Relação entre reconhecimento e compreensão de voz: experimento para análise linguística[1]

Relation between recognition and understanding voice: experiment for linguistic analysis

Edio Roberto Manfio

RESUMO: O objetivo deste artigo é, avaliando o desempenho de um dispositivo que atende a comandos por voz, demonstrar diferenças existentes entre *reconhecimento de voz* e *compreensão de voz*, considerando algumas variantes do Português Brasileiro. Para este estudo, de caráter interdisciplinar, foram consultadas informações sobre o ALiB - Atlas Linguístico do Brasil (CARDOSO, 2014), Fonética e Fonologia (SILVA, 1999) Processamento de Sinais da Fala (YNOGUTI, 1999; HUGO, 1995), o próprio dispositivo (ELECHOUSE, 2014) e resultados de outras pesquisas realizadas pelo autor (MANFIO, 2014; MANFIO; MORENO; BARBOSA, 2014a, 2014b). O procedimento consistiu em gravar três grupos de palavras - comandos - na memória do dispositivo e expô-lo a variantes da cada uma delas, além de outras com grande proximidade fônica. Embora as variantes linguísticas tenham ficado em evidência, o objeto de estudo principal foi o próprio dispositivo. Verificou-se que esse dispositivo realiza apenas o *reconhecimento de voz* - e nunca a *compreensão*. Suspeita-se, ademais, que a maioria dos dispositivos similares no mercado pode apresentar a mesma deficiência.

PALAVRAS-CHAVE: Reconhecimento de voz. Compreensão de voz. Comando por voz. Processamento de Sinais da Fala. ALiB.

ABSTRACT: The purpose of this article is, evaluating the performance of a device that responds to voice commands, demonstrate differences between voice recognition and voice understanding, considering some variants of Brazilian Portuguese. For this study, interdisciplinary nature, were consulted on the ALiB-Linguistic Atlas of Brazil (CARDOSO, 2014), Phonetics and Phonology (SILVA, 1999) Speech Signal Processing (YNOGUTI, 1999; HUGO, 1995), the device itself (ELECHOUSE, 2014) and results of other surveys conducted by the author (MANFIO, 2014; MANFIO; MORENO; BARBOSA, 2014a, 2014b). The procedure

[1] Edio Roberto Manfio é professor grau III na Faculdade de Tecnologia de Garça-SP, Mestre pela Universidade Estadual de Maringá, Doutor pela Universidade Estadual de Londrina e realiza pesquisas interdisciplinares que envolvem as áreas de Linguística e Processamento de Linguagem Natural.

consisted to record three groups of words-commands-in memory of device and expose it to variants of each and other with great phonemic proximity. Although the linguistic variants have been highlighted, the main object of study was the device itself. It was found that this device performs only the voice *recognition* - and never *understanding*. It is suspected that most similar devices on the market can present the same disabilities.

KEYWORDS: Voice recognition. Voice understanding. Voice command. Speech Signal Processing. ALiB.

Introdução

Pedir a uma máquina que realize tarefas apenas com comando por voz é algo presente em muitas obras de ficção científica. No clássico do cinema *Blade Runner* (BLADE Runner, 1982), por exemplo, há uma cena em que o caçador de replicantes Rick Deckard, interpretado por Harrison Ford, dá ordens orais ao computador para que amplie, centralize e enquadre determinada imagem da qual necessita extrair informações relevantes à sua investigação. O computador atende prontamente a todas as suas solicitações até que Deckard consegue finalmente visualizar o que procurava. Embora isso ainda pareça para muitas pessoas um recurso inexistente mesmo em se tratando de século XXI, muito do que foi propagandeado em filmes e livros do gênero há algumas décadas no que diz respeito a comandos por voz[2] já é realidade.

O comando por voz evoluiu tanto nas últimas duas décadas que não são mais necessários equipamentos caríssimos e softwares dedicados e específicos para que ele seja possível. A exemplo disso, há uma infinidade de eletrônicos que operam com esse recurso no dia a dia como os corriqueiros smartphones, sistemas embarcados em automóveis, dispositivos de domótica em residências e os buscadores na internet.

No entanto, embora ambos dependam do processamento de áudio, *reconhecimento de voz e compreensão de voz* não são a mesma coisa e,

[2] Os comandos por voz podem ser atendidos por *reconhecimento de voz* ou por *compreensão de voz*, dependendo da finalidade e/ou robustez dos sistemas empregados. Para o usuário, pouco importa se é um ou outro, bastando apenas que a ordem seja cumprida pela máquina, ou seja, tem que funcionar bem.

na maior parte das vezes, são confundidos pelos usuários pelo fato de estarem intimamente correlacionados. Nesse âmbito, este artigo apresenta um experimento linguístico com um desses dispositivos procurando demonstrar diferenças existentes entre *reconhecimento de voz* e *compreensão de voz* levando em conta algumas variantes do Português Brasileiro.

