o total de erros, os totais parciais de cada tipo de erros, as suas percentagens, o WER, bem como a velocidade de reconhecimento. Como já foi referido, na maior parte das situações reais, mais do que um obter um número (WER ou WKER), é importante fazer uma análise detalhada dos padrões de erros. É interessante correlacionar os parâmetros dos oradores e das características linguísticas e acústicas do *corpus* utilizado. O problema é que existem demasiados parâmetros para correlacionar com os erros e/ou velocidade do reconhecimento: número de palavras nas frases, duração das frase, posição das palavras nas frases, energia da palavra, alinhamento das hipóteses com as referências, prosódia, sotaque do orador, sexo do orador, idade, entre muitos outros. Assim sendo, em muitos casos não podemos ambicionar ser exaustivos, e temos de seleccionar cuidadosamente os aspectos que pretendemos correlacionar. Esta selecção terá de ser guiada pela definição de requisitos da aplicação na qual o ASR será integrado.

Robustez do ASR

O ambiente de operação de um sistema de reconhecimento de voz pode variar desde um ambiente controlado, sem ruído, até um ambiente fabril. A capacidade de um ASR manter um bom nível de reconhecimento em condições adversas dá-se o nome de robustez. O ruído é a parte do sinal que chega ao reconhecedor que não tem significado (ver também secção da Teoria da Comunicação do presente documento): ruído de fundo, ruído do canal. O ruído de fundo pode incluir máquinas, pessoas, eco, entre outros. O ruído do canal refere-se ao efeito do aparelho de captura (microfone, telefone) no sinal de voz. No caso de aplicações telefónicas, quer o microfone do aparelho quer a própria rede introduzem ruído, sendo que a rede ainda atenua as frequências abaixo de 100Hz e acima dos 3100Hz. O ruído varia muito de rede para rede, conforme as condições, qualidade ou o equipamento. Infelizmente, praticamente não existe possibilidade de controlar as características exactas de uma linha

particular na rede. No caso dos telefones móveis, o problema é ainda mais complicado pois o grau de variabilidade é maior ainda e, tipicamente, o nível de ruído de fundo é maior (situações de mobilidade). Para além disso, os utilizadores tendem a alterar as características da sua voz em resposta ao ruído do ambiente (falar mais alto, mais devagar...), o que dificulta também o reconhecimento automático. É, portanto, muito importante fazer a avaliação da robustez do ASR em relação às diferentes condições de degradação.

Um importante indicador do potencial impacto do ruído na precisão do reconhecimento é a relação sinal-ruído (*Signal-to-Noise Ratio*, SNR ou S/N) num dado ambiente. A relação sinal-ruído designa a razão entre um a informação com significado e o ruído ambiente:

$$SNR = \frac{\text{Potência}_{\text{Sinal}}}{\text{Potência}_{\text{Ruído}}}$$

O SNR é normalmente expresso em decibéis (dB). O decibel é o logaritmo da razão entre o valor medido e um valor de referência padronizado e corresponde à mais pequena variação de pressão sonora que o ouvido humano normal pode distinguir em condições normais. Em decibéis o SNR é 20 vezes o logaritmo de base 10 da razão entre as amplitudes RMS (*root-mean-square*) do sinal e do ruído:

$$SNR(dB) = 20 \log_{10} \frac{\text{Amplitude}_{\text{Sinal}}}{\text{Amplitude}_{\text{Ruído}}}$$

Este valor indica uma medida média, no sentido em que SNR instantâneos podem variar consideravelmente. Geralmente, quanto maior for este valor melhor: mais livre de ruído é o sinal. A medição do ruído ambiente num determinado local pode ser determinado conforme [Instituto Ambiente 03], recorrendo a um sonómetro. Preferencialmente, a marca e modelo do equipamento deve estar homologada pelo Instituto Português da Qualidade (IPQ).

A avaliação da robustez do ASR em relação às diferentes condições de degradação do ruído ambiente pode ser feita de forma eficiente com um auxílio de uma ferramenta que permita simular as diferentes condições. As degradações de ruído são introduzidas nos ficheiros de áudio pré-gravados (*corpus*) limpos (sem ruído ou com o ruído mínimo), mantendo, assim, constantes variáveis como o orador ou a informação falada. Podemos, por exemplo, ir aumentando gradualmente o ruído de fundo e submetendo estas amostras de fala ao ASR e analisando a variação da taxa de erro consoante o ruído (ver, por exemplo, [Möller 02a]).

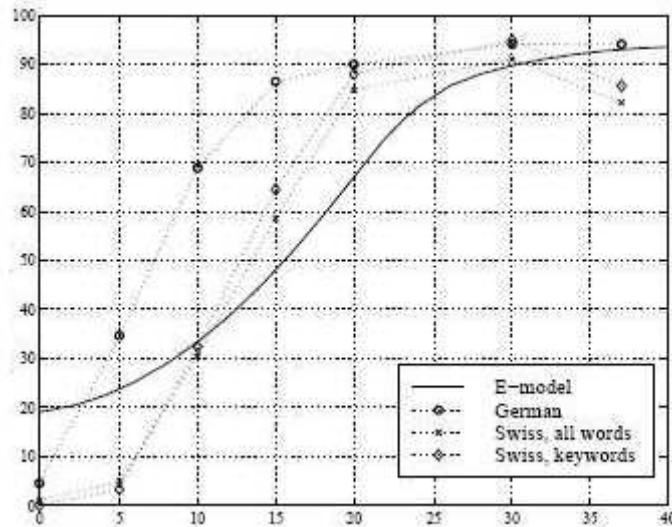


Figura 15 - Taxa de reconhecimento consoante SNR (Fonte: [Möller 02a])

A figura acima, retirada de [Möller 02a], ilustra a degradação da taxa de erro (eixo Y) consoante o SNR (eixo X) para 3 sistemas e compara com a previsão para a comunicação Humano-Humano (linha a cheio: E-model). É perfeitamente claro que quanto maior a SNR (ou seja, amostras mais limpas), maior a taxa de reconhecimento.

A precisão do reconhecimento ressent-se fortemente com a diminuição do SNR. No gráfico, para SNRs abaixo dos 15-20dB, a taxa de reconhecimento começa a cair rapidamente. Esta é também uma razão pela qual as taxas de reconhecimento anunciadas pelos vendedores de ASRs são pouco fiáveis.

Concluindo, para uma real avaliação de um ASR devemos ter em conta o ambiente em que este irá operar:

- natureza do ruído de fundo
- variabilidade e potência do ruído de fundo e do canal
- tipo e qualidade do canal

5.2 Testes de qualidade do TTS

A percepção e reacção dos utilizadores a uma interface de voz também passam muito pela qualidade da síntese de voz que esta apresentar. Importa não só a componente auditiva mas também a componente do conteúdo, ou seja, não só como é transmitido (foneticamente) mas também o que é transmitido. É importante testar e avaliar a componente da síntese de voz em separado da estrutura do diálogo, à semelhança da separação entre a lógica da aplicação (testes funcionais) e o desempenho do reconhecedor. Uma estratégia frequentemente adoptada para melhorar a qualidade perceptível da saída de voz do sistema é intercalar partes de áudio pré-gravado com partes sintetizadas.

Ao nível da avaliação de qualidade, a tarefa também não é fácil dada a multidimensionalidade do próprio termo. Tipicamente, o que é avaliado é a inteligibilidade e a naturalidade da fala gerada. Por inteligibilidade entende-se a sintetização correcta dos fonemas, já a naturalidade está ligada ao tom, à duração, a sintetização correcta dos alofones (variantes fonéticas dos fones, ou seja, diferentes realizações de um mesmo fonema). Por vezes, também se inclui nos parâmetros de teste a capacidade do TTS em processar abreviaturas ou sequências numéricas, vulgarmente designado por *Front-End Processing*.

Estas dimensões (naturalidade, inteligibilidade e capacidades de *Front-End Processing*) não são independentes, o que dificulta ainda mais a avaliação.

Uma regra normalmente aplicada à avaliação de qualidade do TTS é que quanto mais manualmente podermos avaliar, melhor: ou seja, é difícil e pouco desejável automatizar este tipo de testes, no sentido em que é a opinião do utilizador (leia-se “amostra de utilizadores”) que permite comparar qualitativamente dois sistemas TTS. Ver, por exemplo, as recomendações da *International Telecommunication Union* (ITU), nomeadamente a P.800 (*Methods for subjective determination of transmission quality*). A contrapartida é, invariavelmente, o elevado custo de preparar e executar as sessões de avaliação com os utilizadores e a posterior análise dos resultados (esta última fase, sim, pode e deve ser automatizada o mais possível). Estas avaliações com utilizadores devem ser realizadas em ambientes o mais próximo possível do ambiente de operação pretendido para a aplicação. E, claro, a qualidade em si pode não ser o único factor de decisão: outros factores como o domínio da aplicação final (diferentes domínios têm diferentes requisitos), como o preço, a interoperabilidade com o resto do sistema, suporte ou custo em termos de memória/disco (entre outras que se possam considerar importantes para cada caso em particular), são igualmente factores importantes.

Mais recentemente, é proposto que se deve não só avaliar a inteligibilidade e naturalidade como também mais parâmetros como a atitude ou a emoção, por exemplo, [Ambrazaitis 05].

5.2.1 Inteligibilidade

Os testes de inteligibilidade estão mais relacionados com a capacidade de identificar o que foi sintetizado.

Para testar a inteligibilidade podemos usar abordagens como tarefas de compreensão e/ou tarefas de transcrição, as quais incluem testes como o teste das frases semanticamente imprevisíveis (*Semantically Unpredictable Sentence Test* - SUS).

Nas tarefas de compreensão, as frases são sintetizadas e é perguntado aos utilizadores para dizerem qual o assunto que estava a ser tratado. Pode acontecer que apesar de ser “claro” o que é sintetizado, o utilizador tem de fazer um esforço para “acompanhar” o sinal, deixando de prestar atenção ao conteúdo/assunto, perdendo detalhes.

Um tipo de teste comum é o *Diagnostic Rhyme Test* (DRT), no qual se procura testar a inteligibilidade das consoantes no início das palavras. São sintetizadas pares de palavras parecidas, mas com consoantes iniciais diferentes (por exemplo, “penta” e “lenta”), e é pedido ao utilizador que identifica qual a palavra que ouviu.

Como exemplo de testes de transcrição temos o já referido teste das frases semanticamente imprevisíveis (*Semantically Unpredictable Sentence Test* – SUS). No SUS é pedido aos utilizadores para transcrever frase que não têm significado por si próprias ou contexto – procurando minimizar a possibilidade de derivar informação que não a contida apenas no sinal.

5.2.2 Naturalidade

Para testar a naturalidade (e também a inteligibilidade) podemos usar uma escala qualitativa (bom, médio, mau), a que corresponde uma escala numérica (tipicamente de 1 a 5) de forma a obter uma medida média de opinião quantitativa (*Mean Opinion Score* – MOS), a partir dos seguintes conjuntos de dados:

- palavras isoladas (mais frequentes, menos frequentes, média frequência);
- palavras compostas;

- frases curtas;
- frases longas;
- frases com bastante prosódia.

Estes dados podem ser obtidos a partir de um *corpus* (ver secção 5.1.1 do presente documento), seleccionando-os de forma aleatória ou seguindo critérios como frequência, relevância para o domínio, ou outros que se pretendam. Através do valor MOS podemos, então, comparar qualitativamente diferentes TTS.

Como possíveis erros que podemos encontrar, para palavras isoladas temos:

- pronúncia incorreta;
- duração incorreta;
- fraca qualidade sonora;
- palavra ininteligível.

Para frases:

- pronúncia incorreta;
- duração incorreta;
- fraca qualidade sonora;
- ritmo errado;
- pausas erradas.

Quando conduzidos com cuidado, os testes MOS fornecem um indicador precioso na avaliação da qualidade de um sistema TTS. No entanto os resultados MOS só têm validade quando comparamos sistemas com os mesmos utilizadores, as mesmas frases, os mesmos ambientes de teste, as mesmas condições. Dito de outra forma, os valores MOS são muito sensíveis às condições da experiência. E é conveniente, para significância estatística, que as dimensões quer da amostra de utilizadores, quer da amostra de frases, sejam relativamente elevadas.

5.2.3 Qualidade do Front-End Processing

Para avaliar a qualidade do *front-end processing* é comum seleccionar um conjunto de frases, em variados contextos, que contenham abreviaturas, sequências numéricas, datas ou números de telefone, e submetê-las ao sintetizador de fala e analisar o comportamento deste. Esta análise pode ser simplesmente marcar as frases correctas e as incorrectas, ou pode também incluir a atribuição de pesos aos erros consoante a sua relevância para o domínio/aplicação final. Neste tipo de teste, a fatia maior do custo associado está na preparação do conjunto de frases de teste, balanceando o número de casos com a relevância para o domínio. Podemos comparar diferentes tipos de sistemas TTS de uma forma objectiva a este nível.

5.2.4 Outros testes

Como já foi referido, a contrapartida dos testes “manuais” para aferir a qualidade da fala referidos nas secções anteriores é elevado custo de preparar e executar as sessões de avaliação com os utilizadores. Na tentativa de automatizar, reduzir este custo elevado e possibilitar a fácil reprodutibilidade das experiências, têm sido feitos progressos no sentido de estabelecer métricas objectivas. Entre outros testes possíveis, podemos referir o PESQ (*Perceptual*

Evaluation of Speech Quality) [Rix 01], que complementou e substituiu o PSQM (*Perceptual Speech Quality Measurement*) [Dai 00]. O PESQ, que é uma recomendação da *International Telecommunication Union* (ITU), é um algoritmo que procura prever as reacções dos utilizadores às distorções de sinal, particularmente as introduzidas pela rede telefónica. A essência é tentar prever o valor MOS (subjectivo, obtido através dos testes pelos utilizadores), comparando o sinal de voz “distorcido” quando transmitido através de vários *codecs* e meios de transmissão com o sinal original (à entrada da rede). O resultado do PESQ está então, supostamente, correlacionado com o valor MOS subjectivo. Verificou-se, no entanto, que esta suposição é algo optimista [Pennock 02]. Foram então desenvolvidos mapeamentos alternativos para o resultado do PESQ que mostraram possuir uma melhor correlação com os resultados dos testes subjectivos: o PESQ-LQ e o PESQ-LQO. Os resultados do PESQ são calibrados usando uma enorme base de dados de testes subjectivos.

A ideia do PESQ é bastante atractiva na medida em que parece permitir avaliar automaticamente qualquer tipo de sistema de áudio e fornecer uma indicação da qualidade do sistema. No entanto, não é tão simples quanto isso. Embora seja uma excelente indicador, deve ser usado em conjunto com os métodos “tradicionais” indicados nas secções anteriores, e não como um substituto, pois existem factores que não são cobertos pelo algoritmo e que também influenciam a “qualidade” percebida pelos utilizadores humanos - como o volume, por exemplo.

A figura seguinte, retirada de [Rix 01], ilustra a estrutura do modelo PESQ:

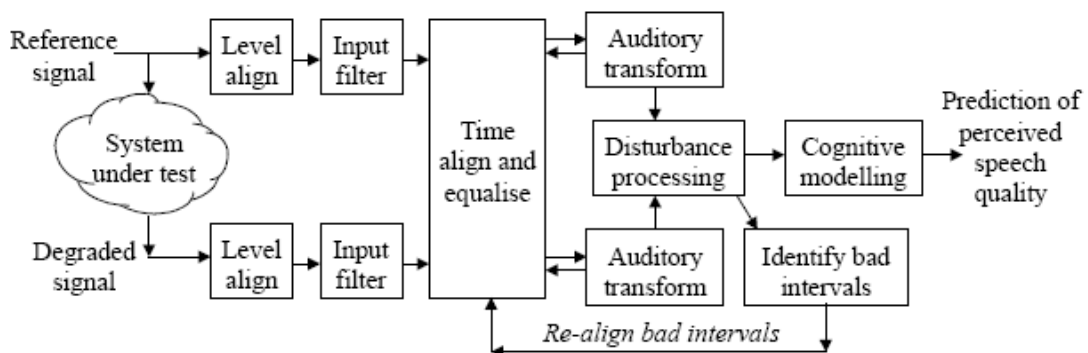


Figura 16 - Estrutura do modelo PESQ (Fonte: [Rix01])

Existem já disponíveis no mercado soluções (plataformas, pacotes integrados, serviços), baseados neste e outros *standards*, para efectuar, de forma automática, o teste da qualidade de fala, como o *Empirix Voice Quality Test Suite* (<http://www.empirix.com>), com o objectivo de diminuir quer os custos associados aos testes subjectivos quer o tempo-para-o-mercado de novas aplicações.

5.3 Testes funcionais do Gestor do Diálogo

Como já foi referido no capítulo anterior do presente documento, os sistemas conversacionais podem ser classificados em 3 grandes grupos consoante os métodos utilizados para controlar o diálogo com o utilizador:

- baseados em máquinas de estados finitos;
- baseados em *frames*;
- baseados em agentes.

Estes métodos de controlo do diálogo são implementados pelo gestor do diálogo (*dialog manager*).

Nos sistemas baseados em máquinas de estados as transições de estado para estado são desencadeadas através de frases do sistema (*prompts*) e respostas (*utterances*) do utilizador. O sistema, perfeitamente determinístico, mantém o controlo do diálogo e desencadeia as acções apropriadas. Este fluxo do diálogo é o mais utilizado e, como tal, será o principal foco da presente secção.

Nos sistemas baseados em *frames*, são colocadas questões ao utilizador de forma a ir preenchendo (*slot filling*) um *template* de forma a executar uma tarefa, tal como apresentar um horário, por exemplo. Neste tipo de sistema o fluxo do diálogo não está predeterminado, dependendo dos pedidos do utilizador e da informação que o sistema necessita.

Os sistemas baseados em agentes são desenhados para permitir uma comunicação complexa entre o utilizador e o sistema de forma a resolver uma determinada tarefa. O modelo de diálogo mantém, tipicamente, um histórico do contexto anterior e, em função disso, evolui dinamicamente em sequências lógicas.

Para além do teste do sistema conversacional como um todo, é também importante, à semelhança do TTS e do ASR, testar o gestor do diálogo individualmente.

Em termos de testes funcionais, o tipo de testes que tradicionalmente é feito ao fluxo do diálogo, isto é, à interface de voz, consiste em interagir manualmente com a aplicação, percorrendo o diagrama de estados directamente ou recorrendo a cenários (*scripts*) de utilização. A figura seguinte ilustra o diagrama de estados de uma hipotética interface de voz muito simples de uma aplicação para uma pizzeria.

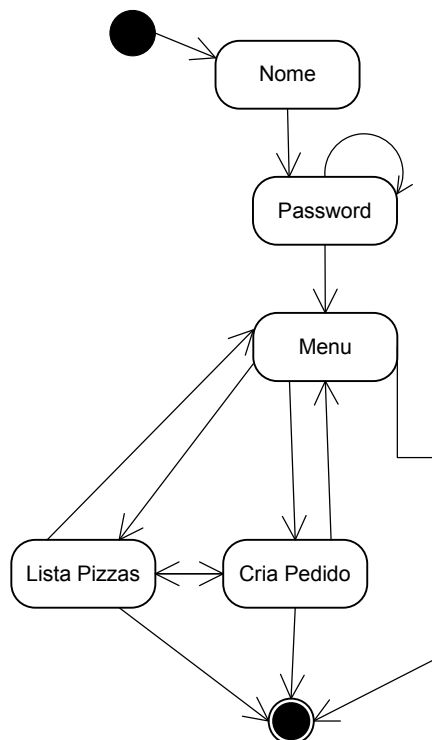


Figura 17 - Diagrama de estados de uma interface de voz simples

O serviço representado na figura acima começa por pedir o nome ao utilizador, valida esse nome e pede a *password*, permitindo, por exemplo, 3 tentativas para a introduzir correctamente. Assim que o fizer, será apresentado ao utilizador um menu principal, na qual ele poderá navegar para as páginas de voz que apresentam a lista de pizzas, criam pedidos ou permitem sair do sistema. É importante notar que o diagrama também evidencia a liberdade que é dada ao utilizador para navegar entre páginas de voz: em qualquer altura, os utilizadores podem aceder a qualquer página de voz. Por exemplo, pode estar a ouvir a lista de pizzas e

avançar logo para os pedidos sem passar pelo menu, ou até sair da aplicação quando quiser: o utilizador tem o controlo. O sistema vai dando sempre *feedback* do reconhecimento dos comandos do utilizador e indica o estado em que está. Em cada passo é fornecido ao utilizador a lista de comandos disponíveis nesse momento para ajudar a navegação e a conclusão das tarefas por parte do utilizador.

