Sistema de Conversão Português/Libras

Leandro Sabino da Silva [1] e Raissa Bezerra Rocha [1,2]

[1]Departamento de Engenharia Elétrica, Universidade Federal de Sergipe - leandrosabino08@gmail.com
[1,2]Instituto de Estudos Avançados em Comunicações - raissa@iecom.com.br

***Resumo*—Este artigo apresenta o desenvolvimento de um sistema de conversão entre os idiomas Português-Libras, desenvolvido para ser utilizado como ferramenta de auxílio a ouvintes e surdos no processo de comunicação entre as modalidades oral e gesto-visual. Por meio da técnica *HMM*, utilizada para treinar e reconhecer os sinais de fala, é possível implementar o reconhecedor necessário ao sistema de conversão. Dessa maneira, os sinais de fala são convertidos para seus respectivos sinais em Libras por meio de um algoritmo de mapeamento. Para avaliar o sistema é necessário verificar a correspondência entre cada sinal de fala e o seu símbolo em Libras, que se limita a letras do alfabeto em Libras. Os resultados do sistema de conversão proposto mostram-se promissores.**

***Palavras-Chave*—Sistema de conversão, Reconhecedor de fala, *HMM*.**

I. Introdução

Desde as telas do cinema à vida real, a ideia de tradutores universais não é recente. Usado em filmes desde a saga *Star Trek* até O Guia do Mochileiro das Galáxias, os tradutores de línguas orais tem desempenhado um importante papel nos sistemas de comunicação. Essas ferramentas são criadas e aprimoradas para auxiliar as pessoas. O Google Tradutor, por exemplo, tem contribuído para uma melhor experiência do usuário nas traduções de línguas orais. Ele trabalha baseando-se na tradução automática por análise estatística e detecta padrões em textos bilíngues criados por tradutores humanos e, dessa forma, determina qual a tradução considerada mais adequada para o texto que lhe é apresentado [1].

No entanto, ainda há pouca atenção no que diz respeito a tradutores de línguas oral para gesto-visuais; o que estabelece um desafio para as modalidades de línguas de sinais na inclusão dos tradutores tradicionais.

Segundo o último Censo realizado pelo IBGE (2010), estima-se que há 9,7 milhões de pessoas com algum grau de deficiência auditiva. Desses, 2,147.366 milhões apresentam deficiência auditiva severa, situação em que há uma perda entre 70 e 90 decibéis (dB) [2].

A Libras (Língua Brasileira de Sinais) é o segundo idioma oficial do Brasil de acordo com a Lei 10.436 de 24 de abril de 2002. Contudo, a difusão da língua entre ouvintes é precária e não atinge todos os espaços.

Por essas razões é proposto um sistema de acessibilidade a pessoas surdas, visto a condição da população com algum grau de surdez no Brasil, para promover uma maior interação entre surdos e leigos com a Libras e estreitar a comunicação. Esse sistema deve reconhecer o que está sendo falado e realizar a conversão para o idioma Libras. A utilização como ferramenta de auxílio na comunicação se estende a diversos ambientes, como restaurantes, shoppings e aeroportos; onde a comunicação rápida e simples entre surdos e ouvintes se faz necessária.

É necessário investigar como as línguas de sinais se organizam e quais as particularidades da Língua Brasileira de Sinais, dado que não existe língua de sinais universal. Dessa maneira, assim como não existe uma língua oral única, também não existe unicidade nas línguas de sinais, visto que as mesmas possuem mecanismos morfológicos, sintáticos e semânticos. As línguas de sinais distinguem-se das línguas orais, principalmente, pelo seu canal de comunicação visual-espacial [3] [4].

O que nas línguas escritas e orais é chamado de palavra, nas línguas de sinais chama-se de sinal. Analogamente às línguas orais que possuem articulações dos sons para formar fonemas, as línguas de sinais possuem pontos de articulações que são expressados por toques no corpo do usuário da língua ou no espaço neutro. Por esse motivo, é relevante observar os cinco parâmetros que compõem a língua de sinais [3] [4]:

- Configuração de Mãos (CM): É a representação inicial de como está a mão dominante.
- Ponto de Articulação (PA): Lugar onde a mão configurada iniciará o sinal.
- Movimento (M): Deslocamento da mão no espaço. Há sinais estáticos que não requerem movimentação.
- Orientação ou Direcionalidade (O/D): Direção que o sinal irá tomar no espaço.
- Expressão Facial e/ou Corporal (EF/C): Existem sinais que necessitam de expressões para dar maior entendimento e até ênfase no contexto das frases sinalizadas. Para isso, usa-se expressões (faciais e/ou corporais) que podem indicar desde intensidade até suavidade nas palavras.