Processamento de áudio: controle por som, reconhecimento e compreensão

Na época em que o longa-metragem foi lançado, a máquina fictícia utilizada por Deckard poderia ser classificada por linguistas e desenvolvedores como dotada de *reconhecimento* e *compreensão de voz* em função da quantidade de comandos que ela reconhecia e as sub-rotinas de cada um deles. Em outras palavras, enquanto ligar e desligar uma lâmpada de determinado cômodo da casa no âmbito da domótica requer no máximo dois comandos como 'liga' e 'desliga', *ampliar*, *centralizar* e *enquadrar* determinada imagem como solicitou o personagem no clássico do cinema implica especificar cada vez mais um comando dado em primeiro plano: primeiro *amplia*, depois, cada qual ao seu tempo, *centraliza* e *enquadra*. Em outras palavras, os itens lexicais envolvidos multiplicam-se e a complexidade aumenta.

Isso não quer dizer que, fora da ficção científica, sub-rotinas por comandos por voz também não possam ser adaptáveis a uma simples lâmpada do mundo real. Pode-se, por exemplo, operar com nuanças como aquilo que se chama popularmente de 'dimerização'[3] que consiste em dar intensidades diferentes ao brilho da lâmpada em funcionamento. Então, em vez de apenas o insosso e dicotômico 'liga-desliga', há 'liga-menos-menos-

[3] A 'dimerização', como é popularmente conhecida no âmbito da eletrotécnica diz respeito ao uso do *dimmer*, dispositivo elétrico-eletrônico capaz controlar a intensidade de uma lâmpada. O homógrafo 'dimerização' - *dimerization* - encontrável nos dicionários diz respeito a um fenômeno químico em que uma reação que forma um dímero (FERREIRA, 2014).

menos-desliga'. Esse exemplo pode também se aplicar a uma série de equipamentos elétricos em uma residência. Um chuveiro elétrico pode vir equipado com o inusitado 'liga-menos-menos-menos-desliga' que, por uma questão de usabilidade, conforto e precisão de comandos, poderia ser programado para 'liga-morno-quente-maisquente-desliga'. Note-se que agora, como representado na Tabela 01, tanto para a lâmpada quanto para o chuveiro, pode-se operar com cinco comandos reconhecíveis e a quantidade de fonemas e respectivas combinações a serem processados multiplica-se:

Tabela 1 – Conjunto de cinco possíveis comandos por voz

lâmpada	**chuveiro**
liga	liga
menos	morno
menos	quente
menos	mais quente
desliga	desliga

Fonte: o próprio autor

É adequado, portanto, explicar o processamento de áudio de um ponto de vista mais voltado à Linguística, uma vez que o objetivo deste artigo é apresentar um experimento que procure demonstrar diferenças entre *reconhecimento* e *compreensão de voz* e, não necessariamente, extinguir dúvidas sobre o processamento de áudio, cuja complexidade envolvida é extremamente grande e as áreas de conhecimento que se inter-relacionam multiplicam-se com o passar dos anos.

Por mais esse motivo é importante explicar que *reconhecimento* e *compreensão de voz* são coisas diferentes não apenas em termos linguísticos, mas também no âmbito do desenvolvimento tecnológico. Em algumas áreas de engenharia, enquanto este último funda-se no entendimento do sentido, do significado da mensagem como um todo (HUGO, 1995), *reconhecimento* está relacionado basicamente a um único fonema ou conjuntos deles - sílabas e palavras (YNOGUTI, 1999). Dos dois, será dada aqui muito mais ênfase ao conceito de *reconhecimento* e o

equipamento utilizado para o experimento será o *Voice Recognition Module V2* (doravante apenas *V2*). O motivo de sua escolha é o fato de ser utilizado amplamente em aplicações básicas de robótica e domótica, oferecer soluções práticas e baratas para aplicações gerais, não depender de conexões de rede como o buscador por voz do Google e, diferentemente, do *IBM Via Voice*, (MANFIO, 2014) não necessitar de Sistema operacional, além de oferecer suporte para voz humana em geral – não depende de um único idioma. No entanto, também possui grandes desvantagens em relação aos outros dois como se vê a seguir.

A propósito, uma afirmação contida do manual do *V2* sobre o *reconhecimento de voz* poderia fazer parte dos critérios de análise de equipamentos que operam com *comandos por voz*. Lá há informação de que *reconhecimento de voz* se dá apenas quando o dispositivo **sabe exatamente** o que o usuário está dizendo - "Voice recognition is something that knows exactly what you were saying" (ELECHOUSE, 2014). Tal informação, no entanto, estaria muito mais vinculada à *compreensão de voz* uma vez que a máquina precisa ser bem mais robusta em termos de software e hardware para fazer isso. Para dar uma ideia, seria necessário implementar, além de aplicações sobre Fonética, Fonologia, Morfologia e Sintaxe, também outras sobre Semântica lexical, Semântica composicional e Pragmática e padrões discursivos (JURAFSKY; MARTIN, 1999). Um dos exemplos mais modernos desse universo seria *Watson* da IBM, que opera a partir de linguagem natural - evidentemente -, aprendizado dinâmico, geração de hipóteses e é considerado pelos próprios desenvolvedores como 'uma extensão natural daquilo que os humanos podem fazer em seu melhor' (IBM WATSON, 2015). O *Watson* talvez seja o mais próximo daquilo que pode ser considerado hoje como *compreensão por voz* – vai além do *reconhecimento de voz*.