No caso de serviços telefónicos, o teste de percorrer o diagrama de estados directamente ou recorrendo a cenários (*scripts*) de utilização, é feito, concretamente, realizando manualmente chamadas telefónicas para o número de acesso do serviço. Esta metodologia, além de enfadonha e morosa, devido à natureza normalmente sequencial da interface, é muito propícia a erros. Se a interface for mais complexa ou se ocorrerem erros de reconhecimento, testar o fluxo do diálogo torna-se ainda mais trabalhoso e complexo. Estes testes costumam demorar entre algumas horas a semanas, dependendo obviamente da complexidade da interface de voz. A solução para pelo menos alguns destes problemas passará por testar o fluxo do diálogo sem recorrer à voz ou ao telefone: interagir com o modelo de estados através de comandos de texto, por exemplo, poderá acelerar e reduzir os erros do processo em grande medida e permitir uma separação maior entre testes funcionais (fluxo do diálogo) e testes ao desempenho do reconhecedor.

Um tipo de testes funcionais do gestor de diálogo mais comum dado os bons resultados obtidos são os testes baseados em cenários. Os cenários podem ser vistos como sequências de interacções (“histórias hipotéticas” [Kaner 03]) entre os utilizadores e o sistema, representando fluxos normais e anormais de eventos dos casos de uso. Através de uma matriz de rastreabilidade podemos anotar o mapeamento entre os cenários de teste e os casos de uso e deve ser identificado pelo menos um caso de teste, e respectivas condições, por cada cenário. A grande dificuldade encontra-se na sequenciação dos cenários e na influência que esta pode ter no teste. Na secção 5.3 (Testes funcionais do Gestor do Diálogo) do presente documento será apresentado um caso concreto de teste baseado em cenários no qual o autor participou.

5.4 Testes de carga e desempenho do sistema

Antes de ser lançada para o mercado, uma aplicação telefónica de voz tem de ser submetida a uma série de testes de carga e *stress*, para avaliar até que ponto cumpre os requisitos. A maior parte das aplicações telefónicas têm um limite máximo de chamadas que podem receber. O número de portos que o equipamento telefónico disponibiliza representa o limite máximo de chamadas simultâneas possíveis, uma vez que um porto serve uma chamada. A experiência mostra que o número de portos é um dos indicadores mais importantes das aplicações de voz.

Para efectuar testes de carga e desempenho das aplicações de voz é necessário, ou pelo menos recomendável, a utilização de uma ferramenta de geração de chamadas (*call generator*) para poder simular a carga desejada. Este tipo de ferramentas necessita de aceder a uma infra-estrutura telefónica que permita realizar múltiplas chamadas de saída simultaneamente. A ferramenta de geração de chamadas necessita igualmente de ser capaz de medir os tempos de resposta da aplicação de voz, tais como os tempos de reconhecimento de voz e DTMF e tempos de reprodução/geração de voz (TTS ou áudio pré-gravado). Quando não existe disponível uma ferramenta deste género, este tipo de teste pode ser efectuado manualmente, pedindo-se a diversas pessoas que liguem para o serviço simultaneamente. Obviamente que este processo manual não é muito escalável e é sujeito a erros na medição dos tempos.

5.5 Avaliação de usabilidade

A Interação Pessoa-Computador é um tema que assume cada vez maior importância. Este tema enfatizou um novo conceito no mundo dos arquitectos e engenheiros de sistemas e soluções informáticas, que é a noção de que os sistemas se devem adequar aos utilizadores e não o contrário. A concepção de interfaces passa por um processo evolutivo, onde os utilizadores devem ser envolvidos em todo o ciclo de desenvolvimento. O principal objectivo desse trabalho é sempre dar resposta às necessidades dos utilizadores com o menor esforço possível destes na aprendizagem e na utilização. É necessário compreender correctamente o problema, os utilizadores, o seu modelo mental e as suas necessidades, bem como definir claramente objectivos de usabilidade e de experiência dos utilizadores. A verdade parece ser que quando se tratam de aplicações reais com utilizadores que pagam pelo mesmo, ou pela sua utilização, não existe uma relação simples e directa entre as taxas de erros, ou qualidades puramente técnicas, e a usabilidade das aplicações de fala.

Existem limitações conhecidas nas avaliações baseadas apenas em corpora, pelo que actualmente é dada cada vez mais importância às avaliações de usabilidade das aplicações de fala. No entanto, trata-se de um tema complexo. Por um lado, as aplicações são criadas para o utilizador, portanto factores de usabilidade, como a satisfação, deveriam ser os derradeiros critérios. Por outro lado, esses mesmos factores de usabilidade podem ser subjectivos e dependentes dos factores técnicos da implementação. Outro factor de complexidade na avaliação de usabilidade advém das interações ou correlações existentes entre as métricas que se seleccionem para essa mesma avaliação (a própria selecção das métricas pode ser um problema).

O trabalho de avaliação de uma interface requer a utilização de técnicas de monitorização da utilização do sistema por utilizadores reais, observação e definição de métricas para avaliação quantitativa e qualitativa. Nesta secção procura-se abordar de uma forma compreensiva algumas das técnicas e métricas de avaliação mais adequadas às interfaces de voz.

5.5.1 Prototipagem

A prototipagem consiste em criar modelos de forma a permitir testar atributos do produto final sem que este esteja concluído. Na maior parte das interfaces com o utilizador este modelo pode ser simplesmente um conjunto de cartões ou desenhos da interface final ou algo mais complexo como código funcional. Quanto mais próximo do produto final o modelo for melhor poderá ser a avaliação, mas também maior será o custo do processo de prototipagem.

Esta técnica pode ser usada em qualquer fase do ciclo de desenvolvimento, tendo em conta que em estados mais avançados do mesmo a construção de novos protótipos tende a ultrapassar o custo de usar versões mais antigas do produto.

Podemos falar de prototipagem rápida, prototipagem modular e prototipagem reutilizável. A primeira é um metodologia em que são desenvolvidos rapidamente protótipos de acordo com alterações no desenho do produto, é feita a avaliação e o protótipo é descartado aquando de uma nova iteração do desenvolvimento do produto. Na prototipagem modular são acrescentadas novas partes ao protótipo existente para acompanhar os desenvolvimentos feitos no produto. Na prototipagem reutilizável, partes do protótipo são incorporados no produto que se está a desenvolver, havendo assim um reaproveitamento do esforço dispendido na construção do protótipo.

Em termos de características do produto final cobertas pelo protótipo, podemos falar de prototipagem horizontal, onde são incluídas muitas funcionalidades/características mas a maior parte não funcionam, e prototipagem vertical, na qual se incluem poucas funcionalidades mas completamente (ou quase) funcionais.

É comum falar-se também em prototipagem de baixa fidelidade e de alta-fidelidade. Na de baixa fidelidade, o protótipo não se “parece” necessariamente com o produto final. Uma vez que os protótipos assim produzidos possuem um custo baixo em termos de tempo e dinheiro, é possível ter mais ciclos de teste, mais utilizadores e mais protótipos. A sua eficácia é maior nos primeiros estágios de desenvolvimento. Um exemplo clássico de prototipagem de baixa fidelidade são os desenhos em papel de cada ecrã da interface. Na prototipagem de alta-fidelidade o protótipo é criado para ser o mais próximo possível do produto final, mas o custo também é bastante mais elevado que na prototipagem de baixa fidelidade. É possível utilizar partes do protótipo no produto final e, sendo um bom protótipo, usá-lo em testes quantitativos. Os protótipos da interface são geralmente construídos com o recurso a pacotes RAD (*Rapid Application Development*) ou pacotes multimédia.

5.5.2 Entrevistas e questionários

Uma forma de identificar o que os utilizadores fazem, pretendem fazer, gostam ou não, é observá-los. Outra forma possível de atingir objectivos semelhantes é recorrer a entrevistas e questionários, que podem ser utilizadas em vários paradigmas de avaliação. As entrevistas e questionários podem ser estruturadas ou mais flexíveis como um diálogo. Quanto às entrevistas, a abordagem mais adequada depende dos objectivos da avaliação, do paradigma e das questões a ser abordadas. Já os questionários são particularmente interessantes para recolher dados demográficos dos utilizadores e as suas opiniões. Uma vantagem dos questionários é que é possível fazê-los chegar “facilmente” a uma população maior.

5.5.3 Avaliação heurística

A avaliação heurística ([Nielsen 1990] e [Nielsen 1994]) é um método da engenharia de usabilidade para detectar problemas de usabilidade no desenho da interface de forma a que possam ser solucionados no processo iterativo de interacção. Enquadra-se no paradigma de avaliação designado por “avaliação preditiva”. Em [Dix 04], por exemplo, é proposto que a avaliação heurística seja utilizada em qualquer etapa do processo de desenvolvimento dada a sua flexibilidade.

Este método envolve a inspecção sistemática da interface do utilizador por um pequeno grupo de peritos que verificam a aderência da interface a princípios de usabilidade (as “heurísticas”), originalmente obtidas empiricamente da análise de centenas de problemas de usabilidade. Esses princípios são baseados em:

- diálogo simples e natural;
- “falar” a linguagem dos utilizadores;
- minimizar a sobrecarga da memória dos utilizadores;
- consistência;
- *feedback*;
- saída bem sinalizada;
- atalhos;
- mensagens de erro claras;
- prevenção de erros;
- ajuda e documentação.

Na realização da avaliação heurística, cada perito inspeciona a interface sozinho, no papel de utilizador. Normalmente não se limita a seguir uma *checklist* com os princípios, adoptando antes uma atitude proactiva. Quando todos os peritos tiverem terminado, juntam as conclusões a que chegaram separadamente.

Esta técnica é rápida e relativamente barata, mas também existem limitações: com o surgimento de novos produtos e modos de interacção (como os telemóveis ou tecnologias colaborativas, entre outros), a lista de princípios de usabilidade pode ser insuficiente ou inadequada.

5.5.4 Wizard of Oz

Nas chamadas experiências Feiticeiro de Oz (*Wizard of Oz*) – WOz [Dahlbäck 93] o utilizador opera com um sistema aparentemente funcional, mas na verdade as suas funcionalidades são fornecidas por um operador humano escondido. O nome provém do livro (e posterior filme com o mesmo nome) *The wonderful Wizard of Oz*, no qual um homem comum escondido atrás de uma cortina fingia ser um poderoso feiticeiro recorrendo à tecnologia, e foi associado pela primeira vez a este tipo de experiência no início da década de 80 por John Kelley [Kelley 84].

A finalidade típica do WOz é gravar e analisar as entradas do utilizador e a saída do feiticeiro. A entrada do utilizador (corpora) é particularmente importante, como se viu anteriormente no presente documento, para o treino e avaliação dos sistemas ASR. O senão desta técnica é que introduz um custo adicional na medida em que é necessário preparar protocolos e treino completo para os feiticeiros, para que estes possam simular o sistema de forma consistente em termos de conteúdo, estilo e tempo de resposta.

5.5.5 Testes de usabilidade

Os testes de usabilidade podem ser definidos como experiências realizadas com a finalidade de encontrar informação específica sobre um determinado desenho/solução de implementação com base na análise estatística dos dados obtidos. No entanto, é dada cada vez mais ênfase à interpretação das observações efectuadas no decurso das experiências do que à “frieza” dos números finais obtidos. Os testes podem ser realizados ao longo de todo o processo de desenvolvimento para validar um determinado desenho, para refinar o mesmo ou para verificar se cumpre os objectivos.

A ideia base no paradigma de testes de usabilidade é relativamente simples: reunir um conjunto de utilizadores, analisar a forma com interagem com o produto e recolher dados como o tempo que demoram a realizar determinada tarefa ou o número de erros, entre outros. As técnicas utilizadas neste paradigma são a observação dos utilizadores, questionários, entrevistas e - a principal - testes com utilizadores. É importante começar por definir quais são os objectivos da experiência (por exemplo, até que ponto os erros de reconhecimento afectam uma utilização multimodal). De seguida é necessário “conhecer” os utilizadores finais do sistema, o seu perfil, e definir com rigor uma amostra representativa dos mesmos para participarem no testes – qual o nível de experiência com sistemas similares, faixa etária, género, nível de escolaridade, só para mencionar algumas características mais óbvias. E pensar, naturalmente, na forma de os recrutar. É comum considerar-se que uma amostra de cerca de 20 utilizadores identifica entre 80 a 90% dos problemas de usabilidade.

O passo seguinte consiste em desenhar a experiência propriamente dita de forma a obter a informação que pretendemos e eliminar (ou minimizar) as variáveis supérfluas. As tarefas que os utilizadores irão realizar devem ser desenhadas com base em cenários reais de utilização do produto. É necessário definir o significado do sucesso em cada uma das tarefas e determinar as condições nas quais as tarefas irão ser realizadas - como o laboratório, o

hardware, o *software*, equipamento de áudio e/ou vídeo, telefones e tudo mais que seja necessário.

Finalmente é necessário designar a equipa responsável pelos testes. O mínimo, obviamente, é um elemento, mas podem ser necessários elementos para receber e acompanhar os utilizadores, para operar o equipamento utilizado, para observar as experiências e outros para analisar os dados.

Após os passos anteriores é necessário executar os testes e recolher os dados. Após receber os utilizadores é importante colocá-los perfeitamente relaxados e confortáveis com as tarefas que irão desempenhar, prestando-lhes todas as informações necessárias e explicando-lhes quer o procedimento quer o facto de poderem parar em qualquer momento ou fazer pausas sempre que necessário, agradecendo-lhes a disponibilidade. Para alguns tipos de testes é necessário ainda que o utilizador dê a sua autorização por escrito para a gravação de áudio e/ou vídeo. Pode ser entregue ao utilizador um questionário pré-testes para obter mais informações sobre o seu perfil.

Quanto aos testes em si mesmo, tipicamente o utilizador executa um série de passos que constituem as tarefas, enquanto os observadores tiram notas do comportamento do utilizador e fazem perguntas ou fornecem informação se necessário. Entretanto podemos ter algum tipo de mecanismo de *logging* que permita analisar a sessão *à posteriori* e gerar dados para as análises estatísticas.

No final das sessões é uma boa altura para conversar com o utilizador sobre eventos que ocorreram ao longo da sessão, para saber a sua opinião ou o que ele estava a pensar em determinada altura. É também nesta ocasião que se deve agradecer ao utilizador e demonstrar o quão importante foi a sua participação, sendo habitual fazer qualquer tipo de oferta como forma de agradecimento.

O trabalho final dos teste de usabilidade consiste em analisar os dados obtidos através dos questionários, dos *logs* do *software* e das observações realizadas ou até dos próprios comentários dos utilizadores. Podemos ter métricas qualitativas ou quantitativas.

Um dos problemas com as avaliações é que podem ser extremamente caras, quer ao nível de tempo quer monetariamente. As tarefas de recrutamento de utilizadores para realizar tarefas com as aplicações, preparação/criação do laboratório e análise estatísticas são particularmente caras. Os diálogos têm de ser transcritos e é necessário fazer observações e anotações. Naturalmente, existe interesse em automatizar e normalizar onde possível, ou utilizar técnicas preditivas, para avaliar e comparar diferentes opções de desenho. Por exemplo, a *framework* PARADISE (*PARAdigm for Dialogue System Evaluation*) [Dahlbäck 97] procura combinar uma série de métricas de desempenho, como taxa média de reconhecimento e taxa de conclusão de tarefas, numa única função de avaliação da desempenho na tentativa de prever a satisfação do utilizador, recorrendo para tal a métodos da teoria de decisão [Hansson 94] para especificar o peso relativo de cada factor. Para aplicar a teoria de decisão é necessário especificar os objectivos do problema de decisão, estabelecer critérios de decisão e formular um modelo ou relações entre objectivos.

O modelo do PARADISE é baseado numa estrutura de objectivos conforme a figura seguinte. O racional é que o desempenho pode ser correlacionado com um critério “externo” como a usabilidade e, assim, o objectivo da aplicação de fala é maximizar um objectivo relacionado com a usabilidade - é comum utilizarem-se indicadores de “Satisfação”, até porque os conceitos de eficácia e eficiência, conforme referidos na norma ISO 9241-11 (que é parte da norma ISO, *International Organization for Standardization*, 9241 - *Ergonomics of Human System Interaction*), são menos subjectivos.

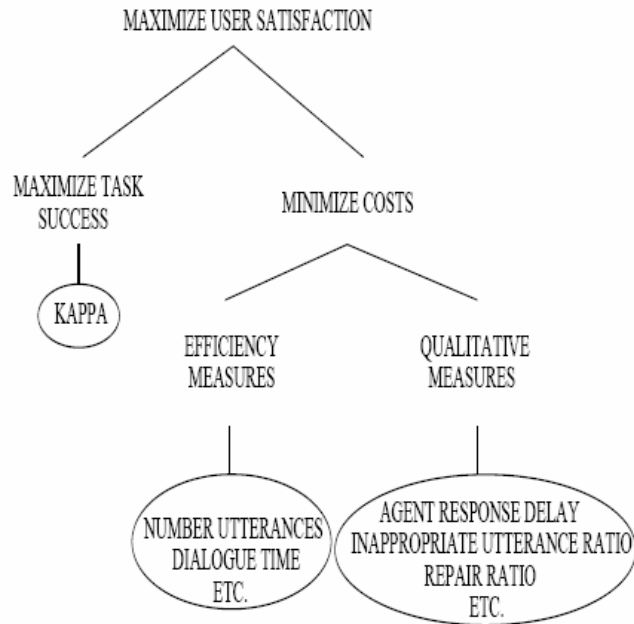


Figura 18 – Estrutura de objectivos no PARADISE (Fonte: [Walker 97])

No modelo da figura acima podemos observar dois factores como relevantes para a “Satisfação” (*User Satisfaction*), nomeadamente taxa de conclusão de tarefas (*Task Success*) e custo do diálogo (*Costs*), sendo que para este último contribuem as medias de eficiência (*Efficiency Measures*) e de qualidade (*Qualitative Measures*). Muito utilizado, este modelo facilita a comparação entre aplicações de fala, mas a generalização é limitada devido ao facto de diferentes soluções de arquitectura influenciarem alguns factores no modelo estatístico [Kamm 99] – por exemplo, o *barge in* parece ser importante em termos de usabilidade e no entanto não é suportado por muitos sistemas. Por outro lado, a utilização do PARADISE “obriga” a um processo intensivo e complexo de recolha de dados.

A referida norma ISO 9241 fornece requisitos e recomendações relacionados com os atributos do *software*, equipamento e ambiente que contribuem para a usabilidade e os princípios ergonómicos subjacentes a esses atributos. Em particular, a norma ISO 9241-11, define usabilidade e descreve a informação que é necessária considerar na especificação ou avaliação de um produto em termos de desempenho e satisfação do utilizador. É dada orientação sob a forma de princípios e técnicas gerais sobre como descrever explicitamente o contexto de utilização e as medidas relevantes de usabilidade. A orientação inclui procedimentos para avaliação de usabilidade, mas não detalha todas as actividades a serem realizadas nem cobre os processos de desenvolvimento dos sistemas.

6 Experiências realizadas

6.1 O corpora no projecto FASiL

6.1.1 O projecto

Designado pelo acrónimo FASiL (*Flexible and Adaptive Spoken Language and Multi-modal Interface*), este projecto pretendia desenvolver a tecnologia de linguagem humana mais avançada até à data, com um motor de reconhecimento e síntese de voz totalmente conversacional e adaptativo, numa altura em que esta área constitui uma das mais promissoras oportunidades de negócio nas chamadas novas tecnologias. Para o efeito, a União Europeia atribuiu-lhe o maior financiamento de sempre a um projecto de investigação tecnológica. O consórcio do projecto, do qual fez parte a Portugal Telecom Inovação, SA, foi liderado pelos ingleses da Vox Generation e o projecto teve a duração de dois anos e um custo global aproximado de investigação aplicada que rondou os 10 milhões de euros. O projecto foi subsidiado através do programa *Information Society Technologies (IST)* – (Tecnologias para a Sociedade de Informação).

A empresa Vox Generation foi responsável pela ideia do projecto e pela constituição de um consórcio de organizações mundiais dotadas de competências complementares. A Portugal Telecom Inovação apoiou a Vox Generation na gestão do projecto, sendo os restantes membros do consórcio a Cap Gemini Ernst & Young Sweden, SpeechWorks UK (adquirida no decorrer do projecto pela ScanSoft), Media Lab Europe, The University of Sheffield, The Royal National Institute for Deaf People e The Royal National Institute of the Blind. Para além do papel administrativo, a Portugal Telecom Inovação apoiou igualmente o desenvolvimento da componente linguística em português europeu, como as gramáticas, recolha de corpora, adaptação e instalação do sistema e, ainda, na condução de testes de usabilidade.