Considerando os aspectos pertinentes no estudo das línguas orais e de sinais, é possível estabelecer a conexão entre ambas as línguas. O sistema de conversão entre idiomas realiza o mapeamento dos sinais de fala para os sinais em Libras que o representam por meio de um algoritmo que estabelece essa conexão. Dessa maneira, o algoritmo é capaz de identificar qual fonema o locutor alimenta na entrada do sistema e retorna uma imagem que representa o sinal em Libras correspondente.

Esse trabalho está organizado da seguinte forma: a Seção II descreve o sistema proposto para a conversão entre idiomas e o sistema de reconhecimento de fala, a Seção III apresenta os

Recebido em julho de 2019. Aceito em outubro de 2019

resultados do reconhecimento de fala e do mapeamento dos sinais em Libras e a Seção IV traz as considerações finais.

II. Sistema Proposto

É apresentado na Figura 1 o diagrama de blocos construído para guiar o desenvolvimento do sistema de conversão.

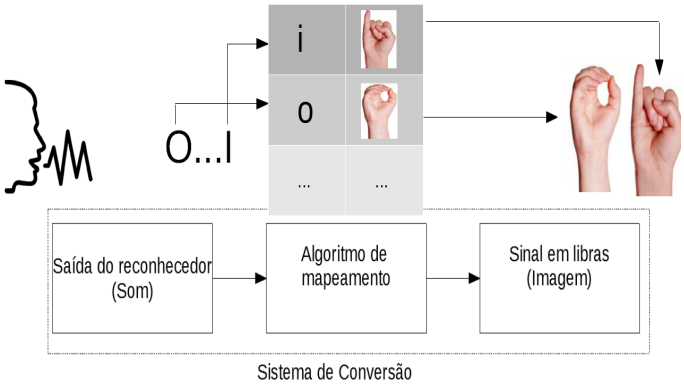

Figura 1: Diagrama de blocos do sistema de conversão.

Após a aquisição do sinal gerado pelo reconhecedor de fala, mapeia-se o sinal para gerar as imagens que correspondem àquele fonema. Esse processo final utiliza um algoritmo de busca para associar fonema à imagem.

Nesse processo, cada palavra transcrita deve ser segmentada sequencialmente em suas letras para a formação dos sinais em Libras. Desse modo a palavra formada deve ser exibida a partir desse sequenciamento de sinais do alfabeto Libras.

A. Reconhecedor de Fala

O principal meio de comunicação do homem é a fala [5]. O emprego da fala na comunicação humana baseia-se em uma sintaxe de léxicos e nomes de um vocabulário. Define-se léxico como uma biblioteca de palavras disponíveis numa língua. A combinação de vogais e consoantes gera os fonemas [21] [22].

O reconhecimento de fala permite que computadores equipados com microfone interpretem a fala humana para transcrição ou comandos [6]. Para isso, mapeia-se um sinal acústico em um conjunto de palavras; isto é, converte-se um sinal acústico em sua representação ortográfica [7] [8] [9]. Durante esse processo, o reconhecimento gera uma sequência de observações que formam a sequência de vetores de características acústicas. No reconhecimento, a sequência das observações da elocução em teste é aceita como verdadeira se possuir alguma medida de verossimilhança acima de um limiar estipulado [10].

O processamento de um sistema de reconhecimento de fala sofre interferência de algumas variabilidades, tais como:

- Variabilidade Fonética: a menor unidade fonética que compõem as palavras são bastante sensíveis ao contexto pois podem modificar o sentido das sentenças. A exemplo da frase 'nada mais é...' que pode ser pronunciada e compreendida como 'nada mais zé' [7] [13] [8].
- Variabilidade Acústica: podem resultar de mudanças no ambiente assim como da posição e características do transdutor (microfone) [7] [13] [8].
- Variabilidade Intra-Locutor: podem resultar de mudanças do estado físico/emocional dos locutores, velocidade de pronúncia ou qualidade de voz [7] [18] [20] [8].
- Variabilidade Entre-Locutor: originam-se da variação linguística da língua nas diversas regiões do país, assim como do tamanho e forma do trato vocal [7] [20] [13] [8].

Entre os principais desafio a serem trabalhados na etapa de reconhecimento de fala, destacam-se:

- Dificuldade para remover barulhos e ruídos.
- Definir com precisão o início e o fim de uma palavra.
- Palavras que possuem o mesmo som, porém com significados diferentes(homônimos).
- Estrutura gramatical e semântica das palavras.