No outro extremo da linha evolutiva de equipamentos que reagem aos sons produzidos por humanos, existem aqueles acionados por som de

palmas - *clappers*[4] - que certamente não foram projetados para operar com *reconhecimento* e *compreensão de voz* e, do ponto de vista linguístico, interessariam pouco neste momento. Com custo significativamente baixo, é possível dotar persianas com esse recurso: uma palma, abrem; duas palmas, fecham - ou vice-versa. Mesmo esses também dependem minimamente de processamento de áudio, ainda que bastante primitivo e simplificado. O circuito eletrônico, nesse caso, deve distinguir entre os estampidos do choque entre as mãos e outros sons pertencentes a uma faixa de frequência e pressão sonora similares como eventuais ruídos, sons de alguns animais, gritos de crianças, gargalhadas ou mesmo enfáticas e espontâneas interjeições como 'ah!', 'eh!', 'uh!' entre outras. Porém, esses dispositivos, embora atendam a comandos por sons, são muito primitivos e não podem ser classificados como dotados dos recursos de *reconhecimento* e *compreensão de voz*[5].

Migrando para sistemas mais sofisticados em relação aos que funcionam com palmas, temos aqueles que atendem a comandos básicos por voz como 'liga' e 'desliga' e que são aplicáveis a quase tudo o que funciona com eletricidade em uma residência: lâmpadas, televisores, ventiladores, condicionadores de ar, exaustores etc. Para esses, o processamento de áudio precisa ser bem mais elaborado porque não são

[4] O *clapper* talvez seja o dispositivo comercial mais conhecido dessa categoria, embora não seja o primeiro da história. Trata-se de um comutador elétrico comandado por som patenteado nos EUA em 1986 que serve para ligar eletrodomésticos: http://en.wikipedia.org/wiki/The_Clapper

[5] Importante salientar nesse momento que, algumas vezes, não há clara distinção entre 'voz' e 'fala' na literatura científica de um modo geral. Em se tratando de Fonética e Fonologia, esses dois termos manifestam-se cada qual em contexto específico, mas, algumas vezes, são equivalentes. Em expressões como 'voz humana' e 'fala humana' pode-se dizer que se trata quase da mesma coisa e alguns linguistas até diriam que não há 'fala' que não seja humana. Quando se distingue, por exemplo, entre um fonema vozeado e outro não-vozeado, é clara a presença das vibrações das cordas vocais em um e não em outro, respectivamente. Não se discute, portanto, sobre fonema 'falado' e fonema 'não-falado'. Em Processamento de Sinais da Fala, 'reconhecimento de voz' e 'reconhecimento de fala' equiparam-se, assim como 'comandos por voz' e 'comandos por fala', ainda que esse último seja bem menos recorrente. De qualquer forma, um mínimo de cuidado com a terminologia deve ser tomado.

apenas sons que os comandam, mas sons minimamente organizados: palavras. É válido lembrar que, em termos de sofisticação e precisão eles são bem diferentes, mas em quantidade de comandos são análogos, como ilustra a Tabela 2:

Tabela 2 – Comparativo entre comandos

	comando por palmas	**comando por voz**
acionar	uma palma	palavra 'liga'
desacionar	duas palmas	palavra 'desliga'

Fonte: o próprio autor

Buscando por mais recursos e por uma funcionalidade um pouco mais adequada, qualquer projetista da área do domótica agora saberia que, caso seu cliente optasse por dotar vários de seus eletrodomésticos com o recurso de comandos por voz, ele teria um problema. Ao entrar em sua sala de estar e dizer 'liga', seriam acionados ao mesmo tempo lâmpada, televisor, ventilador e condicionador de ar. Isso ocorreria porque os sistemas citados estariam dotados de apenas dois comandos programados: 'liga' e 'desliga'.

O passo seguinte seria então especificar: 'ventilador liga' e 'ventilador desliga' ou 'televisor liga' e 'televisor desliga'. Dessa forma, os sistemas, integrados ou não, saberiam qual aparelho ligar. Surge então o primeiro problema que diz respeito ao volume de informação a ser processado. 'Liga' possui bem menos fonemas que 'ventilador liga'. Além disso, necessita de pouco tempo de execução e, por conseguinte, menor espaço para gravação digital, algo bastante crítico quando se trata de processamento de dados de um modo geral. Note-se por esse pequeno detalhe que o processamento de áudio depende de uma série de fatores. Talvez, a um usuário que não se importa em empregar mais e mais palavras a fim de solicitar uma coisa simples, pode ser indiferente dizer "por favor computador, pode por gentileza colocar o ventilador em funcionamento" em vez de "ventilador

liga" tão somente. Aos linguistas, projetistas e desenvolvedores de sistemas isso é um problema.