Ao contrário de muitos sistemas de fala comercializados actualmente, onde o utilizador é obrigado a adaptar-se à forma como o sistema funciona, a tecnologia que se procurou desenvolver no âmbito do FASiL permitiria ao utilizador conduzir a interacção com o sistema de processamento de fala. O sistema de fala procurava compreender e responder a questões levantadas pelos utilizadores e fornecer informação em inglês, português europeu e sueco. Os utilizadores podiam também interagir com o sistema através de textos e imagens, seleccionando o modo mais eficaz de aceder à informação e aos serviços.

6.1.2 Processo de criação dos corpora

Apenas a título de exemplo para se ter uma ideia completa do processo de criação de um *corpus*, no âmbito do projecto FASiL [Gomes 05] foram criados dois corpora, recorrendo à técnica WOz: um unimodal de voz e outro multimodal. O *corpus* unimodal contém diálogo natural e espontâneo entre o utilizador e o sistema (fictício) no domínio da gestão de informação pessoal (correio electrónico, contactos e agenda), com cerca de 12500 frases do utilizador, a que correspondem mais de 100000 palavras para 70 utilizadores. A duração média de cada sessão de gravação foi de 20 minutos, os quais, somando os tempos de recepção do utilizador, explicação da sessão, fecho da mesma e conversa final com o utilizador, ascendem a cerca de 45 minutos por utilizador. Já aqui temos uma medida do custo do processo, pois não era possível a gravação de mais de 3 a 4 sessões por dia. Ou seja, só para a fase de gravações foi necessário um mês para os 70 utilizadores.

A recolha do *corpus* unimodal do FASiL foi realizada com 5 “feiticeiros” e 70 sujeitos que representavam uma amostra dos potenciais utilizadores do sistema: 55% do sexo masculino, 45% do sexo feminino, idades compreendidas entre 18 e 57 anos, de diversas regiões do país e 97% deles usavam frequentemente o *e-mail*.

As interacções com o sistema foram feitas com telefones fixos, estando os utilizadores e os feiticeiros em salas separadas com níveis de ruído controlados. As gravações foram feitas recorrendo a um IVR com a possibilidade de gravar as chamadas com separação de canais e capacidades de *logging*. A figura seguinte ilustra a configuração técnica da plataforma, constituída por 3 camadas, utilizada na experiência WOz. Na “camada” do utilizador, a interface era dada apenas pelo telefone, que podia ser fixo ou móvel. Ao meio temos o IVR como mediador, com as funcionalidades de gravação (*REC*), a síntese de fala (*TTS*) e a capacidade de *logging* (*Logs*). Do lado do “feiticeiro”, temos a informação pessoal do utilizador (*PIM*), que é acedida através de uma aplicação de gestão como o Microsoft Outlook, a gestão da interface de voz (*Voice Manager*) e, como produto do processo, a base de dados de voz (*corpora*).

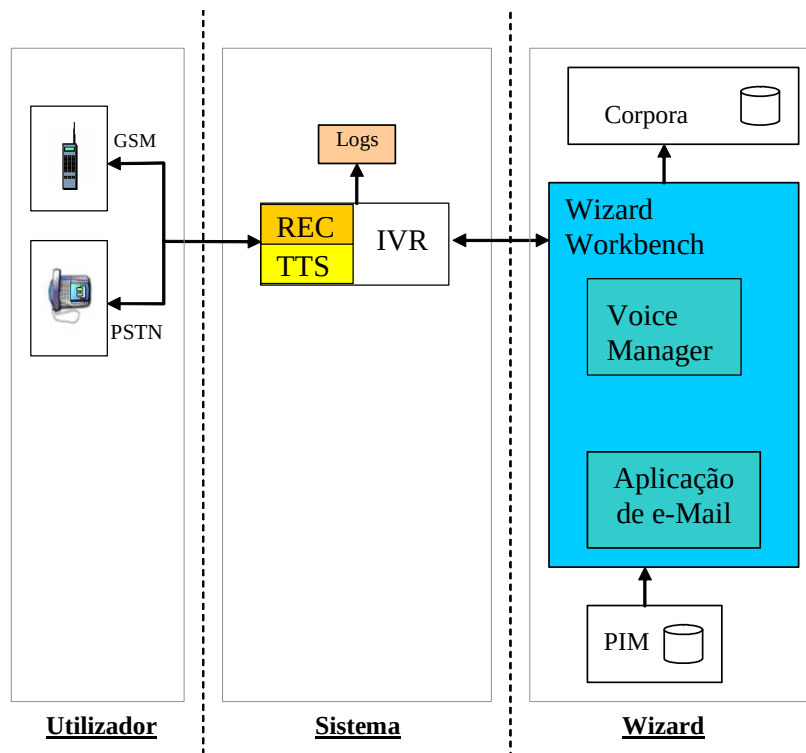


Figura 19 - Plataforma WOz

O resultado de cada sessão de gravação foram dois ficheiros de áudio 8kHz 8bit, um contendo a sessão completa do utilizador e outro com a sessão completa do “sistema”. Todos os ficheiros foram segmentados frase-a-frase usando o “CoolEdit 2000” e depois, recorrendo a uma ferramenta desenvolvida à medida, cada segmento foi transcrito ortograficamente de acordo com regras predefinidas.

Seguidamente procedeu-se ao armazenamento do *corpus* de acordo com uma estrutura acordada: um directório raiz e, no seu interior, um sub directório para cada sessão. Dentro do directório de cada sessão, todos os ficheiros de áudio, transcrições dessa sessão foram colocados juntamente com um ficheiro com a informação que identifica a sessão (*header file*). O *header file* continha informação como o título do projecto, a localização, a língua, a data, o identificador do orador, o identificador do feiticeiro e a data.

No directório raiz de cada CD foi colocado um ficheiro com uma tabela de conteúdos do mesmo, como mostra o exemplo seguinte.

Corpus_TOC.txt
\001\001_0000.txt \001\001_0001.txt \001\001_0001.wav \001\001_0002.txt \001\001_0002.wav \002\002_0000.txt \002\002_0001.txt \002\002_0001.wav \002\002_0002.txt \002\002_0002.wav \003\003_0000.txt (...)

Tabela 1 - Tabela de conteúdos do corpus

Na figura acima podemos observar que a tabela de conteúdos indica claramente quais as sessões (sub directórios) que estão naquele CD/DVD (“\001\”, “\002\”, “\003\” e assim sucessivamente), e os ficheiros que cada sessão contém: o ficheiro <...>_0000.txt é o *header file* e os outros ficheiros representam as frase ordenadas temporalmente sendo o “.wav” o áudio e o “.txt” com o mesmo sufixo a transcrição correspondente.

Após a gravação dos CDs/DVDs finais o *corpus* foi publicado, estando disponível para aquisição através da ELRA.

6.1.3 Lições aprendidas

Da experiência pessoal do autor aquando do seu envolvimento no processo de criação de corpora de voz, resultaram algumas constatações como o facto deste processo ser extremamente moroso e complexo, envolvendo equipas multidisciplinares, participação de utilizadores (e aspectos legais associados) e, por vezes, equipamento caro. Este processo obriga a estudos prévios do tipo de potenciais utilizadores, à definição de amostras representativas dos mesmos, à especificação rigorosa do tipo de dados a recolher, à definição e execução de uma metodologia de recolha e obriga ainda à definição da estrutura final para distribuição e reutilização do corpora. Em determinadas situações, as recolhas de corpora requerem o desenvolvimento de ferramentas próprias e específicas para algumas tarefas, acarretando assim, naturalmente, maiores custos e enfatizando a necessidade de equipas multidisciplinares. As amostras de voz obtidas podem ser complementadas pelas respectivas transcrições e anotações a vários níveis, sendo que para tal é importante a contribuição, ou presença na equipa, de um especialista em linguística. Dependendo dos requisitos, a criação de um *corpus* poderá inclusivamente demorar 1 ou 2 anos (ou até mais).

No decurso da recolha de corpora, o autor participou também numa experiência Feiticeiro de Oz (ver secção 5.5.4 do presente documento), tendo não só contribuído para a especificação da mesma e definição do ambiente necessário, como também feito parte da equipa de “feiticeiros”. A experiência Feiticeiro de Oz revelou-se muito eficaz na captura de discurso espontâneo dos utilizadores, apenas com o senão de ter introduzido um custo adicional não negligenciável, pois foi necessário preparar protocolos e treino completo para os feiticeiros para garantir consistência entre os mesmos em termos de atitude e tempos de resposta durante a experiência.

Apesar do custo, os corpora são de elevada importância, como aliás fica patente nas mais variadas aplicações dos mesmos no campo da pesquisa em linguística e das tecnologias

de fala (síntese e reconhecimento), principalmente se contiver discurso natural e espontâneo. De certa forma, um *corpus* encapsula diferentes perfis de utilização (*usage profiles*).

É pois importante promover a reutilização do corpora e das ferramentas criadas, ou adquiridas, para a captura do mesmo.

6.2 Teste baseado em cenários: o caso do VPA FASiL

Nesta secção irá ser documentado sucintamente um caso real de teste baseado em cenários numa aplicação de voz, realizado pelo autor no âmbito das suas funções na empresa onde desenvolve o seu trabalho.

O objectivo principal era testar o primeiro piloto de um serviço de voz de gestão de informação pessoal – um assistente pessoal virtual (VPA – *Virtual Personal Assistant*), desenvolvido no âmbito do projecto europeu FASiL (ver secção 6.1.1 do presente documento). Dado não estarmos perante a versão final do sistema, interessava também analisar o seu grau de desenvolvimento e conformidade com os objectivos iniciais e especificações funcionais e não funcionais.

A metodologia de trabalho seguida passou pela análise e selecção da documentação disponível, discussão com os restantes parceiros para seleccionar as funcionalidades a testar, preparação do ambiente de testes, execução dos teste e finalmente análise dos resultados para a obtenção de algumas conclusões.

6.2.1 O Assistente Pessoal Virtual

O projecto FASiL produziu um assistente pessoal virtual, de iniciativa mista (utilizador e sistema) com suporte para linguagem natural. Este assistente (VPA – *Virtual Personal Assistant*) permitia uma interacção conversacional, incorporando entrada e saída multimodal num dispositivo móvel (nomeadamente dispositivos 3G), permitindo o acesso e gestão ao correio electrónico, contactos e calendário dos utilizadores.

Para chegar a todos os cidadãos, o VPA era também acessível por cidadãos com dificuldades auditivas e visuais. A investigação centrou-se no desenvolvimento de modelos de língua e gestão sofisticada do diálogo que se adapta à experiência e estilo do utilizador, de forma a obter uma interacção mais eficiente e robusta. As línguas suportadas foram Inglês, Português e Sueco.

No planeamento do projecto estavam previstas duas versões experimentais, antes da versão final:

- VPA 1 – Assistente pessoal virtual com suporte apenas para interacção de voz e funcionalidades limitadas;
- VPA 2 – Assistente pessoal virtual com suporte para interacção multimodal, grande parte das funcionalidades implementadas, reduzido suporte para conversação adaptativa e integração com avatar para cidadãos com dificuldades auditivas.

Aquando dos testes reportados no presente documento a única versão disponível era a VPA 1.

Foi criada uma ficha do produto a testar que sintetizasse a primeira fase do processo de teste que foi a análise do produto. De seguida, reproduz-se essa ficha para um mais fácil acompanhamento das restantes subsecções.

Nome	VPA 1
Descrição resumida	Assistente pessoal virtual para acesso e gestão do correio electrónico, com suporte limitado, nesta versão, para calendário e lista de contactos. A interface com o utilizador é exclusivamente por voz.
Tecnologias	Automatic Speech Recognition (ASR), Text-To-Speech (TTS), VoiceXML, grXML, .NET, J2EE, Java, JavaScript, Servlets, ASP.NET, C, CORBA, Perl
Origem	Implementação pela Vox Generation, módulos de sumarização e adaptação pela Universidade de Sheffield, gestão do diálogo multimodal (ainda não disponível nesta versão) pela MLE, avatar (não disponível nesta versão) pela RNID, corpora linguístico para treino, gramáticas, <i>prompts</i> e adaptação pelos restantes parceiros
Documentação disponível	Especificação funcional, especificação da interface de envio de correio electrónico, especificação da interface de pesquisa de correio electrónico
Código fonte disponível	O código fonte não disponível
Ambiente de testes disponível	O sistema instalado em Inglaterra, pelo que os testes foram efectuados através de chamadas internacionais. Foram criadas caixa de correio, calendário e lista de contactos à medida.
Ferramentas de teste disponíveis?	Não
Dimensão do produto	Como o sistema estava instalado em Inglaterra foi difícil apontar a dimensão com precisão, mas eram necessários 3 CDs de instalação: um para o servidor Linux com o <i>browser</i> de voz e os motores de reconhecimento e síntese de voz, um com componentes para o servidor Windows 2000 e outro com o componente principal da aplicação e alguns módulos opcionais, sendo que o componente principal chegava quase aos 200Mb.
Estado do produto	O VPA 1 foi o primeiro protótipo do sistema VPA. Estavam implementadas apenas as funcionalidades relacionadas com o correio electrónico, identificação e validação dos utilizadores. As funcionalidades relacionadas com o calendário e lista de contactos estavam a começar a ser implementadas.
Nível a que é possível/interessa testar	Ao nível da interface de voz, testar as funcionalidades relacionadas com o correio electrónico e avaliar o estado de desenvolvimento do calendário, lista de contactos, gestão do diálogo e identificação/validação dos utilizadores.

Tabela 2 - Ficha do produto

Após a análise do produto, o passo seguinte consistiu na selecção e revisão da documentação disponível e relevante para os testes.

6.2.2 Revisão da documentação

Aquando da realização dos testes, e devido a contratos existentes entre a Comissão Europeia e o parceiro responsável pelas tarefas de desenvolvimento do código, existia pouca documentação relacionada com a implementação do sistema testado. Era necessário, portanto, tirar o melhor partido possível da documentação existente para conseguir testar o sistema, fazendo uma selecção criteriosa da mesma tendo em visto os objectivos pretendidos.

Os documentos disponíveis considerados relevantes para testar o VPA1 eram a sua especificação funcional e a especificação da interface de voz, constituída neste caso por dois documentos. Devido à acumulação de atrasos nas diversas etapas do projecto, as funcionalidades implementadas no VPA1 foram reduzidas a um mínimo aceitável. Optou-se por disponibilizar no VPA1 apenas as funcionalidades relacionadas com o acesso e gestão do correio electrónico por voz. Assim sendo, os documentos de especificação da interface de voz disponíveis eram a especificação da interface para a pesquisa e para o envio de correio electrónico.

Foi, então, feita uma revisão dos documentos acima mencionados, numa tentativa de avaliar a sua correcção, cobertura, completude, clareza, consistência e existência de ajudas à compreensão dos mesmos. Para tal foi seguida uma lista (*checklist*) para cada um deles, representada na tabela seguinte. No final foi produzida uma tabela de problemas e sugestões para cada.

Ajudas à compreensão	
	Existe introdução?
	Existe tabela de conteúdos?
	O documento está bem subdividido em secções?
	Existe bibliografia/referências claras?
	A paginação está correcta?
Clareza e Consistência	
	A linguagem utilizada é clara?
	Os exemplos utilizados são claros, relevantes e correctos?
	Os diagramas e ilustrações utilizados são claros, relevantes e correctos?
	A terminologia é correcta e clara?
	Existe um glossário?
Correcção	
	Existem pontos incorrectos?
	Existem contradições?
	Existem ambiguidades?
Cobertura e Completude	
	Todos os itens importantes estão incluídos?
	Todos os itens irrelevantes foram omitidos?
	O nível de detalhe técnico é adequado?
	Existe a indicação clara dos objectivos do documento?

Tabela 3 - Checklist para revisão da documentação

Outras características mais específicas de documentos de especificação como a Relevância, a Exequibilidade (*Feasibility*), Independência do código (isto é, se depende de uma arquitectura/tecnologia específica) e a Testabilidade foram revistos pelos parceiros do consórcio ao longo das várias versões. Numa posterior leitura para este trabalho, não resultou nenhum defeito ou sugestão a apontar relativamente a estas características mais específicas. A fase seguinte consistiu na preparação, nomeadamente no que concerne à organização de documentação relevante, preparação da equipa, levantamento dos requisitos do ambiente de teste e preparação do mesmo.

6.2.3 Preparação

Foi decidido internamente na organização que a equipa de testes seria constituída por elementos que fizessem parte da equipa que estava a acompanhar o projecto FASiL. Um aspecto importante a recordar é que a PT Inovação não esteve envolvida em tarefas de programação do sistema.

A preparação consistiu em:

- seleccionar os documentos relevantes já apresentados e leitura atenta dos mesmos;
- trocar opiniões/experiências regularmente com as equipas de teste dos restantes parceiros do consórcio;
- tentar determinar *à priori* os prazos temporais para a execução das tarefas do processo de testes;
- tentar calendarizar as diferentes etapas do processo;
- obter informação sobre teste de *software*.

Quanto ao ambiente de testes, o primeiro problema na sua preparação foi a instalação do próprio sistema. Instalar um sistema “estado-da-arte” com esta dimensão não é tarefa fácil e requer, ainda, conhecimentos muito específicos: diversos servidores, placas telefónicas, sistemas operativos (Windows 2000, Linux), diversas tecnologias (.NET, Java, JavaBeans, Servlets, Perl, VXML VoiceBrowsers, motores de reconhecimento e síntese de voz), políticas de segurança, administração remota - a lista completa seria demasiado maçadora. Os problemas começaram no incompleto e fraco manual de instalação, passando pelos muitos *bugs* severos do próprio sistema e sucessivos atrasos que daí resultaram. Após terem sido gastas, literalmente, 5 a 6 semanas de trabalho, o nível de instabilidade do sistema instalado ainda não permitia a execução de testes. A solução encontrada foi utilizar o sistema instalado no ambiente laboratorial onde estava a ser desenvolvido, em Inglaterra, com a necessária adaptação para português. Todos os testes ao sistema teriam de ser efectuadas através de chamadas telefónicas internacionais. Obviamente que esta situação acarretou custos extra para a organização, com impactos directos no tempo disponível para a realização dos testes – teriam de ser feitos gastando o menor tempo possível.

Para os testes, foi necessário criar uma caixa de correio fictícia, uma lista de contactos e um calendário coerentes com essa caixa.

O processo de criação da caixa de correio passou pela análise cuidadosa dos documentos já referidos, principalmente pelos de especificação da interface de envio e pesquisa, para determinar as suas características:

- mensagens de diferentes datas;
- mensagens novas e outras já lidas;
- diversas mensagens da mesma categoria (tema);
- diferentes números de mensagens de diversos remetentes (uma mensagem de um remetente, 6 mensagens de outro remetente);
- mensagens com diferentes níveis de prioridade (alta, baixa).

Quanto à criação da lista de contactos e do calendário foi mais pacífico, sendo apenas necessário criar a lista com alguns dos remetentes das mensagens da caixa de correio e outros fictícios, com diversos dados como telefone de casa, do emprego, morada, endereço de e-mail; no caso do calendário foi apenas necessário criar compromissos que já tinham ocorrido

e outros que estivessem ainda por ocorrer, com diferentes horas e dias da semana e ainda diferentes meses.

A caixa de correio, o calendário e a lista de contactos foram depois enviados para Inglaterra para poderem ser utilizados com o sistema a testar.

Para efectuar os testes foi utilizado um telefone fixo normal na sala de trabalho. Poder-se-ia ter igualmente usado um telemóvel e, por exemplo, executar os testes num ambiente mais ruidoso. No entanto, esta opção para além de mais cara, também poderia ter influência negativa na taxa de reconhecimento que, nesta fase do projecto, não era um objectivo fundamental de teste.

O passo seguinte consistiu na determinação dos casos de teste, bem como o processo necessário para a sua definição e dos *scripts* de teste.

6.2.4 Casos de teste

A definição dos casos de teste foi, sem dúvida, a mais difícil e morosa etapa de todo o processo.