O sistema de reconhecimento de fala tem por objetivo identificar qual a palavra ou frase foi pronunciada pelo locutor. Dessa forma, o reconhecimento de padrões precisa de uma fase de treinamento e outra de reconhecimento a fim de, inicialmente, gerar uma base de dados de treinamento (fonemas) como modelos de referência para o que pretende ser reconhecido. Na etapa de reconhecimento, os modelos obtidos da fase de treinamento são usados para comparação cuja regra de decisão estipula o que mais se assemelha àquele padrão [15] [23] [24] [7] [11] [12].

Durante a fase de reconhecimento, o sistema utiliza um modelo estatístico de distribuição conjunta $P(W|X)$ entre a sequência de palavras pronunciadas W e a sequência de informações acústicas observadas X que foram geradas durante a fase de treinamento. Dessa maneira o sistema se encarrega de procurar uma estimativa $\widehat{W}$ da sequência de palavras pronunciadas a partir da evidência acústica observada X [15] [8] [25].

Na etapa de decodificação, o reconhecedor de fala usa a Equação 1 para determinar a sequência mais provável dada a sequência observada na entrada; em que X é uma sequência de vetores acústicos e $\widehat{W}$ a sequência de palavras. Cada palavra é convertida em uma sequência de fonemas e para cada fonema há um modelo estatístico que o corresponde [15].

$$\widehat{W} = arg_w max[P(W|X)] = arg_w max[\frac{P(W)P(X|W)}{P(X)}]. \quad (1)$$

Nesse contexto, os modelos ocultos de Markov ou HMM (*Hidden Markov Models*) são utilizados por apresentarem uma maior robustez ao sistema, tendo em vista que são capazes de modelar as variabilidades acústicas e temporais dos sinais de voz [7] [10] [11] [13].

As etapas do reconhecedor incluem: processamento do sinal de voz que compreendem as etapas de pré-ênfase e segmentação do sinal de voz, extração de características,

construção do modelo acústico e decodificação [15]; consoante a Figura 2.

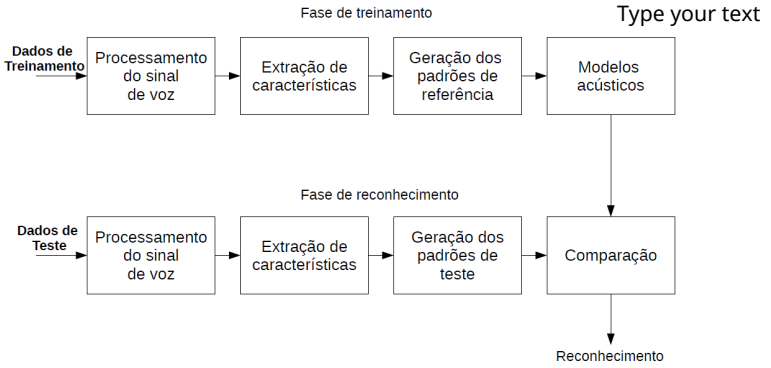

Figura 2: Diagrama de blocos de um sistema reconhecimento de fala [15].

1) Pré-Ênfase e Segmentação: Uma das características que torna o sinal de voz vulnerável a ruído é a baixa amplitude das componentes espectrais do sinal, tendo em vista que sua energia concentra-se nas componentes de baixas frequências. Devido a isso, é necessário pré-enfatizar o sinal passando-o por um filtro de primeira ordem $L(z)$ para acentuar as componentes de frequências mais altas e deixar seu espectro mais plano [5] [15].

$$L(z) = 1 - a_p z^{-1}. \tag{2}$$

Usualmente, o valor de a_p (fator de pré-ênfase) está em torno de um. Neste trabalho utilizou-se um $a_p = 0,95$. Portanto, a pré-ênfase é dada pela Função 3 que relaciona a saída da pré ênfase $s_p(n)$ com a entrada $s(n)$ do sinal original [23] [6] [5] [15].

$$s_p(n) = s(n) - 0,95s(n-1). \tag{3}$$

Na etapa seguinte, é preciso segmentar o sinal de voz em janelas de duração definida, desde que se respeite a característica quase estacionária do sinal de voz [15].