Considerando o exemplo do ventilador e o dotando de recursos 'dimerizados' como o do chuveiro e o da lâmpada citados há pouco, haveria ainda mais trabalho e muito mais variáveis precisariam ser consideradas se o ventilador tivesse que operar juntamente com o condicionador de ar e a iluminação:

Tabela 3 – Possíveis comandos por voz em uma sala de estar

ventilador	**condicionador de ar**	**iluminação**
ventilador liga	ar liga	luz liga
ventilador menos	ar menos	luz menos
ventilador menos	ar menos	luz menos
ventilador menos	ar menos	luz menos
ventilador desliga	ar desliga	luz desliga

Fonte: o próprio autor.

Tomando então a Tabela 3 como uma nova realidade de domótica, haveria a possibilidade ou de um sistema independente para cada um dos aparelhos ou de um único sistema integrado capaz de controlar a todos por meio de um processamento de áudio mais sofisticado que a opção primeira. Note-se que para [ϖε)τΣιλα∪δο}δεζ∪λιγ□][6] há uma complexidade fonética bastante diversa daquela encontrada em [δεζ∪λιγ□]. Além disso, como comentado anteriormente, é necessário mais memória digital e maior capacidade de processamento. Para esses casos, o *reconhecimento de voz* já basta, ou seja, não é necessária a *compreensão de voz*.

Ainda sobre a Tabela 03, considere-se que no Brasil o advérbio 'menos' varia para 'menas' em diversas regiões, tanto na distribuição por nível de escolaridade quanto por faixa etária, tal como indicam as cartas linguísticas do ALiB, M03E e M03G, respectivamente (CARDOSO et al., 2014). O equipamento tem também de dar conta dessa realidade

[6] Para as transcrições realizadas neste artigo foi utilizada a fonte SILDoulos IPA, disponível gratuitamente na página do *Summer Institute of Linguistics* - SIL International (2012)

morfossintática, ou seja, não se trata apenas de variação diageracional (idade), diassexual (vozes masculinas ou femininas) e/ou diatópica (regional), mas de uma palavra que apresenta, ao menos, um fonema diferente. Importante lembrar que 'menas' ocorreria mais provavelmente em situações em que o nome a ser alterado é feminino como *potência*, *velocidade*, *luz*, *altura*, abertura e raramente para *vento*, *volume*, *frio*, *brilho* entre outros – pouco possível que ocorra mentalmente uma construção como 'menas frio' para um aparelho de ar condicionado, por exemplo.

Deve-se salientar ademais que, até esse ponto da discussão, ainda não foram expostos os problemas dos ruídos de ambiência, responsáveis por muitas falhas em dispositivos comandados por voz, além das diferenças existentes entre as vozes das pessoas, como timbre, potência sonora, entonação, prosódia entre outros, assuntos que requerem pesquisas à parte.

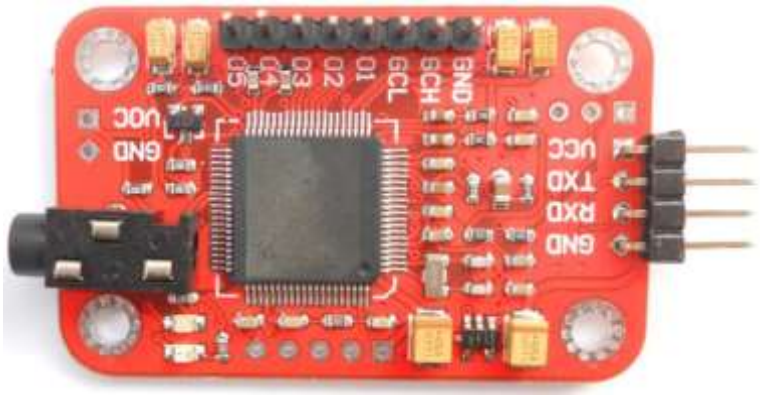
O experimento

Os exemplos de domótica elencados até então são apenas alguns dos muitos empregos do recurso de comando por voz desde sua gênese até a atualidade, mas já possibilitam visualizar uma parte da problemática envolvida entre a idealização de um sistema - prancheta - e sua plena operacionalidade - usuário final e variantes linguísticas correspondentes.

O experimento proposto para o presente estudo tem a ver com domótica, mas pode ser ampliado para outros fins, como disposto adiante. Trata-se basicamente da utilização e aplicação de um dispositivo facilmente encontrável à venda em lojas do ramo - virtuais ou físicas - e é destinado a desenvolvedores, projetistas, hobistas e, claro, pesquisadores.