Para a definição dos casos de teste recorreu-se à documentação referida anteriormente e a algumas das gramáticas e das *prompts* por nós traduzidas e adaptadas para Português. De uma forma simplista, uma gramática documenta o tipo de frases que um utilizador dirá com maior probabilidade num determinado cenário. Assim existiam no sistema gramáticas para envio de *email*, para pesquisa de *email*, comandos de navegação, para pesquisa de contactos, pesquisa de compromissos, criação de compromissos, para identificação no sistema – só para mencionar as mais óbvias. Como estávamos perante um sistema que pretendia ser conversacional, cada gramática possuía partes dinâmicas (criadas em *run-time*) e a sua dimensão ascendia, em alguns casos, às dezenas de milhar de linhas. As gramáticas utilizadas estavam no formato grXML (*Grammar eXtensible Markup Language*).

As *prompts* são as respostas do sistema. São extractos de texto que o sistema sintetiza para produzir a saída (*output*) de voz. Estas *prompts* estavam em ficheiros *Excel* para cada funcionalidade do sistema. À semelhança das gramáticas, o facto do sistema tentar ser conversacional implicava um número astronómico de *prompts*, com milhares de linhas nas folhas de *Excel* para cada funcionalidade do sistema.

Passando agora ao processo de definição dos casos de teste propriamente dito, este foi constituído por três passos: determinar funcionalidades a testar, definição dos casos de teste e a criação dos *scripts* de teste. Analisemos cada um desses passos.

Determinar funcionalidades a testar

Para este primeiro passo o documento mais importante foi naturalmente a Especificação Funcional. Como foi referido anteriormente, este documento apresentava a especificação funcional e a especificação não funcional do sistema VPA1. Quanto aos requisitos não-funcionais não foram efectuados quaisquer testes (ficaram a cargo de outro parceiro). Quanto à especificação funcional, o documento apresentava diversos pacotes de casos de uso, como por exemplo:

- Pacote Identificação e Validação
 - o Identificar Utilizador (*Use Case 1*)
 - o Validar Utilizador (*UC 2*)
 - o Introdução (*UC 3*)
- Pacote Correio Electrónico
 - o Pesquisar mensagens (*UC 4*)

- o Ler mensagens (UC 5)
- o Compor e enviar mensagens (UC 6)
- o Reencaminhar mensagens (UC 7)
- o Responder a mensagens (UC 8)
- o Apagar mensagens (UC 9)
- o Resumir mensagem (UC 10)
- Pacote Personalização
 - o Definir preferências (UC 11)
- Pacote Administração
 - o Gerir utilizadores (UC 12)
- Pacote Gestão do Diálogo
 - o (...)
- Pacote Calendário
 - o (...)
- Pacote Lista de contactos
 - o (...)

Como já foi referido, optou-se por disponibilizar no sistema a testar (VPA1) apenas as funcionalidades relacionadas com o acesso e gestão do correio electrónico por voz, centrando a implementação das funcionalidades relacionadas com o acesso e gestão da lista de contactos e do calendário para o protótipo VPA 2, embora também já existisse um pequeno esforço de implementação dessas funcionalidades no VPA1. Tendo isto em conta e as limitações temporais, os pacotes seleccionados para testes mais “rigorosos” foram o pacote de Identificação e Validação e o pacote de Correio Electrónico. Todos os casos de uso destes pacotes foram alvo de testes. Do Pacote Gestão do Diálogo, por exemplo, foram seleccionados alguns casos de uso considerados importantes: UC13 e UC17.

A seguinte tabela apresenta um resumo das funcionalidades seleccionadas para testar:

Pacote	Caso de Uso	Descrição
Identificação e Validação	UC1	Identificar utilizador
	UC2	Validar utilizador
	UC3	Introdução
Correio electrónico	UC4	Pesquisar <i>email</i>
	UC5	Ler <i>email</i>
	UC6	Compor e enviar <i>email</i>
	UC7	Reencaminhar <i>email</i>
	UC8	Responder <i>email</i>
	UC9	Apagar <i>email</i>
	UC10	Sumariar <i>email</i>
Personalização		
Administração		
Gestão do Diálogo	UC13	Alternar entre serviços
	UC17	Fornecer ajuda
Agenda	UC20	Aceder compromisso
	UC21	Pesquisar compromisso
	UC22	Ouvir detalhes do compromisso
	UC23	Apagar compromisso

Contactos	UC24	Mover compromisso
	UC25	Adicionar compromisso
	UC26	Aceder contactos
	UC27	Ouvir detalhes do contacto

Tabela 4 - Tabela das Funcionalidades Seleccionadas

Definição dos casos de teste

Depois de seleccionadas as funcionalidades, o passo seguinte foi definir os casos de teste para testá-las. Para tal, começou-se por recorrer mais uma vez ao documento de especificação funcional. Neste documento, para cada caso de uso são explicados o fluxo normal, os eventuais fluxos alternativos e ainda informação relacionada. Este tipo de dados foi explorado para a definição dos casos de teste. Por exemplo para o caso de uso 4 (UC 4), Pesquisar *email*, temos que o fluxo normal é o utilizador especificar os critérios de pesquisa que podem ser, segundo a informação relacionada, por nome (do remetente, que tem que existir na lista de contactos), data, prioridade ou por lido/não lido, depois, seguindo o fluxo normal, avança para o caso de uso 5 (UC 5), que é a leitura de mensagens. O caso de uso de 5 diz-nos que podemos ter fluxos alternativos conforme o número de mensagens retornados pela pesquisa efectuada: 0 mensagens, 1 mensagem, número médio de mensagens ou número grande de mensagens. Do que foi dito atrás retiram-se quase de imediato alguns casos de teste para a pesquisa de mensagens:

- Pesquisa por mensagens de um destinatário específico;
- Pesquisa por mensagens de um destinatário que não exista na lista de contactos;
- Pesquisa por categoria;
- Pesquisa por mensagens com uma data específica;
- Pesquisa por mensagens recebidas num determinado intervalo de datas;
- Pesquisa por mensagens não lidas;
- Pesquisa por mensagens de prioridades alta ou baixa;
- Pesquisa por mensagens combinando critérios (por exemplo, pesquisa por mensagens de prioridade alta recebidas numa data específica);
- Pesquisa que não retorna mensagens;
- Pesquisa que retorna um número médio de mensagens (entre 5 a 10);
- Pesquisa que retorna um grande número de mensagens (mais de 10).

A lista poderia continuar. E estamos só a falar de um/dois casos de uso... o número de casos de teste possíveis tende a ser enorme, mesmo para um subconjunto de todas as funcionalidades disponíveis.

Outra fonte de informação para a criação dos casos de teste provém da experiência pessoal com aplicações de correio electrónico – afinal o sistema a testar (VPA 1) é isso mesmo com uma interface de voz. Assim os casos de teste podem ser criados reflectindo sobre a forma como utilizamos este tipo de aplicações, por exemplo depois de lermos uma mensagem poderemos querer responder ou reencaminhar essa mensagem:

- Responder a mensagens;
- Reencaminhar mensagens;
- Apagar mensagem;

- Criar e enviar mensagens;
- Criar e enviar mensagens com prioridade alta;
- Criar e enviar mensagens para vários destinatários;
- Criar e enviar mensagens com conhecimento para vários destinatários;
- Criar e enviar mensagens com cópia confidencial para vários destinatários;
- Conjuguar vários comandos, como criar e enviar mensagens para vários destinatários e cópia confidencial para um destinatário específico.

Mais uma vez a lista poderia continuar: o número de casos de teste possíveis tende a ser enorme, mesmo para um subconjunto de todas as funcionalidades disponíveis. Nesta altura, e pelo que foi exposto atrás, foi também possível observar que as funcionalidades do pacote de correio electrónico podem ser agrupadas segundo dois grandes grupos de casos de teste – grupo de envio de mensagens e grupo de pesquisa de mensagens.

Outra técnica utilizada para determinar casos de teste foram os testes exploratórios [Bach 03], embora tenha sido mais utilizada na criação das *scripts* de teste para os pacotes de casos de uso do Calendário e Contactos. Nesta técnica, os casos de teste e as *scripts* de teste podem ser criados através da exploração do sistema, usando-o efectivamente apenas para navegar na sua interface ou durante a execução de outros testes, obrigando neste caso a uma atitude mais pró-activa durante a execução dos mesmos. Por exemplo, ao explorar sem “compromissos” o sistema foram detectados potenciais problemas ao tentar aceder a um contacto inexistente, tendo sido de imediato criado um caso de teste correspondente.

Havia ainda a necessidade de reduzir o número de casos pois, como já foi referido atrás, o número de potenciais casos de teste era enorme e não é possível/viável testar tudo. O primeiro passo foi dividir os casos de teste relacionados com o pacote de casos de uso de correio electrónico em dois grandes grupos – enviar e pesquisar. Depois tentou-se seleccionar os casos de teste que tivessem a maior probabilidade de revelar defeitos, em especial os defeitos que tivessem maior impacto (ainda para mais tratando-se de um protótipo intermédio) e em particular defeitos que ocorrem nas funcionalidades mais usadas do sistema – daí a aposta no pacote de correio electrónico. Para reduzir o número de casos de teste também se aplicou, onde possível, o raciocínio por trás da técnica de divisão em classes de equivalência, por exemplo: no caso do envio de mensagens para múltiplos destinatários, testar para dois destinatários é o mesmo que testar para três, logo para testar múltiplos destinatários bastará usar dois. Outra técnica aplicada foi a conjugação de vários comandos, como já se viu atrás: em vez de termos um caso de teste para enviar uma mensagem para múltiplos contactos e outro caso de teste para mandar uma mensagem com prioridade alta, a ideia é juntar os dois casos de teste num só, mandar uma mensagem para múltiplos contactos com prioridade alta.

De seguida são apresentados exemplos das tabelas contendo os casos de teste criados para testar o sistema VPA1. É possível observar não apresentam a forma canónica “Entrada> Saída Esperada” pois desde a entrada à saída existe um conjunto variável de passos que podem falhar antes sequer de haver qualquer saída. Por exemplo, para o caso de teste de envio de mensagem a saída esperada seria uma nova mensagem na caixa de correio do destinatário, mas no sistema esta operação envolve a confirmação implícita da operação a realizar, a confirmação do destinatário, ditar e confirmar o fim de mensagem, sendo que nestes passos o sistema pode falhar e interessa saber em que passo falha. Tratando-se de um sistema conversacional não faz sentido afirmar que, no caso de apagar uma mensagem, a entrada é “utilizador diz: ‘apagar mensagem’” e a saída é “mensagem apagada”, porque o utilizador não pode chegar ao sistema e dizer “apagar mensagem”: tem de efectuar uma pesquisa pela mensagem que quer apagar, o sistema tem de reconhecer o comando correctamente, o utilizador tem de pedir para apagar, o sistema depois pede confirmação... Para o teste ser

considerado bem sucedido todos estes passos têm de se ser bem sucedidos. Se o sistema apagar a mensagem, tal como previsto pela saída esperada do caso de teste, o teste pode ter falhado na mesma se as *prompts* estiveram erradas, por exemplo.

Nas tabelas seguintes, denominadas matrizes de rastreabilidade, cada caso de teste vai testando (passe a redundância) progressivamente mais funcionalidades que o anterior (o problema da sequenciação dos cenários não é trivial). A cada caso de teste corresponde uma *script* – explicadas na subsecção seguinte deste documento. Para cada caso de teste são apresentados os casos de uso cobertos por esse caso de teste.

Começando para o grupo de testes relacionados com o envio do correio electrónico.

Script	Caso de Teste	Casos de uso
1	Enviar email	UC6
2	Enviar email - múltiplos contactos	UC6
3	Enviar email - múltiplos contactos e cópia para um contacto	UC6
4	Enviar email - múltiplos contactos, CC & BCC para um contacto	UC6
5	Enviar email - múltiplos contactos, CC para um contacto + rever contactos	UC6
6	Enviar email - múltiplos contactos, BCC para um contacto + rever a mensagem e regravá-la	UC6
7	Enviar email - múltiplos contactos, definir como prioritário	UC6
8	Enviar email - múltiplos contactos, apagar um contacto	UC6
9	Enviar email - múltiplos contactos, apagar um contacto sem especificar o nome na primeira fase	UC6
10	Enviar email - múltiplos contactos, CC & BCC para um contacto, gravar mensagem, rever destinatários, rever mensagem	UC6
11	Enviar email - múltiplos contactos & CC para contactos, alterar contacto, rever tudo	UC6
12	Enviar email – múltiplos contactos & BCC para um contacto + rever a mensagem, regravá-la e cancelar	UC6
13	Enviar email - múltiplos contactos, gravar mensagem, remover um contacto que não tinha sido especificado	UC6

Tabela 5 - Matriz de Rastreabilidade – Casos de Teste para Envio de Correio Electrónico

A tabela seguinte lista os casos de teste para o grupo de funcionalidades de pesquisa de correio electrónico.

Script	Caso de Teste	Casos de Uso
1	Pesquisar mensagens com alta prioridade, navegar nos resultados e responder ao remetente	UC4,UC5,UC8
2	Nova pesquisa por <i>emails</i> de um remetente específico	UC4,UC5
3	Nova pesquisa por <i>emails</i> de um remetente inexistente	UC4,UC5
4	Nova pesquisa por <i>emails</i> de uma categoria específica	UC4,UC5
5	Nova pesquisa por <i>emails</i> de um remetente específico e refinar por prioridade	UC4,UC5
6	Nova pesquisa por emails recebidos num determinado intervalo de tempo, refinar e efectuar nova pesquisa	UC4,UC5
7	Nova pesquisa conjugando critérios: múltiplos remetentes, ouvir a mensagem e apagá-la	UC4,UC5,UC9
8	Nova pesquisa conjugando critérios: remetente e período de tempo	UC4,UC5
9	Pesquisar por mensagens de baixa prioridade, navegar nos resultados e pedir um resumo	UC4,UC5,UC10
10	Pesquisar por mensagens, refinar por mensagens novas e efectuar uma nova pesquisa	UC4,UC5
11	Nova pesquisa por novas mensagens (>5 resultados), navegar nos resultados, ouvir e apagar uma mensagem, voltar à anterior e efectuar nova pesquisa	UC4,UC5, UC9

12	Nova pesquisa por novas mensagens (>10 resultados), voltar ao anterior depois de ouvir a primeira mensagem	UC4,UC5
13	Nova pesquisa conjugando critérios: múltiplos remetentes, refinar por data, cancelar	UC4,UC5
14	Pesquisar por mensagens prioritárias, navegar nos resultados com diferentes tipos de conteúdo	UC4,UC5
15	Nova pesquisa por emails de um remetente específico, refinar por prioridade (resultado=0)	UC4,UC5
16	Pesquisar por mensagens prioritárias, navegar nos resultados e reencaminhá-la para um destinatário	UC4,UC5, UC7

Tabela 6 - Matriz de Rastreabilidade – Casos de Teste para Pesquisa de Correio Electrónico

Outras matrizes semelhantes foram criadas para as funcionalidades relacionadas com o calendário, com a lista de contactos, com a validação dos utilizadores e com a gestão do diálogo.

6.2.5 Criação das scripts de teste

O processo de criação das *scripts* (sequências) de teste foi bastante complicado. Para cada caso de teste identificado foi criada uma *script* com os passos necessários para atingir o objectivo: que comandos dizer e que resposta (*prompt*) esperar em cada passo. Mas esta tarefa foi bastante difícil pois não existia documentado um diagrama de estados do sistema que fornecesse a informação de qual, ou quais, as gramáticas e *prompts* activas em cada estado. Assim, por exemplo, se o utilizador diz “pesquisar mensagem”, não se sabia qual a gramática activa nesse instante nem qual a *prompt* que iria ser tocada – teve-se de tentar deduzir e convém realçar que tínhamos diversas folhas de *prompts*, perfazendo um total de milhares de *prompts*, e diversas gramáticas gigantescas com dezenas de milhares de linhas. No caso das *prompts* essa missão foi auxiliada pelos documentos de “Especificação da Interface de Pesquisa de Correio Electrónico” e “Especificação da Interface de Envio de Correio Electrónico”. Estes documentos apresentavam os estados possíveis da interface de voz para as funcionalidades de pesquisa e envio de correio electrónico, respectivamente. Para cada estado eram listadas as pré-condições e as *prompts* possíveis. Para aquilo que o utilizador pode dizer, houve que procurar nas gramáticas...

Ilustra-se agora um exemplo de criação de uma *script* para o caso de teste 1 do envio de correio electrónico (ver Tabela 5).

O utilizador pretende enviar um email:

Foi necessário procurar nas gramáticas as secções com frases/comandos para esta acção. Esta acção correspondia a um estado no documento de especificação da interface de envio de correio electrónico. Neste estado, a *prompt* que ia ser tocada após o comando do utilizador era “Enviando uma mensagem. Diga não se estiver incorrecto”. Como o utilizador não especificou remetentes a interface avança para outro estado, cuja *prompt* é “Para quem gostaria de enviar?”. Como o utilizador não disse nada entre as duas *prompts* a sequência seria qualquer coisa como:

Utilizador: “Enviar mensagem”

Sistema: “Enviando uma mensagem. Diga não se estiver incorrecto.” <pausa> “Para quem gostaria de enviar?”

Neste momento o utilizador pode especificar os destinatários. O estado seguinte da interface indicava a *prompt* tocada: “Para <repete nomes dados pelo utilizador>. Diga não se estiver errado”. A interface “avança” então para um outro estado, cuja *prompt* era “Por favor grave a sua mensagem após o sinal, e pressione qualquer tecla para terminar”. A sequência aqui será:

Utilizador: “Para X e Y”

Sistema: “Para X e Y. Diga não se estiver errado.” <pausa> “Por favor grave a sua mensagem após o sinal, e pressione qualquer tecla para terminar”

A partir deste momento o utilizador grava a sua mensagem e carrega numa tecla para terminar. A interface transitava para novo estado, cuja *prompt* principal era “Está pronto para enviar esta mensagem agora?”. A sequência aqui seria:

Utilizador: <grava mensagem>

Sistema: “Está pronto para enviar esta mensagem agora?”

Agora o utilizador confirma o envio e a interface avança para o estado final: “A sua mensagem foi enviada.”

A tabela seguinte apresenta a *script* completa. O exemplo atrás indicado era para o caso de teste mais simples possível para o grupo de envio de mensagens (ver Tabela 5) ...

Tarefa:	Enviar mensagem					
Passo	Utilizador	Sistema	Resultado (OK/Not OK)	Gramática	Folha de <i>Prompts</i>	Estado
1	"Criar mensagem"	Enviando uma mensagem. Diga não se estiver incorrecto....Para quem gostaria de enviar?		send_email_fsm_parser_pt.xml	Fasil Prompts - Search E-mail_v2.xls	1a, 2ª
2	"[nome]"	"Para [nome]. Diga não se estiver errado... Por favor grave a sua mensagem após o sinal, e pressione qualquer tecla para terminar		send_email_fsm_parser_pt.xml	Fasil Prompts - Search E-mail_v2.xls	C2a, 8ª
3	gravar mensagem e carregar numa tecla para terminar	Está pronto para enviar esta mensagem agora?			Fasil Prompts - Search E-mail_v2.xls	10
4	"Sim"	A sua mensagem foi enviada.			Fasil Prompts - Search E-mail_v2.xls	11c

Tabela 7 - Script (Sequência) para Enviar Mensagem

Podemos observar da tabela as seguintes colunas:

- Utilizador: comandos de voz retirados das gramáticas e acções que o utilizador pode/deve executar;
- Sistema: saída de voz do sistema;
- Resultado: apenas uma pequena indicação se o passo teve sucesso (OK) ou não (Not OK). Este campo não pretende ser um completo registo de resultados, isso será registado em outras folhas de Excel que o capítulo seguinte explicará em mais detalhe;
- Gramática: indicação da gramática onde se encontram os comandos que o utilizador pode dizer no passo correspondente da *script*;
- Folha de *Prompts*: indicação do ficheiro onde se encontram as *prompts* do sistema desta *script*;
- Estado: estado da interface de voz indicado pelos documentos de especificação da interface de envio ou pesquisa de mensagens.

Para a criação das *scripts* das funcionalidades relacionadas com o Calendário, a Lista de Contactos, a Gestão de Diálogo e Identificação e Validação dos Utilizadores foi utilizada também a (já referida) técnica de testes exploratórios [BACH 03], visto não existirem documentos de especificação da interface, à semelhança do que acontecia para o envio e pesquisa de mensagens. Não esquecer que para estas funcionalidades o rigor exigido era também menor. Na técnica exploratória, navegamos pelo sistema para atingir um determinado objectivo, por exemplo procurar um compromisso, procuramos nas gramáticas aquilo que o utilizador pode dizer e, através das *prompts* que o sistema vai tocando, vai-se refinando a *script* de teste analisando as folhas de *prompts* e as gramáticas.