Distribuído pela Elechouse (2014), o *V2* é dotado de um processador específico para esse fim. Ele funciona por interface USB, consome pouca energia e apresenta relativa facilidade quanto à programação e configuração. A Figura 1 dá uma noção de sua aparência física.

Figura 1- Voice Recognition Module V2



Fonte: Elechouse (2014)

O que ele faz é bastante simples ao observador, embora seja sabidamente a junção de uma série de tecnologias que vem sendo desenvolvidas nas últimas décadas. Sua função é gravar e armazenar três grupos de comandos por voz, cada qual com cinco peças[7] de até 1.3 segundos de extensão em tempo de execução. Após gravados e armazenados, os comandos podem acionar praticamente qualquer dispositivo elétrico por meio de suas portas lógicas.

Tabela 4 – Amostra de possíveis grupos com 5 peças (comandos) cada

grupo 01	**grupo 02**	**grupo 03**
Amarelo	primeiro	direita
Vermelho	segundo	frente
Verde	terceiro	para
Azul	quarto	trás
Branco	quinto	esquerda

Fonte: o próprio autor

Por limitações técnicas do próprio *V2* e pela necessidade de vários circuitos adicionais, a utilização de comandos como os exibidos na Tabela 03 serão feitos em outro experimento. Neste, a opção foi por algo mais

[7] Para o caso deste equipamento, o comando por voz diz respeito a uma palavra ou expressão no idioma correspondente. No manual do *V2*, cada comando por voz é denominado 'peça' e o que mais importa é o tempo de duração. A palavra 'geossociolinguística', por exemplo, pronunciada de modo espontâneo, pode não ser uma 'peça' válida por ultrapassar o máximo de 1.3 segundo estabelecidos pelo fabricante.

simples - eletronicamente falando - pois quando se trata do acionamento de equipamentos que operam com nuanças ajustáveis como ventiladores, chuveiros, cadeiras ou camas reclináveis entre outros o projeto fica mais complexo e oneroso.

Portanto, para testá-lo, em um dos grupos forma gravados os nomes de cinco cores: vermelho, verde, azul, amarelo e branco. Essas palavras têm algumas vantagens: são relativamente curtas, conhecidas por todas as pessoas - mesmo em tenra idade - e acionam no experimento lâmpadas com as cores respectivas, o que visualmente é incontestável – salvo para algumas pessoas com problemas de visão.

Além disso, a complexidade linguística envolvida em 'vermelho, verde, azul, amarelo e branco' - cinco palavras - é maior que em 'liga-menos-menos-menos-desliga' - apenas três palavras/comandos. Para esse último caso, é válido notar que, embora para dimerizar seja necessário repetir sequencialmente a palavra 'menos' mais de uma vez, a função do reconhecedor é apenas distinguir entre 'liga' e 'menos'. Quem vai decidir se é ou não para dimerizar é o software, responsável por calcular o intervalo de tempo entre as palavras pronunciadas.

É claro que, fora de um contexto experimental e científico como esse, acionar cores pode ainda ser considerado pouco útil em termos de domótica, embora a cromoterapia[8] ocupe espaço considerável entre as terapias alternativas. Porém, como foi dito, as palavras poderiam ser outras e em vez de lâmpadas coloridas, elevadores - (grupo 02/Tabela 4), veículos sobre rodas (grupo 03/Tabela 4) ou eletrodomésticos (Tabela 3) poderiam ser acionados, substituições relativamente fáceis de serem feitas no âmbito da eletroeletrônica. Em outras palavras, a eletricidade não questiona o que ela está acionando, ela simplesmente é encaminhada até seu destino por meio de condutores e semicondutores e lá faz o seu trabalho.

[8] A título de curiosidade, a Cromoterapia está entre as terapias alternativas reconhecidas pela Organização Mundial de Saúde desde a década de 1970. Uma das obras mais consultadas nessa área é a de René Nunes denominada *Compêndio Científico de Cromoterapia*.

Os inconvenientes deste experimento são os mesmos elencados há pouco quando se discutiu sobre equipamentos que funcionam por comandos por voz: o ruído de ambiência e as diferenças existentes entre as vozes. O próprio manual (ELECHOUSE, 2014) do *V2* deixa claro que possivelmente ele não atenderá aos comandos de um amigo, mesmo pronunciando as mesmas palavras. Quanto aos ruídos de ambiência, eles devem ser minimizados ao máximo durante a gravação e execução das palavras, caso contrário o processamento de áudio fica claramente comprometido.

Além do *V2*, as lâmpadas também contaram com alguns cuidados técnicos e todas foram montadas em um gabinete – Figura 2 - com revestimento opaco e branco que permite visualizar suas cores apenas quando estão ligadas. Cada uma delas tem uma baia própria e isolada que impede a contaminação entre as cores quando estão acionadas simultaneamente e potência suficiente apenas para permitir boa resolução cromática mesmo em ambiente iluminado. O *V2* também foi alojado no interior do gabinete e as conexões com o computador eram feita por meio de aberturas na tampa traseira.

Figura 2 - Gabinete cromático

Fonte: Elechouse (2014)

Tomadas todas as precauções básicas que pudessem permitir melhor funcionamento do experimento, os testes começaram a ser feitos. Primeiramente, procedeu-se com a gravação de cada uma das peças, como são denominadas no manual do *V2* (ELECHOUSE, 2014). Então vermelho, verde, azul, amarelo e branco foram pronunciadas ao microfone no *V2* do modo mais natural possível, embora seja um fato que o grau de automonitoramento é sempre presente e significativo quando a pessoa sabe que está sendo gravada, independentemente de ser um informante ou o próprio pesquisador.

De qualquer forma, o software que controla as gravações está programado para solicitar sempre duas amostras de cada peça, então as compara e, caso sejam equivalentes, finaliza o registro e automaticamente permite a gravação da peça seguinte. Analogamente, é quase como o linguista durante uma transcrição: sempre ouve mais de uma vez. Note-se por esse detalhe que o *V2* tem critérios interessantes quanto à sua

concepção no quesito funcionalidade uma vez que não permite continuar a gravação se o usuário também não adotar critérios mínimos quanto à pronúncia.

Gravadas todas as peças, o *V2* foi colocado à prova. Devidamente conectado ao gabinete com as lâmpadas, funcionou muito adequadamente para cada uma das cores e as respectivas lâmpadas foram acionadas e desacionadas com bastante êxito. As transcrições aproximadas das peças são [ϖε}∪μελιY], [∪ϖε}δZι], [α∪ζυω], [αμα∪PEλY] e [∪βPα)κY]. A Tabela 5 apresenta o modo como os comandos se distribuem. Pode-se nela perceber que, embora o comando para ligar e desligar seja o mesmo - a própria palavra - pode-se tranquilamente dizer que há cinco comandos para o gabinete, do mesmo modo como há para cada um dos eletrodomésticos listados na Tabela 3.

Tabela 5 – Relação dos comandos e cores utilizados no experimento

cores	**para ligar**	**para desligar**
vermelho	vermelho	vermelho
verde	verde	verde
azul	azul	azul
amarelo	amarelo	amarelo
branco	branco	branco

Fonte: o próprio autor

Algumas variantes também foram experimentadas, principalmente das palavras 'vermelho' e 'verde' por possuírem maior incidência de variação no Brasil: o /R/ em coda silábica em interior de palavra tem essa característica – Cartas Fonéticas F04 C5 e C6 (CARDOSO et al., 2014). Mesmo com as típicas limitações de um falante que não é nativo da região correspondente, o sistema respondeu bem a ponto de as falhas serem desconsideradas em análise final. Isso não quer dizer que o sistema opere com *compreensão de voz*, e sim comprova que o método de processamento de sinal certamente leva em conta o tempo de gravação e pressão sonora de cada peça. Algumas das variações simuladas que obtiveram êxito como comandos estão na Tabela 6.

Tabela 6 – Relação dos comandos e cores utilizados no experimento

peça gravada	**variantes simuladas como comandos**				
[ɷε}∪μελιY]	[ɷEΞ∪μελιY]	[ɷEη∪μελιY]	[ɷεΞ∪μελιY]	[ɷεη∪μελιY]	[ɷεP∪μελιY]
[∪ɷε}δZι]	[∪ɷεΞδι]	[∪ɷEηδι]	[∪ɷεΞδZι]	[∪ɷεηδZι]	[∪ɷεPδZι]

Fonte: o próprio autor

No entanto, embora tenha acionado a lâmpada vermelha com a peça de áudio [ɷε}∪μελιY] assim como ocorreu com todas as outras lâmpadas e respectivas peças, também acionou a lâmpada vermelha com as execuções [πε)∪τελιY] e [φε∪δελιY]. O mesmo se aplicou à lâmpada amarela, ou seja, foi facilmente acionada com a peça de áudio [αμαPEλω], mas quase igualmente com as execuções [Σι∪νEλY] e [μα∪}μEλY]. Da mesma forma, o *V2* 'entendeu' [α∪ζυω] com [τα∪τY] e [∪βPα)κY] com [∪μα)κY], tal como demonstra Tabela 7. Está claro que, adotados os critérios adequados, ele funciona bem com as peças previamente gravadas, mas também aceita os comandos quando a execução tem fonemas, prosódia e pressão sonora similares ou aproximados. Em termos linguísticos mais técnicos, quando há pares mínimos como 'faca/vaca' ou 'sumir/zunir' em que estão em jogo Contraste em Ambiente Idêntico – CAI – ou Contraste em Ambiente Análogo - CAA (SILVA, 1999) – respectivamente, o sistema vai mal.