Para não tornar a leitura deste documento maçadora e não tornar o seu tamanho incomportável, as restantes *scripts* não serão reproduzidas.

6.2.6 Execução dos testes

A execução dos testes deveria ser realizada no menor intervalo de tempo possível devido, essencialmente, ao já referido facto de o sistema ter de ser acedido através de chamadas telefónicas internacionais.

Para executar os testes e registar os resultados foram criadas tabelas em Excel para cada grupo de funcionalidades: envio de mensagens, pesquisa de mensagens, calendário, lista de contactos, gestão do diálogo e identificação/validação dos utilizadores. As tabelas eram constituídas pelas seguintes colunas:

- *Script*: identificador numérico de qual o caso de teste em causa (visto que a cada caso de teste corresponde uma *script*);
- Finalidade: descrição breve do caso de teste;
- Casos de Uso: casos de uso testados pela *script*;
- ID Responsável: identificador numérico do responsável pelos testes;
- Data: data da realização dos testes;
- Hora: hora de realização dos testes;
- OK/Not OK: resultado do teste, passou não passou;

- Passo: no caso de ter falhado o teste, indica qual o passo da *script* que falhou;
- Entrada: comando de voz ou acção do utilizador no passo que falhou. Neste campo concretizava-se o preenchimento efectuado para os comandos de voz – por exemplo, na *script* está “enviar para <nome>”, por isso neste campo colocava-se o que realmente foi dito, como “enviar para Rui Gomes”;
- Saída Esperada: saída que está indicada na *script* no passo que falhou;
- Saída: a saída que realmente ocorreu no passo que falhou;
- Severidade: indicador do nível de severidade da falha (baixa, média ou alta);
- Reproduzível (S/N): indicava-se se a possibilidade ou não de reproduzir o problema detectado pelo teste;
- Observações: campo de escrita para colocar as ocorrências anormais e explicar o que se passou.

A tabela seguinte ilustra a folha de Excel para a gestão do diálogo.

Script	Finalidade	Casos de Uso	ID_Resp.	Data	Hora	(OK/NOT OK)	Passo	Entrada	Saída Esperada	Saída	Severidade	Reproduzível (S/N)	Observações
1	Diálogo – Ir para o nível de topo	UC13											
2	Diálogo – Ir para o Calendário	UC13											
3	Diálogo – Ir para a Lista de Contactos	UC13											
4	Diálogo – Ir para o Correio Electrónico	UC13											
5	Diálogo – Pedir ajuda no nível de topo	UC17											
6	Diálogo – Pedir ajuda no Correio Electrónico	UC17											
7	Diálogo – Pedir ajuda no Calendário	UC17											
8	Diálogo – Pedir ajuda na Lista de Contactos	UC17											

Tabela 8 - Tabela para registar os resultados dos testes

Depois de definidas estas tabelas Excel procedeu-se à realização dos testes. Como já foi referido o sistema estava instalado em Inglaterra, tendo os testes sido efectuados através de chamadas internacionais: ou seja, o responsável pelo teste realizava a chamada e percorria a *script*, anotando os diversos passos e resultados.

À semelhança das *scripts* de teste, para não tornar a leitura deste documento maçadora e não tornar o seu tamanho incomportável, as restantes folhas de resultados são omitidas.

6.2.7 Análise dos resultados obtidos a partir da execução dos testes

Após a execução dos testes procedeu-se à análise dos resultados obtidos. Como foi referido na secção anterior, foram criadas folhas Excel para apontar os resultados e ocorrências dos testes. A partir dessas folhas realizou-se então a análise. Os resultados foram analisados por grupo de funcionalidade através de pequenos relatórios de incidentes (*Test incident reports*) e, no final, foram analisados os resultados globais. O facto dos resultados estarem anotados nas folhas Excel apresentadas no capítulo anterior facilitou a sua transmissão (*report*) e compreensão pelos os restantes parceiros, em particular para os envolvidos nas tarefas de desenvolvimento.

Para cada do grupo de funcionalidades, foi calculado o tempo total dos testes e o tempo por *script*, o número e percentagem de falhas. Para cada *script* que falhou, foi apontada a razão da falha, a sua severidade e a possibilidade de reprodução. No final foram elaboradas considerações gerais e atribuída uma classificação numa escala de não-satisfaz, satisfaz, bom, muito bom.

Após analisar cada grupo de funcionalidades, foi realizada uma análise global dos resultados e custos do processo de testes encetado.

Em termos tempo gasto no processo temos de contabilizar o tempo dispendido nos vários passos do processo:

- análise da documentação - 3 dias (8x3=24 horas);
- tempo de preparação do ambiente de testes - 1,5 dias (12horas);
- tempo de criação das sequências (*scripts*) – 2 semanas (8*10=80horas). A criação dos *scripts* foi, como já referido, muito difícil e morosa. Para cada passo de cada *script* foi necessário analisar o caso de uso em causa, nomeadamente os fluxos normais e alternativos apresentados no documento de especificação funcional, depois determinar nas gramáticas o que o utilizador podia dizer (não esquecer que as gramáticas tinham dezenas de milhar de frases possíveis) e, por fim, localizar por dedução qual o estado da interface nos documentos de especificação da interface para poder determinar a saída (*prompt*) do sistema (não esquecer que existiam milhares de *prompts* possíveis). E isto para cada passo. Existem um total de cerca de 250 passos!
- o tempo de execução dos testes foi de cerca 7 horas, o que corresponde sensivelmente a 1 dia de trabalho. Este tempo de 7 horas tem de ser também contabilizado em termos de chamadas internacionais.
- o tempo de criação da documentação do processo de teste foi de cerca de 30 horas, o que inclui a análise dos resultados.

Portanto o tempo total do processo de testes foi de 153 horas, sendo que mais de 50%, 52% para ser mais preciso, deste tempo foi dispendido na criação das sequências (*scripts*) de teste.

Foram ainda efectuadas algumas estatísticas para análise, recorrendo, por exemplo a tabela com os resultados da execução das *scripts*. Por exemplo, para o envio de mensagens:

Script	Caso de Teste	Resultado
1	Enviar <i>email</i>	Ok
2	Enviar <i>email</i> - múltiplos contactos	Ok
3	Enviar <i>email</i> - múltiplos contactos e cópia para um contacto	Ok
4	Enviar <i>email</i> - múltiplos contactos, CC & BCC para um contacto	Not Ok
5	Enviar <i>email</i> - múltiplos contactos, CC para um contacto + rever contactos	Ok
6	Enviar <i>email</i> - múltiplos contactos, BCC para um contacto + rever a mensagem e regravá-la	Ok
7	Enviar <i>email</i> - múltiplos contactos, definir como prioritário	Ok
8	Enviar <i>email</i> - múltiplos contactos, apagar um contacto	Ok
9	Enviar <i>email</i> - múltiplos contactos, apagar um contacto sem especificar o nome na primeira fase	Ok
10	Enviar <i>email</i> - múltiplos contactos, CC & BCC para um contacto, gravar mensagem, rever destinatários, rever mensagem	Ok
11	Enviar <i>email</i> - múltiplos contactos & CC para contactos, alterar contacto, rever tudo	Not Ok
12	Enviar <i>email</i> - múltiplos contactos & BCC para um contacto + rever a mensagem, regravá-la e cancelar	Ok
13	Enviar <i>email</i> - múltiplos contactos, gravar mensagem, remover um contacto que não tinha sido especificado	Not Ok

Tabela 9 - Resultados do Envio de Mensagens

A nível global, foram executados 47 *scripts*, das quais 12 encontraram falhas do sistema, o que equivale a cerca de 25% das *scripts* encontrarem falhas.

Em termos de passos, foram criados 248 passos, o que dá uma média de 5,2 passos por *script*. Dos 248 passos, 212 passaram com sucesso e 12 falharam (os restantes 24 não chegaram a ser testados pois seguiam-se aos que falharam). Podemos extrapolar estes resultados para, estatisticamente, dizer que por cada 1000 passos, 50 falham (ou seja, 5% dos passos falharam). Em termos médios, tipicamente o passo que falha é o quarto.

Olhando para o tempo, podemos dizer que cada *script* demorou em média cerca de 9 minutos e cada passo 1 minuto e 42 segundos. Ora, sendo tipicamente o quarto passo a falhar, podemos dizer que 5 minutos era o tempo crítico ao fim do qual estatisticamente começavam a aparecer falhas (3 passos iniciais a 1 minuto e 42 segundos cada perfaz 5 minutos e 6 segundos para chegar ao passo que tipicamente falha).

Os problemas de reconhecimento foram a principal razão das falhas do sistema, mas também as *prompts* denotaram necessitar de atenção. O grupo de funcionalidades de envio e pesquisa de correio electrónico foram, sem surpresa, os que apresentam maior maturidade. As funcionalidades relacionadas com a identificação e validação dos utilizadores e navegação pela aplicação (parte da gestão do diálogo) também apresentaram uma boa robustez, servindo de base para futuros desenvolvimentos. Por outro lado, as funcionalidades da lista de contactos bem como do calendário encontravam-se ainda bastante atrasadas, bem como o reconhecimento de voz relacionado com essas áreas. Em termos de casos de uso específicos, foi impossível apontar um caso de uso como sendo mais problemático que outro.

6.2.8 Lições aprendidas

Os testes baseados em cenários efectuados tinham por finalidade aquilatar o grau de desenvolvimento do sistema, verificar a conformidade com os requisitos funcionais e não funcionais especificados para o primeiro piloto do sistema e, finalmente, verificar superficialmente a taxa de reconhecimento do motor de reconhecimento de voz utilizado.

Para atingir os objectivos propostos analisou-se a documentação disponível, foi preparado o ambiente de testes, definiram-se os casos de teste, procedeu-se à execução dos testes e finalmente foram analisados os resultados para a obtenção de algumas conclusões.

Numa primeira etapa, foi feita uma revisão dos documentos, numa tentativa de avaliar a sua correcção, cobertura, completude, clareza, consistência e existência de ajudas à compreensão dos mesmos. A quantidade de documentação disponível relacionada com o sistema era francamente limitada para a criação/execução dos testes. Uma lição a retirar para futuros processos de teste é a importância dos procedimentos ao nível da documentação, nomeadamente a criação de todos os documentos necessários, com qualidade ao nível da clareza, completude, consistência e correcção.

A fase seguinte consistiu na preparação, nomeadamente no que concerne à organização de documentação relevante, preparação da equipa, levantamento dos requisitos do ambiente de teste e preparação do mesmo. Quanto ao ambiente de testes propriamente dito, o primeiro problema na sua preparação foi a instalação do próprio sistema. Instalar um sistema “estado-da-arte” de grandes dimensões não foi tarefa fácil e o processo de testes foi sendo atrasado devido à instabilidade e problemas da sua instalação. É importante, pois, agilizar os processos de instalação dos sistemas a testar.

A definição dos casos de teste foi, sem dúvida, a mais difícil e morosa etapa de todo o processo. Foram seguidos três passos: determinar as funcionalidades a testar, definição dos casos de teste e a criação das *scripts* de teste. Para executar estes passos, recorreu-se à documentação disponível e a um conjunto de técnicas como os testes exploratórios [Bach 03], técnica de divisão em classes de equivalência e selecção dos casos de teste baseados na probabilidade de revelar defeitos (em especial os defeitos que tivessem maior impacto e defeitos que ocorrem nas funcionalidades mais usadas do sistema). O último passo, o processo de criação das *scripts* de teste, foi bastante complicado. Para cada caso de teste identificado foi criada uma *script* com os passos necessários para atingir o objectivo: que comandos dizer e que resposta esperar em cada passo – esta informação teve de ser deduzida a partir de milhares de *prompts* e diversas gramáticas com dezenas de milhares de linhas.

Para executar os testes e registar os resultados foram criadas tabelas em Excel para cada grupo de funcionalidades. A execução dos testes consistiu em realizar chamadas telefónicas para o sistema e “correr” os cenários.

Este processo foi extremamente importante para o projecto, permitindo de uma forma relativamente rápida, avaliar o grau de maturidade do protótipo e identificar as áreas problemáticas numa etapa inicial do sistema (não esquecer que se tratava do primeiro protótipo). As repercussões da descoberta destes problemas numa etapa mais adiantada num projecto europeu desta dimensão poderiam ser bastante graves. O teste baseado em cenários revelou-se eficaz e relativamente rápido. No entanto, a dimensão da “rapidez” parece ser sensível às condições de teste, como por exemplo, estar na posse da documentação adequada e de qualidade.

6.3 Avaliação de usabilidade - o caso do projecto FASiL

Nesta secção irá ser documentado sucintamente um caso de avaliação de usabilidade de uma aplicação de voz realizado pelo autor no âmbito do projecto FASiL (descrito na subsecção 6.2).

O objectivo fundamental da avaliação no âmbito do projecto FASiL foi a determinação do grau de usabilidade, acessibilidade e aceitabilidade do VPA. O processo de avaliação nas três línguas (Português, Inglês e Sueco) envolveu uma metodologia semelhante:

- Teste preliminar com três utilizadores independentes para determinar se o sistema cumpria os níveis mínimos de funcionalidade, desempenho e estabilidade previamente acordados pelo consórcio;
- Avaliação heurística efectuada por especialistas de usabilidade, que verificaram se o sistema cumpria as “regras de ouro” de Jakob Nielsen (<http://www.useit.com>), habitualmente utilizadas neste tipo de avaliações. Estas regras representam boas práticas e permitem avaliar aspectos como a simplicidade dos diálogos, o tipo de linguagem utilizada, a carga de memória necessária para interagir com o sistema e a prevenção de erros, entre outros;
- Teste de Utilizadores finais para avaliar a usabilidade e aceitabilidade do sistema.

Os testes específicos de acessibilidade foram realizados por utilizadores com necessidades especiais, quer na área da visão quer na área auditiva, mas apenas para o piloto inglês.

Os objectivos gerais traçados para a avaliação eram a recolha de informação sobre a usabilidade e acessibilidade do sistema nas 3 línguas, incluindo a interacção com o sistema, com a interface de voz, a interface gráfica e a interface multimodal e também a possibilidade de a interface fornecer funcionalidades equivalentes para um espectro alargado de utilizadores, incluindo pessoas com dificuldades visuais e auditivas. No entanto, não foi possível testar todas as modalidades com a extensão desejada devido, essencialmente, a limitações temporais. Estas limitações tiveram ainda um impacto negativo ao nível dos *bugs* na versão disponível para os testes com utilizadores finais, pois, praticamente, não houve muito tempo após a avaliação heurística para os corrigir.

A PT Inovação executou os testes de avaliação do piloto português. Tendo em conta a sua área de actividade, um objectivo particular da avaliação teve a ver com o estudo e a tentativa de compreender que factores afectam a experiência de utilização e de aceitação deste tipo de serviços multimodais. Em particular, havia interesse em conhecer melhor a influência da acuidade de reconhecimento / compreensão na apreciação do utilizador e também a utilização – como e quando – da multimodalidade [Oviatt 99].

6.3.1 A avaliação heurística

Este método, conforme descrito na secção 5.5.3 do presente documento, envolve a inspecção sistemática da interface do utilizador por um pequeno grupo de peritos que verificam a aderência da interface a princípios de usabilidade (as “heurísticas”). Nesta secção será descrita a experiência do autor na sua utilização no âmbito do projecto FASiL.

Participantes

A avaliação heurística foi realizada por cada um dos parceiros do projecto, tendo a PT Inovação participado com 3 peritos.

Desenho e materiais da avaliação

O sistema foi avaliado nas três línguas seguindo um processo iterativo. Foram acordados um plano de avaliação e um calendário para as várias etapas.

Foi utilizado um PC com ligação à Internet, colunas e um microfone, bem como equipamento de gravação para registar as sessões.

Método

Foi fornecido aos peritos um plano de testes, a lista de funcionalidades a explorar, uma ficha para reportar os resultados e a lista das heurísticas a utilizar. Os peritos inspeccionaram a interface e anotaram todos os aspectos relevantes com que se depararam na ficha de resultados. Esta ficha era constituída por quatro colunas: a primeira com a identificação da parte do sistema, a segunda para descrever brevemente o problema de usabilidade encontrado, a terceira coluna para identificar o princípio de usabilidade violado, e a última coluna para escrever recomendações para eliminar o problema. No final da avaliação heurística, os peritos encontravam-se e elaboravam as conclusões com base nos seus resultados individuais.

Resultados da avaliação heurística

A inspecção da interface pelos peritos revelou vários aspectos que, para facilitar a desmistagem, foram divididos em 3 secções: expectativas dos utilizadores (a), aspectos de usabilidade (b) e defeitos (c). Os defeitos e expectativas dos utilizadores foram resultados adicionais em relação aos aspectos de usabilidade, mas que foram tidos como relevantes dado o estado “embrionário” do protótipo e a escassez de tempo para o resto do projecto.

a) Expectativas dos utilizadores

Os peritos tentaram efectuar algumas tarefas para as quais o sistema não estava desenhado (por exemplo, procurar emails por assunto – que é uma funcionalidade permitida pelo Outlook). O sistema deveria disponibilizar não só as funcionalidades esperadas, mas estas deveriam também ser facilmente identificadas e acessíveis em termos de acções necessárias para o utilizador realizar uma tarefa. Daqui resultou a recomendação de estudar melhor o modelo mental do utilizador em futuras versões do sistema. Foram identificadas várias situações de comandos fora das gramáticas (isto é, daquilo que o sistema esperava) e estas situações mostraram alguma falta de robustez do sistema para lidar com elas.

Relativamente às expectativas dos utilizadores foram elaboradas uma lista com as funcionalidades em falta e outra lista com as funcionalidades que estariam mais distantes do modelo mental dos utilizadores e respectivas razões para isto.

b) Aspectos de usabilidade

Aqui os principais problemas encontrados advinham do facto de as *prompts* do sistema condicionarem o utilizador a adoptar um estilo de diálogo directo (comandos simples) em vez de conversacional, como se pretendia.

No final foi coligida uma lista dos principais problemas de usabilidade encontrados, comentários e recomendações.

c) Defeitos

Um dos principais problemas com que a equipa se confrontou foi o facto de a versão do sistema disponibilizada não era estável e apresentava comportamentos distintos entre as várias sessões de avaliação. Estes defeitos tinham ainda o efeito colateral nefasto de mascarar problemas de usabilidade... Por exemplo, o comando “Responder a todos” não funcionava (defeito), mas numa posterior versão do sistema “corrigida” detectou-se que o sistema não informava o utilizador que estava a responder a todos (problema de usabilidade).

Relativamente às recomendações finais dos peritos, estas foram divididas em dois grupos:

- recomendações principais;
- recomendações secundárias.

A lista de recomendações finais é extensa e fica fora do âmbito do presente documento e desta subsecção em particular, pelo que não há necessidade de a apresentar.

6.3.2 Avaliação do VPA2

A PT Inovação executou os testes de avaliação do piloto português. Era consensual no consórcio a ideia de tentar fazer uma avaliação de usabilidade com utilizadores finais o mais cedo possível. No entanto, o reduzido intervalo de tempo decorrido entre a avaliação heurística e a avaliação com utilizadores finais não permitiu efectuar grandes correcções/alterações à versão do sistema utilizada na referida avaliação heurística. Devido a este facto, foi acordado entre os parceiros não convidar utilizadores pouco experientes com este tipo de sistemas nem de fora das próprias organizações, pois eram as suas imagens que poderiam ser postas em causa... por outro lado, dada a instabilidade do sistema, convidar utilizadores a deslocarem-se propositadamente para assistir a bloqueios do sistema poderia levá-los a pensar que estariam a perder o seu tempo, embora as boas práticas recomendem que os utilizadores devam ser exteriores à empresa – alguns estudos referem que os utilizadores “internos” habitualmente são condescendentes na apreciação.

Participantes

Em função do exposto anteriormente, os parceiros do projecto procuraram recrutar utilizadores com alguma experiência em usar aplicações de e-mail e gestão pessoal, sendo o número mínimo 10 utilizadores. Em Portugal, os utilizadores foram seleccionados na PT Inovação: 4 mulheres e 8 homens com idades compreendidas entre os 19 e os 46 anos (média 30,6 anos; desvio-padrão 6,4 anos), sendo 75% licenciados. Três deles tinham experiência anterior em aplicações de voz. Nenhum dos utilizadores estava envolvido no projecto ou familiarizado com o sistema. Todos os 12 utilizadores eram utilizadores frequentes de aplicações de e-mail. Um deles tinha ligeiras dificuldades auditivas.