Tabela 7 – Comparativos entre as peças de áudio e execuções aproximadas ou similares

Peça	**execuções aproximadas**	
[ɷε}∪μελιY]	[πε)∪τελιY]	[φε∪δελιY]
[∪ɷε}δZι]	[∪λε}δZισ]	[∪ɷε}τΣι]
[α∪ζυω]	[α)∪νY]	[τα∪τY]
[αμα∪PEλY]	[Σι∪νEλY]	[μα}∪μEλY]
[∪βPα)κY]	[∪μα)κY]	[∪βα)κY]

Fonte: o próprio autor

A peça [∪ϖε}δZι], no entanto, tal como mostra a Tabela 7, sofreu menores problemas porque palavras em português com disposição de fonemas, prosódia e pressão sonora similares ou aproximados são bastante raras – elas muitas vezes coincidem com palavras que rimam. Porém, isso não significa que o *V2* também não possa entender [∪ϖε}δZι] com alguma outra eventual execução inusitada como [∪λε}δZισ] [9] ou a locução [∪ϖε}τΣι] [10], pouquíssimo utilizada na fala atualmente.

4 O experimento em campo

Originalmente, o gabinete cromático foi idealizado e montado tão somente para demonstrar a funcionalidade do *V2* em termos linguísticos. Isso seria feito apenas junto às bancas de professores em algumas versões do SEDATA na UEL como extensão do projeto de pesquisa e/ou na presença de outros pesquisadores interessados. No entanto, devido à sua versatilidade nos quesitos visual - não há como não vê-lo - e portabilidade - dimensões compatíveis com malas de vários automóveis - ele começou a viajar. Ao longo de todo o segundo semestre de 2014, eventos científicos, palestras, reuniões, divulgação de vestibular a até uma feira de inovação tecnológica cederam espaço ao experimento que acabou surpreendendo - o pesquisador - pela capacidade de prender a atenção das pessoas de um modo geral.

Em outras palavras, ele interessantemente serviu para divulgar várias coisas ao mesmo tempo: a presente pesquisa, o projeto de Regime de Jornada Integral junto à Fatec Garça, o projeto do Centro de Pesquisa Tecnológica de Garça, o PPGEL e, claro, o ALiB.

Durante suas andanças, muitas dúvidas foram tiradas, muita curiosidade foi suscitada, várias ideias surgiram, um significativo número de pessoas testou o experimento emprestando suas vozes e certa quantidade de conclusões que agora fazem parte do corpo deste artigo puderam ser

[9] Trata-se da 2ª pessoa do plural verbo 'ler' no infinitivo pessoal
[10] 'Ver-te'

tiradas a partir das observações linguísticas. Entre os fatos que mais chamaram atenção está o de que ficou evidente o cuidado que os desenvolvedores desse tipo de equipamento devem ter com as variantes diassexuais: quando as peças eram gravadas com vozes masculinas, ele funcionava mal com as mulheres e vice-versa. Outro fato interessante tem a ver com o desempenho do equipamento quando exposto a uma variedade do português que não foi estudado pelo ALiB. Durante a *V Feira de Ideias e Inovação* promovida pela AINTEC-UEL o experimento foi testado por um angolano, que interessantemente teve êxito com todas as cores, exceto o verde – foi o único homem em todo o semestre que não conseguiu acionar este comando.

Embora possam parecer detalhes pouco significativos a outras áreas de pesquisa, esses dois fatos certamente tem potencial de gerar produtivos estudos e arrisca-se a dizer que, no mínimo, poderiam ser temas para dissertações de Mestrado em que duas respostas poderiam ser buscadas: (i) se os equipamentos estão preparados para dar conta das diferenças entre vozes masculinas e femininas (ii) se há de fato, tecnicamente falando, algo de peculiar na palavra 'verde' considerando-se todas as distribuições de variantes no Brasil (diassexual, diageracional, etc). Para essa última dúvida teórica, é válido lembrar que o buscador por voz do Google, inversamente, só identifica 'verde' para palavras com proximidade fônica[11].

Considerações finais

O objetivo de apresentar um experimento linguístico que procurasse demonstrar diferenças existentes entre *reconhecimento* e *compreensão de voz* por meio de um dispositivo dotado da capacidade de processamento de áudio foi feito com bastante êxito. Verificou-se, entre outros detalhes que, fica clara a diferença entre *reconhecimento de voz* e *compreensão de voz* e não há dúvidas de que entre as tarefas do *V2* não se encontra a de

[11] Um estudo sobre comandos de voz on-line estava sendo conduzido quando do fechamento deste artigo.

compreensão - ele não 'sabe' o que foi dito. Em outras palavras, comporta-se – na melhor das hipóteses - como um estrangeiro tentando identificar com seu ouvido ainda pouco treinado às nuanças entre palavras com proximidade fônica. Desconsiderando-se a melhor das hipóteses, seu desempenho está bastante próximo daquele de personagens surdos de quadros humorísticos. Sistemas mais sofisticados que operam com linguagem natural escrita identificam há muito os campos semânticos – entre outras coisas – situação em que 'chinelo' é facilmente descartado por não pertencer ao grupo 'cores'. Os buscadores na internet, entre muitos, usam esse recurso paralelamente a outros (MANFIO; MORENO; BARBOSA, 2014a, 2014b).