Consideramos que a amostra foi representativa do público-alvo: profissionais da área tecnológica, grau elevado de mobilidade e utilizadores experimentados do correio electrónico.

Preparação

Como parte da experiência, foi fornecido aos facilitadores/observadores um documento descrevendo o procedimento da experiência. Este documento indicava exactamente quais os passos que eles deveriam cumprir, o conteúdo do diálogo que deveriam ter com o utilizador em cada passo e o comportamento que deveriam seguir durante as sessões. Este documento permitiu “normalizar” os facilitadores/observadores ao longo de sessões distintas.

Para os utilizadores foi preparado um documento com informação acerca do sistema e um questionário para preencher antes do teste. Este questionário tinha a finalidade de recolher informação demográfica e o nível de experiência dos utilizadores com as tecnologias envolvidas.

Foi criado um pequeno conjunto de tarefas para o utilizador executar utilizando o protótipo desenvolvido. Estas tarefas pretendiam ser representativas do tipo de acções que os utilizadores finais iriam executar no sistema final e estavam ordenadas da mais fácil para a mais complicada, de forma a ir aumentando a confiança do utilizador com o sistema. Assim, por exemplo, era pedido aos utilizadores para navegar nos emails, encontrar emails específicos, ler, responder, entre outras. As questões colocadas ao utilizador para cada tarefa foram baseadas numa escala de 1 a 5 (escalas de Likert [Likert 32]).

Foi também criado um questionário pós-teste para tentar capturar as opiniões gerais dos utilizadores (o que acharam da facilidade de utilização, intuitividade e confiança no sistema).

Na avaliação portuguesa foi utilizado o seguinte equipamento:

- *Hardware*
 - o 1 pc
 - o 1 telefone fixo
 - o 1 microfone
 - o 1 IVR
- *Software*
 - o Internet Explorer 6.0
 - o SnagIt 5.2.2 (para captura de ecrãs)

O utilizador utilizava o PC para interagir com a interface gráfica e o telefone para interagir por voz com o sistema. Tanto o utilizador como um par de observadores estavam na mesma sala, sendo que os observadores se posicionaram atrás do utilizador para tirar notas da interacção.

Método

Cada sessão de avaliação estava programada para durar entre uma hora e meia a duas horas para cada utilizador. As sessões experimentais decorreram em ambiente laboratorial com os doze utilizadores já referidos, que interagiram com o VPA através de um conjunto de tarefas previamente definidas que simulavam a actividade usual de utilizadores de correio electrónico. Foi desenhada uma caixa do correio apropriada para a avaliação, preenchida com um conjunto de mensagens relacionadas com os cenários propostos.

Os testes foram efectuados individualmente, ou seja, apenas um utilizador de cada vez interagiu com o sistema. Cada teste foi acompanhado por dois elementos do grupo de trabalho do projecto, funcionando como facilitadores/observadores da experiência. Os observadores anotaram os tempos de execução, o tipo de interacção (voz ou gráfica), os problemas detectados e os comentários dos utilizadores.

Cada teste foi constituído, essencialmente, por 3 partes:

- **Introdução:** nesta fase de introdução era importante deixar os utilizadores descontraídos. O sistema e a experiência eram apresentados aos utilizadores e estes assinavam um termo de responsabilidade em como aceitavam que os seus dados e resultados do teste fossem usados para tratamento estatístico, sendo-lhes garantido o anonimato. Após a autorização, os utilizadores preenchiam um questionário pré-teste com questões pessoais (essencialmente);
- **Execução das tarefas:** foi pedido aos participantes que realizassem um conjunto de tarefas utilizando o sistema, quer a interface *web* quer a de voz. Enquanto o utilizador efectuava as tarefas, os observadores tiravam notas da experiência de interacção desse mesmo utilizador e dos seus comentários. Após completar cada tarefa os utilizadores manifestavam a sua opinião através de um questionário baseado numa escala de Likert – escala utilizada muitas vezes na forma de questionários para recolha de atitudes e que mede o grau de concordância /discordância com um conjunto de declarações não ambíguas. A escala utilizada ia de 1 a 5, em que 1 correspondia a “Muito satisfeito” a 5 “Muito Insatisfeito”;

- Pós-teste: após a conclusão de todas as tarefas, era pedido ao utilizador para preencher um questionário pós-teste. Após o questionário era realizada uma entrevista informal acerca das impressões gerais do sistema, bem como eventuais sugestões.

A plataforma INOVox (plataforma IVR desenvolvida pela PT Inovação) foi utilizada para gravar as intervenções orais dos utilizadores e as interacções gráficas gravadas em vídeo com o SnagIt 5.2.2. Uma ferramenta, desenvolvida pela VOX para tratamento dos *logs* do sistema permitiu a análise posterior da interacção de voz completa entre o sistema e o utilizador e também obter dados objectivos sobre o desempenho do sistema, no que se refere ao reconhecimento e compreensão das solicitações do utilizador.

Tarefas

As tarefas propostas, típicas do domínio do correio electrónico, foram descritas como cenários abertos de forma a permitir total autonomia e iniciativa aos utilizadores na sua concretização. Estas tarefas pretendiam ser representativas do tipo de acções que os utilizadores finais iriam executar no sistema final e estavam ordenadas da mais fácil para a mais complicada: desde a simples navegação pela caixa do correio até a pesquisas sucessivas de mensagens, permitindo criar, responder e reenviar mensagens e gerir a lista de destinatários.

Resultados

Foram realizadas um total de 55 sessões de avaliação da interface multimodal distribuídas pelos diversos parceiros da seguinte forma:

- 23 em Inglês;
- 20 em Sueco;
- 12 em Português.

Concretamente ao caso nacional, apesar do número reduzido de participantes, a experiência permitiu obter resultados que acreditamos serem interessantes.

Um dos aspectos mais salientes da observação da interacção dos utilizadores com o sistema foi a aproximação idêntica: comandos de voz para executar as tarefas e utilização da componente gráfica para ler as mensagens ou pesquisar mais rapidamente a lista de contactos e para ultrapassar problemas de reconhecimento ou a lentidão do sistema. A maioria dos utilizadores utilizou frases curtas e simples.

Terminada a sessão era pedido a cada utilizador a apreciação global do sistema e da sua experiência. Para além de questões abertas em que os utilizadores referiam os aspectos positivos e negativos do sistema, foram avaliados nove aspectos: intuitividade, facilidade de utilização, confiança no sistema, satisfação global, controlo sobre a interacção, interactividade em particular o grau de reconhecimento da fala, qualidade da voz sintetizada (intervenções do sistema e leitura das mensagens) e apreciação da qualidade do diálogo.

De seguida é apresentado um pequeno resumo dos resultados na avaliação e no questionário pós-teste. Para tal foi utilizada uma escala de 1 a 5, sendo 1 equivalente a muito satisfeito e 5 muito insatisfeito. A média foi arredondada para o inteiro mais próximo.

a) Intuitividade, facilidade de utilização e confiança no sistema

A primeira tabela mostra os resultados combinados das várias línguas e a seguinte o caso português.

	Intuitividade	Facilidade de utilização	Confiança no sistema
Média:	2	3	3
Moda:	2	2	2
Mediana:	2	2	2

Tabela 10 - Intuitividade, facilidade de utilização e confiança no sistema (combinada)

	Intuitividade	Facilidade de utilização	Confiança no sistema
Média:	2	3	3
Moda:	2	2	2
Mediana:	2	2.5	3

Tabela 11 - Intuitividade, facilidade de utilização e confiança no sistema (português)

A primeira linha das tabelas apresenta a média obtida a partir do somatório dos resultados obtidos nos questionários dividido pelo número de resultados. A segunda linha (moda) apresenta o valor mais frequente das respostas. A última linha apresenta a mediana, que é o valor que “divide a meio” os dados ordenados, ou seja, 50% dos dados estão abaixo da mediana e, obviamente, os restantes 50% estão acima.

Os resultados indicam que os utilizadores acharam o sistema intuitivo. No entanto alguns utilizadores encontraram áreas pouco intuitivas, tendo sido tolerantes com esta situação.

Em relação à facilidade de utilização, as opiniões ficaram pelo “meio”, ou seja, nem muito fácil, nem muito difícil. Sendo que o mesmo se aplica à “Confiança no sistema”. No entanto, não foi possível deixar de registar os problemas existentes ao nível no reconhecimento de fala.

b) Satisfação

As tabelas seguintes ilustram os resultados para os níveis de satisfação

	Geral	Navegação	Interacção	Voz das prompts	Voz do emails	Conversação
Média:	3	3	3	2	2	3
Moda:	4	2	4	2	2	2
Mediana:	3	2	3	2	2	3

Tabela 12 - Níveis de satisfação (combinada)

	Geral	Navegação	Interacção	Voz das prompts	Voz do emails	Conversação
Média:	3	3	3.5	2	2.5	3
Moda:	4	2	4	2	2	3
Mediana:	3	2.5	4	2	2	3

Tabela 13 - Níveis de satisfação (português)

Ao nível da satisfação geral, os utilizadores mostraram-se nem muito satisfeitos nem insatisfeitos, embora um boa parte tenha indicado insatisfação (moda=4). Uma das razões para alguma insatisfação tinha a ver a relativa lentidão do sistema.

Quer para a navegação, quer para a interacção, os resultados foram semelhantes: nem satisfeitos nem insatisfeitos, embora no caso particular do português, pareceu existir alguma insatisfação.

De uma forma geral, os utilizadores mostraram-se satisfeitos com a voz utilizada nas *prompts* do sistema e na leitura dos e-mail.

Os utilizadores não se mostraram nem satisfeitos, nem insatisfeitos ao nível da conversação com o sistema. Os resultados aqui poderiam ter sido melhores não fossem os problemas de reconhecimento que influenciaram negativamente o fluxo do diálogo entre os utilizadores e o sistema. No entanto, não se determinou correlação estatisticamente significativa entre o grau de reconhecimento percebido pelo utilizador e o grau objectivo de reconhecimento do sistema.

c) Uso futuro do sistema

Aproximadamente 44% dos utilizadores afirmaram que utilizariam o sistema VPA no futuro.

d) Comentários dos utilizadores aos pontos positivos do sistema

- “A mobilidade possibilitada pelo acesso telefónico ao e-mail”;
- “O sistema reage a frases complexas”;
- “Opções de acessibilidade”.

e) Comentários dos utilizadores aos pontos negativos do sistema

- “Nível de reconhecimento baixo quando se usa um diálogo natural”;
- “Tempos de resposta lentos e problemas de reconhecimento”;
- “Difícil recuperar de erros”.

f) Comentários dos utilizadores sobre como melhorar o sistema

- “Melhorar os tempos de resposta”;
- “Oferecer a possibilidade de mudar o assunto das mensagens por voz”;
- “Melhorar o reconhecimento de voz”.

g) Impressões iniciais dos utilizadores

- “Sistema com muito potencial, mas ainda com alguns problemas”;
- “Divertido, Versátil, Potencial”;
- “Ainda falta muita coisa, mas foi interessante conversar com o computador”;
- “Não muito bom”;
- “Inovador”;
- “Muito fácil de aprender”.

6.3.3 Lições aprendidas

Conforme referido, os objectivos gerais traçados para a avaliação eram a recolha de informação sobre a usabilidade e acessibilidade do sistema. Tendo em conta a área de actividade da empresa onde o autor desenvolve a sua actividade profissional, um objectivo particular da avaliação era o estudo e a tentativa de compreender que factores afectam a experiência de utilização e de aceitação deste tipo de serviços. No decurso do processo de avaliação, o autor participou numa avaliação heurística e no teste com utilizadores finais.

A avaliação heurística foi efectuada por 3 peritos que inspeccionaram sistematicamente a interface do utilizador, para verificar a sua aderência a princípios de

usabilidade. Foi fornecido aos peritos um plano de testes, a lista de funcionalidades a explorar, uma ficha para reportar os resultados e a lista das heurísticas a utilizar. Os peritos inspeccionaram a interface e anotaram todos os aspectos relevantes com que se depararam na ficha de resultados. As sessões foram registadas recorrendo a equipamento de gravação. No final da avaliação heurística, os peritos encontraram-se e elaboraram as conclusões com base nos seus resultados individuais. A inspecção da interface pelos peritos revelou vários aspectos que, para facilitar a despistagem, foram divididos em 3 secções: expectativas dos utilizadores (a), aspectos de usabilidade (b) e defeitos (c). A avaliação heurística revelou-se muito eficaz na detecção de problemas com a vantagem adicional de ser pouco cara, em termos de tempo e complexidade.

O teste com utilizadores pretendeu avaliar a usabilidade e aceitabilidade do sistema. Embora as boas práticas recomendem que os utilizadores devam ser exteriores à empresa, foi acordado entre os parceiros não convidar utilizadores de fora das próprias organizações nem pouco experientes com este tipo de sistemas, dada a instabilidade do sistema. Foram seleccionados na empresa 12 utilizadores, tendo os seus dados demográficos sido devidamente caracterizados. Nenhum dos utilizadores estava envolvido no projecto ou familiarizado com o sistema. Na preparação da experiência foram criados documentos para os diferentes intervenientes (protocolos para os observadores, instruções, questionários e cenários para os utilizadores) e foi definida e posta em prática toda a logística necessária à execução dos testes (equipamento, ambiente de teste, convidar os utilizadores, horário das sessões, entre outros). Cada sessão de avaliação teve a duração de cerca de 2 horas para cada utilizador. As sessões experimentais decorreram em ambiente laboratorial, tendo os utilizadores interagido com o sistema através de um conjunto de tarefas previamente definidas com base em cenários reais de utilização do produto. Enquanto o utilizador efectuava as tarefas, os observadores tiravam notas da experiência de interacção desse mesmo utilizador e dos seus comentários. As interacções do utilizador foram gravadas recorrendo a ferramentas próprias, o que permitiu a análise posterior da interacção completa entre o sistema e o utilizador e também obter dados objectivos sobre o desempenho do sistema, no que se refere ao reconhecimento e compreensão das solicitações do utilizador. A etapa final consistiu em analisar os dados obtidos através dos questionários, dos logs do software e das observações realizadas e dos comentários dos utilizadores.

O método empregue na realização deste trabalho, por si só, demonstrou-nos a necessidade de refinar sucessivamente o sistema através da realização de mais iterações no ciclo de desenvolvimento, bem como sujeitar o sistema a um teste mais alargado em termos de utilizadores finais. No entanto, as avaliações podem ser extremamente caras, quer ao nível de tempo quer monetariamente. As tarefas de recrutamento de utilizadores, preparação do laboratório e análise estatística revelaram-se particularmente caras.

7 Processo de teste de interfaces de voz num contexto empresarial: situação actual e propostas de melhoria

7.1 Contexto empresarial

Nesta subsecção será apresentada uma breve descrição do departamento onde o autor desempenha funções profissionais e a situação actual do mesmo ao nível quer das aplicações desenvolvidas, quer ao nível do processo de teste implementado.

7.1.1 O departamento

A Portugal Telecom Inovação, SA (PT Inovação), constituída em 1999, é o braço para a Investigação e o Desenvolvimento do Grupo Portugal Telecom. Tem como objectivo principal garantir o processo de inovação nas empresas do Grupo PT, tanto em Portugal como no estrangeiro, através da prestação de serviços às mesmas, contando com 50 anos de experiência. A PT Inovação é detida em 100% pela Portugal Telecom SGPS, SA (mais informação em <http://www.ptinovacao.pt>). A sede da empresa situa-se em Aveiro, contando com pólos no Porto, Lisboa e São Paulo, no Brasil.

O autor desempenha funções profissionais exactamente no pólo do Porto da PT Inovação, onde trabalham cerca de 40 engenheiros. Os principais departamentos deste pólo são a) sistemas de *customer care* pré-pago e b) serviços e plataformas de voz, sendo este ultimo, especificamente, onde se encontra o autor. O objectivo deste departamento é colaborar com as Empresas do Grupo PT na implementação de Serviços Interactivos de Fala que procuram ser inovadores e competitivos. A abordagem para atingir tal objectivo consiste na focalização na área das tecnologias para gestão da interacção e integração de modos de contacto e, ainda, na integração de produtos existentes em soluções de valor acrescentado usando tecnologias de fala. As principais actividades do departamento podem ser agrupadas em consultoria em soluções de atendimento, desenvolvimento de sistemas para atendimento e implementação de serviços de atendimento automático.

De acordo com o modelo CMMI-SW (*Capability Maturity Model Integration for Software Engineering*), é possível dizer que a empresa se encontra no nível 3 de maturidade, pondo em prática medidas para atingir o nível 4. O modelo CMMI-SW foi desenvolvido pelo Software Engineering Institute (SEI) da Carnegie Mellon University (CMU) com o intuito de ajudar as organizações a melhorar os seus processos e serviços e preconiza 5 níveis de maturidade. No nível 3, “Definido”, o conjunto de processos da organização está estabelecido e é melhorado ao longo do tempo. Cada área de negócio e cada projecto adapta os processos de acordo com as suas necessidades e de acordo com regras definidas para essa adaptação. O nível 4, “Gestão quantitativa”, pressupõe a existência e recolha de métricas detalhadas do processo de desenvolvimento de *software* e qualidade dos produtos. A secção 7.1.3 do presente documento detalha o processo de desenvolvimento de *software* na empresa e, aí, se justificará o nível de maturidade da empresa de acordo com o CMMI-SW.

7.1.2 Serviços e aplicações

Como foi referido na secção anterior, o principal objectivo do departamento de serviços e plataformas de voz do pólo do Porto da PT Inovação é colaborar na implementação de serviços interactivos de voz. De seguida é apresentada uma lista dos principais marcos do departamento:

- Início do desenvolvimento da plataforma INOVox, ainda no Inesctel (1995);
- Serviço Informativo “118” com reconhecimento de fala (1995);
- Síntese de Fala em *Unified Messaging* (2000);
- Portais de Voz (2001);
- Serviços Multimodais, como no caso do projecto europeu MUST (2002);
- Messaging Rede Fixa: conversão SMS para Voz (2003);
- Plataforma de Áudio-conferência (2004);
- Validação do Orador (2005);
- *Voice Activated Dialling* (VAD) 2006.

Destas aplicações, no entanto, as que mais se aproximam do conceito de sistema conversacional são o serviço multimodal desenvolvido no projecto europeu MUST e o ainda por terminar (à altura da conclusão da presente dissertação) VAD. O primeiro é conhecido já de conferências e artigos (ver, por exemplo [Almeida 02] e [Gomes 02]). O segundo, VAD, é um serviço *call-in* que permite aos utilizadores, quer da rede fixa quer da móvel, a marcação por fala usando linguagem natural. Assente em tecnologia de reconhecimento de fala independente do orador, permite aos clientes a realização de chamadas de saída através de simples comandos vocais. O reconhecimento de fala (ASR) é independente do orador e baseado no produto Nuance/Scansoft OSR3. A conversão de Texto para Fala (TTS) é baseada no produto Nuance/Scansoft RealSpeak.

De seguida apresenta-se um exemplo de interacção com o VAD. Suponhamos que alguém liga para a empresa para falar com o autor do presente documento:

VAD: Com quem deseja falar?

Utilizador: “Rui Gomes”

VAD: Não entendi o que disse. Por favor diga o primeiro e último nome da pessoa com quem deseja falar...

Utilizador: “Rui Gomes”

VAD: Deseja transferir a chamada para Rui Gomes?

Utilizador: “Sim”

VAD: Por favor aguarde, a sua chamada vai ser transferida para Rui Gomes

Se o reconhecimento falhar duas vezes, pelo nome ser de difícil reconhecimento (neste caso é guardado para posterior análise fonética) ou por ser um colaborador novo, a chamada é passada para o atendedor humano (a designada estratégia *fallback*).

Um serviço deste tipo permite a redução de custos na empresa, maior rapidez e menor atendimento humano, maior disponibilidade, facilidade de utilização do serviço (particularmente em situações de mobilidade) e ter um único número para acessos fixo ou móvel, entre outras.

A esmagadora maioria dos serviços e aplicações desenvolvidos no departamento são relativamente simples em termos de estratégia de gestão de diálogo e são baseados em máquinas de estados finitos. Nos sistemas baseados em máquinas de estados as transições de estado para estado são desencadeadas através de frases do sistema (*prompts*) e respostas (*utterances*) do utilizador, sendo perfeitamente determinísticos. A quantidade de estados dos serviços desenvolvidos é reduzida, o que permite “relaxar” as estratégias de teste utilizadas.

7.1.3 O sistema de qualidade e o processo de testes

A PT Inovação iniciou em 1999 um projecto global de construção de um Sistema de Qualidade para a empresa, em que o processo de desenvolvimento de *software* é parte determinante. Dos requisitos que este Sistema de Qualidade deveria cumprir podemos destacar:

- Documentação comum a toda a organização;
- Permitir visões personalizadas;
- Integrar certificação NP EN ISO9001 e acreditação NP EN ISO/IEC 17025 nos laboratórios;
- Automatização progressiva do processo de gestão documental;
- Interface de acesso por *browser* a todos os colaboradores;
- Ser útil;
- Melhorar a eficácia dos processos.

Em 2002 foi obtida a certificação da empresa ISO 9000:2000 e já em 2006 a certificação ambiental ISO 14001\2004. A melhoria contínua e a qualidade fazem parte da responsabilidade da empresa e de cada um dos seus colaboradores, tendo em vista garantir a satisfação dos clientes como razão de sucesso do negócio, o desenvolvimento profissional dos colaboradores através de uma adequada conjugação entre formação e prática e na disponibilização de um ambiente de trabalho salutar e desafiante e a gestão adequada dos recursos naturais, energéticos e resíduos produzidos.

O Sistema de Qualidade da PT Inovação assenta em procedimentos organizados em processos. Os procedimentos descrevem as regras de actuação e as respectivas responsabilidades definidas por papéis. Executando-se os papéis dos procedimentos produzem-se os registos de trabalho. Podemos classificar os processos por "grupos" tendo em conta o seu posicionamento na lógica de funcionamento da empresa:

- Processos de produção - Executam os diversos produtos disponibilizados aos clientes, integrando, por exemplo o Processo de Desenvolvimento de Sistemas (PDS);
- Processos de gestão global - Coordenam a gestão da empresa, ao nível global, da Área e da Acção;
- Processos de gestão dos recursos humanos - Coordenam as necessidades relativas aos recursos humanos, no seu ciclo de vida na empresa;
- Processos de relacionamento com o cliente - Define e coordena as interacções com os clientes;
- Processos de relacionamento com fornecedores - Define a forma como se executam os processos de aquisição e relacionamento com os fornecedores;
- Processos de suporte - Coordenam a execução das actividades de suporte técnico.

Concentrando-nos apenas no Processo de Desenvolvimento de Sistemas, pois é o mais relevante para a presente dissertação, podemos dizer que é um processo de produção que tem como objectivo fundamental o desenvolvimento de sistemas, podendo esse desenvolvimento ser feito a partir de um conjunto de requisitos apresentados por um cliente externo ou identificados internamente na própria organização.

Assim, um sistema poderá dar resposta:

- a necessidades identificadas por um cliente;
- a necessidades identificadas pelo cliente em conjunto com a organização;
- a futuras necessidades de mercado identificadas pela organização;
- a necessidades internas identificadas no seio da própria organização.

O PDS pode assim ser aplicado ao desenvolvimento de *software*, *hardware*, soluções híbridas ou soluções integrando mais que um sistema. Este processo integra dois sub-processos paralelos que podem ser utilizados em alternativa para o desenvolvimento de um sistema, o PDS BASE e PDS SIMPLES (PDS para sistemas simples). É da responsabilidade do responsável da acção optar por um deles. O conjunto de actividades que compõe cada um dos PDSs encontram-se, obviamente, descritas em documentos próprios e facilmente acessíveis. Os resultados previstos no PDS, obrigatórios ou opcionais, podem ser substituídos por resultados obtidos a partir de ferramentas aprovadas pelo responsável do processo. Todas as actividades descritas nos dois sub-processos PDS identificam as formas de verificar a adequação dos resultados, nomeadamente através de revisões e/ou aprovações. O PDS disponibiliza um conjunto de *checklists* que podem ser utilizadas para suporte da revisão técnica dos resultados das diversas actividades. O processo PDS possui ainda um conjunto de indicadores.

Em termos de Execução, os principais papéis / funções internas ao PDS são:

- Analista de Sistemas;
- Arquitecto de Sistemas;
- Programador;
- Gestor de Documentação do Cliente;
- Testador;
- Instalador;
- Gestor de Configurações;
- Administrador de Base de Dados.

A caracterização das funções de cada um destes papéis, bem como os requisitos mínimos para o seu preenchimento, estão devidamente definidos no manual do PDS. No caso do testador as suas funções, conforme a manual de qualidade da empresa, são as seguintes:

- Elaborar, em colaboração com outros elementos da equipa, as especificações de testes;
- Colaborar na revisão das especificações;
- Garantir a correcta execução dos testes;
- Registar os problemas detectados;
- Garantir a correcção dos problemas pela equipa de desenvolvimento.

Ainda de acordo com o manual de qualidade, este papel deve ser desempenhado por um quadro técnico, com conhecimento das funcionalidades do sistema a ser testado e da metodologia de testes adoptada. No caso de se ter adoptado uma ferramenta de suporte deve ter conhecimento das funcionalidades da respectiva ferramenta. São aconselhadas as ferramentas CVS (Gestão de versões e configurações) e TestDirector (Especificação e registo de testes).

O CVS é um conjunto de ferramentas para fazer a gestão de versões de *software* e de *hardware*. Estas ferramentas fornecem o controlo de versões, a coordenação do trabalho em grupo e resolução de conflitos num ambiente multi-utilizador de múltipla edição. Associado a estas ferramentas, existe um conjunto adicional que permite a visualização e administração do repositório através da *web* (WebCVS), entre outras.

O TestDirector é uma ferramenta utilizada para especificar testes, registar a sua execução e gerir os defeitos eventualmente detectados. A ferramenta permite, caso seja necessário, obter relatórios da especificação dos testes, registo da execução e lista de defeitos. Esta potente ferramenta ajuda a organizar e a dirigir todas as fases do processo de testar o *software* (que tende a ser complexo), incluindo o planeamento, a criação, a execução dos testes e o controlo dos defeitos. Os utilizadores podem utilizar a versão do TestDirector disponível mediante acesso na Intranet da empresa.

O TestDirector mantém os testes numa base de dados e simplifica a sua gestão, ajudando a dirigir todas as fases do processo de testes, incluindo o planeamento, a especificação, a execução dos testes e o controlo dos defeitos. A partir do TestDirector consegue-se construir grupos de testes (*test sets*) para se cobrir objectivos específicos. Por exemplo, pode-se criar um ou mais grupos de testes que verifiquem um módulo (Testes de Módulo), todos os requisitos de um sistema (Testes de Sistema), novas versões do *software* (novas versões de Testes de Sistema), ou que verifiquem apenas um conjunto de funcionalidades específicas, como resultado por exemplo da correcção de um *bug*.

O fluxo de trabalho quando se utiliza o TestDirector consiste, essencialmente em 3 fases:

- Planear e Especificar os Testes;
- Executar os Testes;
- Controlar os Defeitos.

A figura seguinte ilustra a dinâmica entre as 3 fases do fluxo de trabalho do TestDirector:

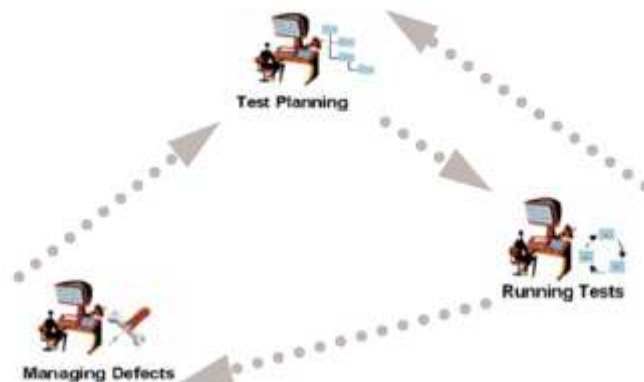


Figura 20 - Fluxo de trabalho do TestDirector

Do PDS fazem parte artefactos como a Especificação de Requisitos, Relatório de Concepção, Especificação de testes, Plano de testes, Modelo para Registo de testes, Manual de Utilização, Manual de Instalação, Relatório de Instalação, entre outros.

Uma vez que se insere no tema da presente dissertação, serão abordados com mais detalhe os testes no PDS.

No PDS, qualquer tipo de testes deve passar pelas seguintes fases seguintes: Planeamento, Especificação, Execução e Registo. Os tipos de testes previstos são Testes de Módulo, Testes de Integração, Testes de Sistema, Testes de Instalação e Testes de Aceitação. Os testes de módulo são da responsabilidade do implementador desse módulo. A execução

dos restantes tipos de testes é atribuída ao responsável pela acção e/ou testadores por este designados. Os testes de sistema e os testes de instalação, no caso desta ser da responsabilidade da PT Inovação, são obrigatórios, sendo portanto também o seu planeamento e especificação obrigatórias. Os testes de aceitação são em princípio da responsabilidade do cliente, podendo no entanto a empresa colaborar na sua especificação e execução, a pedido deste. O registo da execução dos testes e dos defeitos é feito, tipicamente, no TestDirector. Quando em qualquer dos testes referidos anteriormente são identificadas pelos responsáveis pelos testes discrepâncias entre os resultados esperados e os resultados actuais do teste, estas discrepâncias são registadas como defeitos (com diferentes níveis de severidade) e remetidas aos membros da equipa de desenvolvimento para análise e correcção. As causas são identificadas e no caso do tipo de defeito detectado impedir a continuação dos testes, estes ficam suspensos até serem executadas as correcções necessárias, que são registadas também no TestDirector. Antes da passagem à instalação para entrega de um sistema é feita uma revisão dos testes, de forma a garantir que o processo de testes satisfaz o que está definido no PDS e que o sistema a entregar ao cliente satisfaz os requisitos deste. Este processo procura garantir, por exemplo, que a especificação de testes de sistema cobre todos os requisitos identificados nos documentos de requisitos, que os testes foram realizados e registados, que os defeitos foram documentados, entre outros.

7.2 Situação actual do teste de interfaces de voz

No caso concreto dos serviços de voz desenvolvidos no departamento da empresa onde o autor desempenha funções profissionais, e, em parte, dada a relativa simplicidade ao nível de gestão do diálogo dos mesmos, a metodologia adoptada para efectuar os testes de sistema (ver secção anterior do presente capítulo) consiste em “manualmente” percorrer a máquina de estados que constitui o serviço, procurando definir casos de teste que cobram todos os estados e transições. Cada caso de teste descreve uma sequência de entradas e saídas e pode cobrir vários estados e transições. Depois, é estabelecida uma chamada telefónica para o serviço, e procuram-se executar esses casos de teste. São executados testes para passar e para falhar. Simultaneamente, nos serviços em que existem esses componentes, procura-se testar o desempenho do reconhecedor de fala e a qualidade da síntese de fala.

Ao nível da usabilidade os testes são inexistentes, na maior parte dos casos, ou, na melhor das hipóteses, feitos *ad hoc* (salvo honrosas excepções). No entanto, obviamente, no desenvolvimento das interfaces de voz procuram-se seguir as melhores práticas de usabilidade.

Nos serviços de menor complexidade e/ou menor peso orçamental, os testes de sistema são efectuados pelos "desenvolvedores" dos mesmos.

7.3 Propostas de melhoria do processo de teste de interfaces de voz

7.3.1 A equipa de testes

A primeira e, quiçá, mais profunda e importante proposta é a constituição de uma equipa de testes transversal a todo o departamento. A equipa deveria possuir competências nas áreas mencionadas ao longo desta dissertação, como sejam a Engenharia de Software (particularmente teste e avaliação de produtos de *software*), Tecnologias de Fala e Usabilidade.

As responsabilidades de tal equipa iriam desde a especificação, planeamento e execução dos testes, acompanhamento dos defeitos, passando ainda pelas avaliações de usabilidade adequadas, garantindo o PDS (ver secção 7.1). Obviamente, em relacionamento

estreito com a equipa/departamento de qualidade e com as equipas de programadores dos diversos projectos e reportando aos responsáveis de acção.

Em função da dimensão do departamento, do número de projectos e dimensão dos mesmos, a equipa poderia ser constituída por 3 elementos (no mínimo 2).

7.3.2 Teste de ASR e TTS

Relativamente aos módulos de reconhecimento automático e de síntese automática de fala (ASR e TTS), o seu teste por parte das organizações é extremamente complicado. Tipicamente, um sistema deste tipo é escolhido em função de testes *ad hoc* e/ou em função da relação preço-qualidade. Por exemplo, uma metodologia rudimentar que já foi utilizada na empresa, consistiu em ter um IVR que efectuava chamadas para si próprio e tocava *prompts* com o TTS, gerando assim fala a partir de textos de entrada, para testar o desempenho do ASR instalado. Contudo, conforme ficou demonstrado no capítulo 5, em concreto nas secções 5.1 e 5.2, o teste e avaliação destes componentes é extremamente caro. Por outro lado, actualmente, ainda não é viável comercialmente desenvolver sistemas conversacionais pois não existe receptividade no mercado para tal. Assim sendo, é complicado justificar o custo elevado do teste e avaliação de diferentes sistemas ASR e TTS, tendo em vista requisitos específicos para sua adopção. É inconcebível, do ponto de vista de retorno financeiro, estar regularmente a investir em novos sistemas deste tipo. Como tal, a solução passa por estabelecer parcerias com os maiores fabricantes destes sistemas, para que estes possam fornecer versões de demonstração e/ou novas versões dos existentes na empresa. O trabalho, depois, consiste em retirar o máximo partido dos sistemas disponíveis, no sentido da afinação de parâmetros, construção de gramáticas, modelos, entre outros, para que se consiga atingir a qualidade pretendida em cada serviço desenvolvido.

Um recurso extremamente valioso, como referido anteriormente (ver secção 5.1.1 do documento), são os corpora de voz. Existem na empresa este tipo de recursos, que foram criados no âmbito de projectos europeus de grande dimensão. Não são, no entanto, em número nem dimensão suficientes – aliás, são manifestamente escassos e cobrindo poucos domínios de aplicação. Não obstante, devem ser aproveitados ao máximo para testar, ainda que de forma muito superficial, quer o desempenho do ASR quer a qualidade do TTS (embora este ultimo implique sessões com utilizadores), conforme preconizado no capítulo 5 da presente dissertação. Uma outra proposta, ao nível dos corpora, passa por estabelecer parcerias com universidades e institutos de investigação, para intercâmbio quer de *know-how* quer dos próprios corpora, visto que ambas as partes poderiam beneficiar muito com esta situação. Não existem no tecido empresarial nacional muitas empresas que trabalhem sobre este tipo de tecnologias e recursos. Isto permitira adquirir competências e diversificar os domínios cobertos pelos corpora de ambas as partes, visto que são recursos com custos de criação extremamente elevados e, simultaneamente, de uma utilidade extrema. Até para promover o Português enquanto língua que se pretende afirmar neste domínio. Obviamente, que existe o problema de ser um departamento pequeno e é difícil afectar recursos (um grupo pequeno) a este tipo de actividade.

7.3.3 Testes funcionais

Conforme já mencionado, a metodologia concreta adoptada na empresa para efectuar os testes de sistema consiste em, “manualmente”, percorrer a máquina de estados que constitui o serviço, procurando definir casos de teste que cubram todos os estados e transições. Melhorias nos resultados do processo poderiam ser obtidas através da aplicação de uma combinação de técnicas de teste de “caixa-negra”, sendo a principal o teste baseado em cenários (conforme

exemplificado na secção 5.3 para o caso concreto do projecto FASiL). Os cenários podem ser vistos como sequências de interações entre os utilizadores e o sistema, representando fluxos normais e anormais de eventos dos casos de uso. Através de uma matriz de rastreabilidade podemos anotar o mapeamento entre os cenários de teste e os casos de uso e deve ser identificado pelo menos um caso de teste, e respectivas condições, por cada cenário. Uma das dificuldades é a sequenciação dos cenários e a influência que esta pode ter no teste.

Na combinação de técnicas, e para além dos testes baseados em cenários, poderíamos incluir a identificação de classes de equivalência (para entradas válidas e inválidas), adivinhação de erros e testes exploratórios. A adivinhação de erros envolve a construção de uma lista de erros que são espectáveis numa determinada área do sistema e desenhar um conjunto de casos de teste para a sua verificação. Nos testes exploratórios, o testador baseado na sua experiência, “navega” pela interface de voz em busca de situações problemáticas. Estas situações são anotadas, e posteriormente, são analisadas e servem como base para a especificação de casos de teste.

Para além da combinação de técnicas, deve ser estudada e explorada a automatização do processo onde possível, tendo em visto reduções de tempo e custo, bem como a promoção uma maior robustez e reprodutibilidade dos testes. Por exemplo, testar o fluxo do diálogo sem recorrer à voz ou ao telefone, interagindo com o modelo de estados através de comandos de texto poderá acelerar e reduzir os erros do processo em grande medida e permitir uma separação maior entre testes funcionais (fluxo do diálogo) e testes ao desempenho do reconhecedor.

Naturalmente, e uma vez que já é prática na empresa, todo este processo deve ser norteado pela utilização do TestDirector.

7.3.4 Base de dados de erros

A detecção de falhas é um problema complexo que deve ser abordado com um conjunto de técnicas dinâmicas e estáticas. Uma proposta de técnica estática para procurar reduzir custos, melhorar a qualidade e diagnosticar o processo de desenvolvimento (entre outras) é a construção de uma base de dados de falhas mais frequentes, com base na experiência anterior, associada a um sistema de classificação das mesmas tal como no conceito de *Orthogonal Defect Classification* (ODC) [Chillarege 92], constituindo assim “perfis de defeitos” com as falhas mais frequentes.

O conceito ODC foi inventado pela IBM (*International Business Machines*) no início da década de 90 do século XX. A ideia base do conceito é categorizar as falhas, convertendo a informação das mesmas em métricas valiosas do processo de desenvolvimento de *software* e/ou do produto. É possível categorizar as falhas em classes que, colectivamente, indicam que parte do processo necessita atenção [Chillarege 92]. O ODC define um conjunto de atributos para as falhas e um conjunto de regras para analisar essas mesmas falhas. Depois, com base na análise, podem ser tomadas acções para melhorar a qualidade. No ODC cada defeito é caracterizado por oito atributos: quando o testador encontram e reporta uma falha, pode preencher atributos como “Actividade”, “Tipo de activação” e “Impacto”. Quando o desenvolvedor corrige ou responde a uma falha, pode preencher os atributos “Idade”, “Origem”, “Qualificador”, “Tipo de Falha” e “Alvo”. Os seus significados são os seguintes:

- Actividade (*Activity*): indica qual a actividade do processo na qual a falha é descoberta;
- Tipo de activação (*Fault trigger*): descreve o ambiente ou condição que levaram à exposição da falha;
- Impacto (*Impact*): refere qual o impacto nos utilizadores;

- **Alvo (*Target*):** representa a identidade de alto nível da entidade que foi alvo de correcção;
- **Tipo de falha (*Fault type*):** representa o tipo de correcção feita. Exemplos de tipos de falhas preconizados no ODC: atribuição, teste, interface, temporização, algoritmo, função;
- **Qualificador (*Qualifier*):** aplica-se ao tipo de falha e especifica se a correcção feita se deveu a informação ou código em falta, incorrecto ou externo;
- **Origem (*Source*):** indica se a falta foi encontrado em código desenvolvido, reutilizado ou adquirido externamente;
- **Idade (*Age*):** indica a versão da origem da falha.

O ODC pode ser integrado no processo de desenvolvimento estabelecido na empresa, permitindo depois a criação de casos de teste baseado baseados na análise de padrões dos “perfis de erros” e a monitorização de qualquer passo desse mesmo processo, da eficiência dos testes e das correcções, do estado do produto, entre outras.

7.3.5 Avaliação de usabilidade

Ao nível da usabilidade, conforme referido, os testes são inexistentes na maior parte dos casos, ou, na melhor das hipóteses, são *ad hoc* (salvo honrosas excepções).