Rick Deckard, em *Blade Runner*, provavelmente teria acertado seu equipamento com uma marreta se obtivesse esses resultados, pois além do transtorno ocasionado pelo mau funcionamento, sua investigação ficaria prejudicada. No filme, como comentado anteriormente, a máquina parece não ter dúvidas quanto ao que o detetive pronuncia pois atende ao comando por voz com a respectiva ação, mesmo levando em conta que há apenas uma fração de segundo entre uma e outra. Considerando-se um continuum evolutivo que parte dos equipamentos acionados por som - como os *clappers* - até aqueles que realizam a *compreensão de voz*, o equipamento fictício do filme está claramente mais próximo deste último recurso: a *compreensão*.

Nesse mesmo continuum, o experimento aqui apresentado parece deixar claro que, embora tenha ocorrido o *reconhecimento de voz* seguido - por frações de segundo - da execução do *comando por voz*, a *compreensão de voz* é inexistente. O próprio *reconhecimento* é precário uma vez que há identidade entre palavras que possuem certa similaridade fonética, mas bastante divergentes entre si em quesitos semânticos. Ainda sugere que, possivelmente, dentre os muitos equipamentos disponíveis no mercado que operam com *reconhecimento* e *comando por voz*, a grande

maioria pode apresentar a mesma falha pois os métodos utilizados no processamento do sinal de áudio são similares.

Referências

BLADE Runner. Direção: Ridley Scott. Produção: Michael Deeley. Roteiro: Hampton Fancher, David Peoples. Música: Vangelis. 1982. EUA.

CARDOSO, Suzana Alice Marcelino da Silva et al. *Atlas Linguístico do Brasil*: Introdução. Londrina: Eduel, 2014. v.2.

ELECHOUSE. *Voice Recognition Module V2 Manual*. Disponível em: <http://www.elechouse.com/elechouse/images/product/Voice%20Recognition%20Module/Manual.pdf>. Acesso em: 2 jul. 2014.

FERREIRA, Aurélio Buarque de Holanda. *Dicionário Aurélio Eletrônico da Língua Portuguesa*. Curitiba: Editora Positivo, 2014. 1 CD.

HUGO, Marcel. *Uma interface de reconhecimento de voz para o sistema de gerenciamento de central de informação de fretes*. 1995. 60 f. Dissertação (Mestrado em Engenharia de Produção) – Universidade Estadual de Santa Catarina. 1996.

IBM WATSON. *What is Watson*? Disponível em: <http://www.ibm.com/smarterplanet/us/en/ibmwatson/what-is-watson.html>. Acesso em: 13 jan. 2015.

JURAFSKY, Daniel; MARTIN, James H. *Speech and language processing:* an introduction to Natural Language Processing, Computacional Linguistics and Speech Recognition. Prentice Hall, Englewood Cliffs, New Jersey, 1999.

MANFIO, Edio Roberto. O ALiB no *IBM Via Voice*: Pesquisa geossociolinguística associada a aplicativo de reconhecimento de fala. In: XIII SEDATA - SEMINÁRIO DE DISSERTAÇÕES E TESES EM ANDAMENTO, 13., 2014, Londrina. *Caderno de Resumos...* Londrina: UEL, 2014. v.1. p. 35.

MANFIO, Edio Roberto; MORENO, Fabio Carlos; BARBOSA, Cinthyan Renata Sachs Camerlengo de. Tecnologia Interativa Conversacional sobre Assuntos Linguísticos - Tical: Linguagem e Significação. In: SEMINÁRIO DE ESTUDOS SOBRE LINGUAGEM E SIGNIFICAÇÃO, 9., SIMPÓSIO DE LEITURA DA UEL "Convenções e Ousadias da Linguagem", 9., 2014, Londrina. *Caderno de Resumos...* Londrina: UEL, 2014a. p. 54-55.

MANFIO, Edio Roberto; MORENO, Fabio Carlos; BARBOSA, Cinthyan Renata Sachs Camerlengo de. Professor Tical: Robô de Conversação sobre Dialetologia e Geossociolinguística. In: CIDS - CONGRESSO INTERNACIONAL DE DIALETOLOGIA E SOCIOLINGUÍSTICA – Variação, Atitudes linguísticas e Ensino, 3., 2014, Londrina. *Caderno de Resumos...*. Londrina: UEL, 2014b. p. 48.

SIL INTERNATIONAL. *Products*. Disponível em: <http://www.sil.org/>. Acesso em: 3 jun. 2012.

SILVA, Thaïs Cristófaro. *Fonética e Fonologia do Português*: roteiro de estudos e guia de exercícios. São Paulo: Contexto, 1999.

YNOGUTI, Carlos Alberto. *Reconhecimento de Fala Contínua Usando Modelos Ocultos de Markov*. 1999. 138 f. Tese (Doutorado em Engenharia Elétrica) – Unicamp. Campinas, 2000.