Seria interessante incluir no PDS (Processo de Desenvolvimento de Sistemas) da empresa a noção de avaliação heurística. A avaliação heurística ([Nielsen 1990] e [Nielsen 1994]) tem por finalidade detectar problemas de usabilidade no desenho da interface de forma a que possam ser solucionados no processo iterativo de interacção. Dada a sua flexibilidade e “leveza” pode ser utilizada em qualquer etapa do processo de desenvolvimento, tal como propõe [Dix 04]. Esta avaliação heurística poderia ser levada a cabo pela equipa de testes referida na secção 7.3.1, com o apoio de *checklists* criadas para o efeito contendo as “regras de ouro” que representam as boas práticas de usabilidade e uma tabela onde se possam anotar os resultados. Segue-se o exemplo de uma tabela para o efeito:

Área do Sistema	Problema de Usabilidade	Princípio Violado	Comentários / Recomendações

Tabela 14 - Tabela para anotação dos resultados da avaliação heurística

A tabela tem quatro colunas: na primeira deve-se colocar a parte do sistema sob análise, na coluna 2 deve-se ser colocado o potencial problema de usabilidade detectado, a terceira coluna contém o princípio de usabilidade violado (socorrendo-se o avaliador da *checklist* das “regras de ouro”), por fim, a quarta coluna reserva espaço para comentários e/ou recomendações que o avaliador pretenda incluir.

A acompanhar a inclusão das avaliações heurísticas, ou decorrente das mesmas, poder-se-ia tornar mais iterativo e ágil o próprio PDS, com um maior nível de prototipagem de forma a que o *feedback* das avaliações heurísticas fosse incluído no processo, procurando que este se centre mais no utilizador do que o que acontece actualmente.

Quanto às avaliações com utilizadores finais, o seu custo é demasiado elevado para a dimensão dos sistemas de voz actualmente desenvolvidos no departamento e para o tempo-para-o-mercado (*time-to-market*) imposto para essa dimensão - pelo menos com um número de utilizadores que possibilitem significância estatística à posterior análise de resultados. No entanto, com o expectável surgimento de oportunidades para desenvolver sistemas mais complexos e interactivos, a avaliação com utilizadores finais tornar-se-á uma necessidade absoluta. Nessa altura, o principal entrave será a reduzida dimensão do próprio departamento, pois os aspectos logísticos serão pertinentes. Então, a proposta é a criação de um laboratório de usabilidade transversal à organização, ou seja, na sede, por exemplo, criar um laboratório com as condições técnicas ao nível de espaço e equipamento para realizar as avaliações com utilizadores finais e, ainda, a obtenção de competências para a análise dos resultados obtidos tendo em vista a sua inclusão nos projectos.

8 Conclusões e trabalho futuro

As pessoas, empresas e até as economias dependem fortemente do sucesso do desenvolvimento e utilização de sistemas de *software*, sendo necessário tentar desenvolvê-los otimizando recursos e com cada vez mais qualidade. Assim, o teste de *software* está rapidamente a tornar-se numa área crítica da Engenharia de Software. Testes completos e consistentes são extremamente importantes para se obter a qualidade desejada.

Os avanços registados nos últimos anos nos sistemas de reconhecimento de fala e tecnologias de desenvolvimento potenciaram o surgimento de inúmeras oportunidades para as empresas fornecerem novas interfaces baseadas na voz para aplicações existentes ou novos tipos de aplicações. Contudo no caso das interfaces de voz, a verificação e validação são campos de estudo ainda em aberto. Foi exactamente este tema que a presente dissertação procurou abordar. Apesar da necessidade de testes ao longo de todo o processo, foi dada ênfase ao nível de testes de sistema.

Os principais tipos de testes de sistema neste tipo de aplicações são:

- Testes funcionais (*functional testing*);
- Testes de carga e desempenho do sistema (*load and performance testing*);
- Testes de desempenho do reconhecimento de fala e qualidade da saída;
- Testes de usabilidade.

Para atingir os objectivos propostos no início do trabalho de dissertação (enunciados no princípio do presente documento), começou-se por procurar fazer uma abordagem ao conhecimento de base da área, como a linguística, a neurolinguística, a teoria da comunicação ou até o estudo da anatomia Humana que possibilita a produção de voz e a sua captação. Procurou-se também fazer uma pequena síntese da história das tecnologias de fala.

Com um carácter mais técnico, foram estudados:

- o funcionamento dos sistemas conversacionais
 - o que são
 - quais os principais tipos
 - tecnologias
 - arquitecturas mais comuns
- a síntese automática de fala, reconhecimento automático de fala e gestão do diálogo
 - o que são
 - como funcionam
 - quais são os tipos mais comuns
 - qual o estado da arte destes componentes fundamentais
 - o tipo de testes e métricas a eles associados
- o tema da usabilidade nas interfaces de voz
 - processo de desenvolvimento orientado ao utilizador

- o melhores práticas
- o técnicas e métricas de avaliação

Seguidamente foram exploradas as melhores práticas ao nível dos testes apropriados a cada componente e à interface, indicando como se realizam, os pontos fortes, as dificuldades e os mais recentes desenvolvimentos. Neste particular, foi exposta ainda a experiência profissional do autor em projectos europeus da área na entidade empregadora, nomeadamente ao nível da criação de corpora linguístico, dos testes funcionais e de usabilidade. Já os testes de desempenho do reconhecedor de fala e do sintetizador são, como decorre do estudo elaborado, extremamente caros em recursos e tempo pelo que não existe na entidade empregadora experiência prática a um nível que não sejam os testes *ad-hoc*.

Ainda no âmbito da entidade empregadora, foi estudado o Sistema de Qualidade, principalmente ao nível do Processo de Desenvolvimento de Sistemas e, mais detalhadamente, os testes: desde os intervenientes, passando pelas actividades, artefactos e ainda ferramentas. Foi ainda feita uma breve análise à metodologia concreta aplicada às interfaces de voz desenvolvidas no departamento no qual o autor desempenha a sua actividade profissional.

No momento, não existe na entidade empregadora um produto de voz, ou projecto em curso, com complexidade adequada à aplicação dos resultados do estudo efectuado ao nível de testes. Assim, procurou-se efectuar uma análise à situação actual do departamento e elaborar um conjunto de propostas aplicáveis ao teste de interfaces de voz tendo em vista a melhoria do processo na entidade empregadora.

Em termos de trabalho futuro, seria interessante e de grande contributo a aplicação dessas propostas na empresa, a avaliação do impacto das mesmas e aplicação de medidas correctivas por forma a atingir melhorias no processo que consubstanciem melhorias na qualidade dos produtos. Outra vertente a explorar no âmbito de trabalho futuro prende-se com a geração automática de casos de teste, tendo em vista ganhos ao nível do esforço e custo necessários ao processo de teste, bem como completude da bateria de testes.

O trabalho de dissertação foi realizado em tempo parcial, tentando o autor conciliar a realização do mesmo com a sua actividade profissional. A verdade é que o regime parcial de realização da dissertação alonga (naturalmente...) os prazos das diversas etapas até à conclusão, por um lado e, por outro lado, dificulta a concentração e dedicação necessárias ao tema escolhido. O exercer de uma actividade profissional numa empresa nem sempre possibilita a conjugação de áreas, interesses e horários de trabalho com o Mestrado. Neste caso, temos portanto de coordenar os tempos de trabalho na empresa e, em horário pós-laboral, os tempos de trabalho na Tese. Em horário pós-laboral nem sempre é fácil o acesso a pessoas e recursos (bibliográficos e outros), não esquecendo os reflexos negativos na vida sócio-familiar. A determinada altura, o grande Adamastor é a própria pessoa e não as questões técnicas da área de estudo. Não é possível deixar de referir, pois marca de forma indelével o resto da vida, a luta pessoal para levar a bom porto a presente dissertação. Poderia ter sido diferente, poderia talvez ser melhor, mas é-o!

Referências e Bibliografia

- [Abrunhosa 94] Abrunhosa, M., Leitão, M., (1994), *Um Outro Olhar Sobre o Mundo*, Filolofosofia 11º Ano, Edições Asa
- [Almeida 02] Almeida, L., Gomes, R., et al, (2002), *Implementing and evaluating a multimodal and multilingual tourist guide*, Proc. CLASS workshop, pp.1-7, Copenhaga
- [Ambrazaitis 05] Ambrazaitis, G. (2005), *Intonation and attitudes: A method for basic research and TTS evaluation*, Speech Synth Course 2005, Universidade de Lund
- [Amith 00] Amith, J. (2000), *Legal, ethical, and policy issues concerning the recording and publication of primary language materials*, Workshop on Web-Based Language Documentation and Description, Institute for Research in Cognitive Science, Universidade da Pensilvânia.
- [Anderson 01] Anderson, Eve, Breitenbach, Stephen, et al, (2001), *VoiceXML – early adopter*, Wrox Press Ltd.
- [Binder 97] Binder, J.R., et al, (1997), *Human Brain Language Areas Identified by Functional Magnetic Resonance Imaging*, The Journal of Neuroscience, Vol. 17, No 1, pp. 353-362
- [Bird 01] Bird, S., Liberman, M., (2001), *A formal framework for linguistic annotation*, Speech Communication, Vol. 33, No. 1-2, pp. 23 - 60
- [Bird 02] Bird, S., et al, (2002), *TableTrans, MultiTrans, InterTrans and TreeTrans: Diverse Tools Built on the Annotation Graph Toolkit*, Proc. LREC2002, pp. 364-370, Las Palmas
- [Branco 05] Branco, G., Almeida, L., Beires, N., Gomes, R., (2005), *A Portuguese spoken and multi-modal dialog Corpora*, Proc. INTER/EUROSPEECH 2005, pp. 1529-1532, Lisboa
- [Burnstein 03] Burnstein, Ilene, (2003), *Practical Software Testing*, Springer
- [Caseiro 02] Caseiro, D., Trancoso, I., (2002), *Large Vocabulary Continuous Speech Recognition Using Weighted Finite-State Transducers*, Proc. Portugal for Natural Language Processing (PorTAL), Faro
- [Chillarege 92] Chillarege, R., et al, (1992), *Orthogonal Defect Classification - A Concept for In-Process Measurements*, IEEE Transactions on Software Engineering , Vol. 18, pp. 943-956

- [Cole 96] Cole, R.A., et al, (1996), *Survey of the State of the Art in Human Language Technology*, Cambridge University Press
- [Cox 98] Cox, S.J., et al, (1998), *Towards speech recognizer assessment using a human reference standard*, Computer Speech and Language, Vol. 12, No. 4, pp. 375-391
- [Dahlbäck 93] Dahlbäck, N., Jönsson, A, and Ahrenberg, A, (1993), *Wizard of Oz Studies - Why and How*, Knowledge-Based Systems, Vol. 6, No. 4, pp. 258-266
- [Dai 00] Dai, R., (2000), *A Technical White Paper on Sage's PSQM Test*, Sage Document, 2000
- [Damásio 92] Damásio, A.R., Damásio, H., (1992), *Brain and Language*, Scientific American, Vol. 267, No. 3, pp. 63-71
- [Damásio 93] Damásio, A.R., Tranel, D., (1993), *Nouns and verbs are retrieved with differently distributed neural systems*, Proc. National Academy of Sciences USA, Vol. 90, No. 11, pp. 4957-4960
- [Damásio 99] Damásio, A.R., (1999), *The fabric of the mind: a neurobiological perspective*, NWO/Huygens Lectures
- [Dix 04] Dix, A., et al, (2004) *Human Computer Interaction*, 3rd Edition, Prentice Hall
- [Flanagan 95] Flanagan, J., (1995), *Research in speech communication*, Proc. National Academy of Sciences USA, Vol 92, pp. 9938-9945
- [Furui 00] Furui, S., Juang, B., (2000), *Automatic Recognition and Understanding of Spoken Language—A First Step Toward Natural Human–Machine Communication*, Proc. IEEE, Vol.88, No.8, pp.1142-1165
- [Furui 03] Furui, S., (2003), *Recent advances in spontaneous speech recognition and understanding*, Proc. IEEE Workshop Spontaneous Speech Processing and Recognition, pp. 1-6, Tóquio
- [Godfrey 92] Godfrey, J., E. Holliman, J. McDaniel, (1992), *SWITCHBOARD: Telephone Speech corpus for research and development*, Proc. IEEE Int. Conf. Acoust. Speech Signal Processing, pp. 517-520, San Francisco
- [Gomes 02] Gomes, R., et al, (2002), *User-friendly Multimodal Services – A MUST for UMTS, Going the Multimodal route: making and evaluating a multimodal tourist guide service*, Proc. Eurescom Summit 2002, Heidelberg

- [Gomes 05] Gomes, R., et al, (2005), *The FASiL Speech and Multimodal Corpora*, Proc. INTER/EUROSPEECH 2005, pp. 1605-1608, Lisboa
- [Graff 00] Graff, D., Bird, S., (2000) *Many Uses, Many Annotations for Large Speech Corpora: Switchboard and TDT as Case Studies*, Proc. LREC2000, pp 427-433, Atenas
- [Greenberg 00] Greenberg, S., et al, (2000), *An introduction to the diagnostic evaluation of switchboard-corpus automatic speech recognition systems*, NIST Speech Transcription Workshop, College Park
- [Guerra 93] Guerra, J.P.F., Viera, J.A.S., (1993), *Aula Viva*, Português-B 10º Ano, Porto Editora
- [Hansson 94] Hansson, S., (1994) *Decision Theory: A Brief Introduction*
- [Hoge 99] Hoge, H. et al, (1999), *SpeechDat multilingual speech databases for teleservices: across the finish line*, Proc. EuroSpeech, Vol. 6, pp. 2699-2702, Budapeste
- [Instituto Ambiente 03] Instituto do Ambiente, (2003), *Procedimentos Específicos De Medição De Ruído Ambiente*, Ministério das Cidades, Ordenamento do Território e Ambiente
- [IPL 2002] IPL, (2002), *An Introduction to Software Testing*, IPL Information Processing, Ltd
- [Kamm 99] Kamm, C., Walker, M., Litman, D., (1999) *Evaluating Spoken Language Systems*, Proc. American Voice Input/Output Society (AVIOS), San Jose
- [Kaner 03] Kaner, C., *The power of 'What If...' and nine ways to fuel your imagination: Cem Kaner on scenario testing*, Software Testing and Quality Engineering Magazine, Vol. 5, No.5, pp. 16-22
- [Kelley 84] Kelley, J., (1984), *An iterative design methodology for user-friendly natural language office information applications*, ACM Transactions on Office Information Systems, Vol. 2, No. 1, pp. 26–41
- [Klein 98] Klein, M., et al, (1998), *Supported Coding Schemes*. MATE Report D1.1
- [Leech 98] Leech, G., et al, (1998), *Survey and guidelines for the representation and annotation of dialogue*, LE-EAGLES-WP4-4 Integrated Resources Working Group
- [Likert 32] Likert, R. (1932), *A Technique for the Measurement of Attitudes*, Archives of Psychology No.140, pp. 55

-
- [Maeda 02] Maeda, K., (2002), *Creating annotation tools with the annotation graph toolkit*, Proc. LREC2002, pp. 1914-1921, Las Palmas
- [Maeda 04] Maeda, K., Strassel, S., (2004), *Annotation Tools for Large-Scale Corpus Development: Using AGTK at the Linguistic Data Consortium*, Proc. LREC2004, Lisboa
- [Makhoul 95] Makhoul, J., Schwartz, R., (1995), *State of the art in continuous speech recognition*, Proc. National Academy of Sciences USA, Vol. 92, No. 22, pp. 9956-9963
- [Markowitz 96] Markowitz, J. A., (1996) *Using Speech Recognition*, Prentice Hall
- [Möller 02a] Möller, S., Boulard, H., (2001), *Analytic Assessment of Telephone Transmission Impact on ASR Performance Using a Simulation Model*, Speech Communication No 38, pp. 441-459
- [Möller 02b] Möller, S., Kavallieratou, E., (2002), *Diagnostic Assessment of Telephone Transmission Impact on ASR Performance and Human-to-Human Speech Quality*, Proc. LREC 2002, Vol. 4, pp. 1177-1184, Las Palmas
- [Monteiro 95a] Monteiro, M., dos Santos, M.R., (1995), *Psicologia 1ºVol.*, Psicologia 12º Ano, Porto Editora
- [Monteiro 95b] Monteiro, M., dos Santos, M.R., (1995), *Psicologia 2ºVol.*, Psicologia 12º Ano, Porto Editora
- [Moran 97] Moran, D.B., Park, S., (1997), *Multimodal User Interfaces in the Open Agent Architecture*, Proc. International Conference on Intelligent User Interfaces, Orlando
- [Nanjo 05] Nanjo, H., Kawahara, T., (2005), *A New ASR Evaluation Measure and Minimum Bayes-Risk Decoding for Open-domain Speech Understanding*, Proc. IEEE-ICASSP, Vol.1, pp.1053-1056, Philadelphia
- [Nielsen 90] Nielsen, J., Molich, R. (1990). *Heuristic evaluation of user interfaces*, Proc. ACM CHI'90, pp. 249-256, Seattle
- [Nielsen 94] Nielsen, J. (1994). *Heuristic evaluation*, Usability Inspection Methods, John Wiley & Sons, New York
- [Oviatt 99] Oviatt, S., (199), *Ten myths of multimodal interaction*, Communications of the ACM, Vol. 42, No. 11, pp. 75-81
- [Pennock 02] Pennock, S., (2002), *Accuracy of the Perceptual Evaluation of Speech Quality (PESQ) algorithm*, Proc. MESAQIN 02, Praga

- [Pereira 03] Pereira, J.R., et al, (2003), *Neuroanatomia Funcional: Anatomia das áreas activáveis nos usuais paradigmas em ressonância magnética funcional*, Acta Médica Portuguesa, Vol. 16, No. 3, pp. 106-117, Lisboa
- [Preece 02] Preece, J., Rogers, Y., (2002) *Interaction Design: Beyond Human-Computer Interaction*, John Wiley & Sons, Inc.
- [Rix 01] Rix, A., et al, (2001), *Perceptual Evaluation of Speech Quality (PESQ) – A new method for speech quality assessment of telephone networks and codecs*, IEEE ICASSP, Vol. 2, pp. 749–752
- [Sa 03] Sa, F., et al, (2003), *Reconhecimento Automático de Fala Contínua em Português Europeu Recorrendo a Streams Audio-Visuais*, Proc. COOPMEDIA 2003, Porto
- [Shannon 48] Shannon, C.E., (1948), *A Mathematical Theory of Communication*, Bell System Technical Journal, Vol. 27, pp. 379-423 e 623-656
- [Shinozaki 03] Shinozaki, T, Furui, S., (2003), *An assessment of automatic recognition techniques for spontaneous speech in comparison with human performance*, Proc. SSPR2003, pp. 95-98, Tóquio
- [Sjölander 00] Sjölander, K., Beskow, J., (2000) *WAVESURFER - An open source speech tool*, Proc. ICSLP 2000, Vol. 4, pp. 464-467, Beijing
- [Skantze 03] Skantze, G., (2003), *The use of speech recognition confidence scores in dialogue systems*, Graduate School of Language Technology, Speech Technology 1, Universidade de Gotemburgo
- [Trancoso 00a] Trancoso, I., (2000), *Speech Synthesis and Recognition*, Proc. ICASSP'2000, Istambul
- [Trancoso 00b] Trancoso, I., et al, (2000), *Da Escrita à Fala - Da Fala à Escrita*
- [Trancoso 04] Trancoso, I., Caseiro, D., (2004), *Spoken Language Processing using Weighted Finite State Transducers*, Proc. SWIM'2004 - Special Workshop in Maui: Lectures by Masters in Speech Processing, Maui
- [Trancoso 98] Trancoso, I, et al, (1998), *Spoken Language Corpora for Speech Recognition and Synthesis in European Portuguese*, Proc. RECPAD'98 - 10th Portuguese Conference on Pattern Recognition, Lisboa

-
- [Turunen 04]** Turunen, M., (2004), *Jaspis - A Spoken Dialogue Architecture and its Applications*, Tese de Doutorado, Universidade de Tampere
- [Vieira 01]** Vieira R., Lima, V., (2001), *Linguística Computacional: Princípios e Aplicações*, JAIA 2001, SBC - Jornadas de Atualização em Inteligência Artificial, Fortaleza
- [Walker 97]** Walker, M., et al, (1997), *PARADISE: A Framework for Evaluating Spoken Dialogue Agents*, Proc. ACL-97 35th Annual Meeting of the Association for Computational Linguistics, pp. 271-280, Madrid
- [Wang 05]** Wang, Ye-Yi, et al, (2005), *Spoken Language Understanding - An introduction to the statistical framework*, IEEE Signal Processing Magazine
- [Weinschenk 00]** Weinschenk, S., Barker, D.T., (2000), *Designing Effective Speech Interfaces*, John Wiley & Sons, Inc.
- [Ynoguti 99]** Ynoguti, C.A., (1999), *Reconhecimento de Fala Contínua Usando Modelos Ocultos de Markov*, Tese de Doutorado, Universidade Estadual de Campinas