Por isso, foi definida uma janela de *Hamming* de 25ms para a segmentação do sinal por se mostrar mais eficiente que as demais janelas de segmentação, assim como observado em [16] e [15]; definida pela Equação 4.

$$J(n) = \begin{cases} 0,54 - 0,46\cos[2\pi n/(N_A - 1)], & 0 \leq n \leq N_A - 1 \\ 0, & \text{caso contrário} \end{cases} \tag{4}$$

2) Extração de Características: Extrai-se as informações mais relevantes do sinal de voz. Os coeficientes MFCC's (*Mel Frequency Cepstral Coeficients*) atendem ao requisito por serem capazes de representar o sinal de voz baseado no ouvido humano [5] [15] [16]. A Figura 3 ilustra a obtenção desses coeficientes.

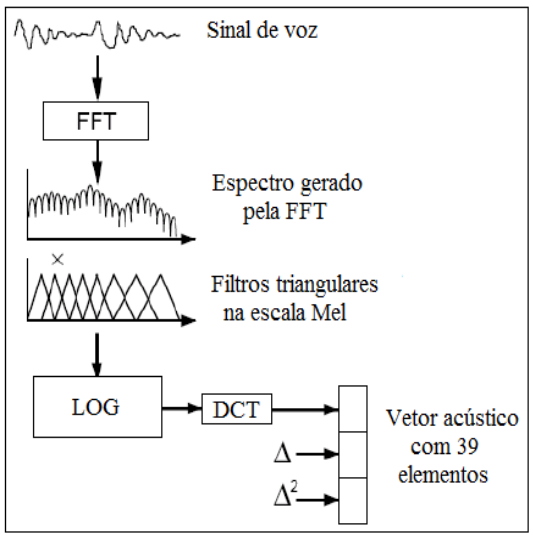

Figura 3: Extração de Características [8].

Para cada janela do sinal de fala, aplica-se uma FFT (*Fast Fourier Transform*) para obtenção de seu espectro de frequências. Em seguida submete-se o espectro a um conjunto de filtros triangulares na escala Mel que resulta na atenuação das componentes de altas frequências pois estas não contribuem para uma boa representação do comportamento do ouvido humano. Obtém-se a compressão logarítma seguida de uma DCT (*Discrete Cosine Transform*) para diminuir a correlação entre elementos do vetor. A seguir, acrescenta-se aos coeficientes MFCC's as suas derivadas de primeira ordem que representam a velocidade de variação do espectro mel-cepstral e as derivadas de segunda ordem que representam a sua aceleração. Ao fim desse processo, obtém-se um vetor de 39 parâmetros por janela [8] [15] [11].

3) Modelo Acústico: A partir das características extraídas da etapa anterior, é possível criar um modelo matemático que represente cada segmento da fala que, para este trabalho, é tratado a nível fonético. Devido às fontes de variabilidade do sinal de voz, é necessário submeter o sistema ao treinamento de um modelo que melhor generalize a descrição de um fonema a partir de um extenso conjunto de sentenças faladas pela maior diversidade de pessoas possíveis [17] [15].

Dada uma sequência de palavras W, o modelo calcula a verossimilhança com a sequência de vetores X dada; $P(X|W)$. Cria-se modelos acústicos para cada fonema da língua e os mesmos procuram, a partir do vetor que representa o sinal sonoro, deduzir qual a sequência de fonemas que gera aquele vetor [15] [8] [25] [13]. Com intuito de facilitar a manipulação matemática da verossimilhança, é comum utilizar a função de log-verossimilhança negativa, que equivale a aplicar a função logarítmo e transformar o sinal. Dado que o valor numérico da verossimilhança é menor que um, o logaritmo desse valor é negativo. Assim, a transformação do sinal garante que a função log-verossimilhança negativa seja um número positivo e em uma escala mais adequada. Dessa maneira, a hipótese

que tem a maior verossimilhança é aquela com menor log-verossimilhança negativa [17] [19]. Nas Tabelas I e II a quarta coluna refere-se ao cálculo da função log-verossimilhança negativa .

Na literatura é comum ver que a modelagem acústica, para cada fonema caracterizado por um HMM, é obtida a partir dos coeficientes mel-cepstrais. Dessa forma, pode-se obter um HMM composto pela união dos modelos de fonemas a fim de formar palavras [15] [8].

Esse monofone contém uma topologia do tipo esquerda-direita (*left-right*) em que três estados são emissores de símbolos com uma dada distribuição de probabilidade enquanto que os estados de entrada e saída não emitem símbolos. Apesar de não emitir símbolos, esse último estado é responsável pela união entre HMMs para dar origem a HMMs formados a partir da união de fonemas que por sua vez dá origem às palavras; conforme ilustrado na Figura 4 [11] [24] [8] [13].

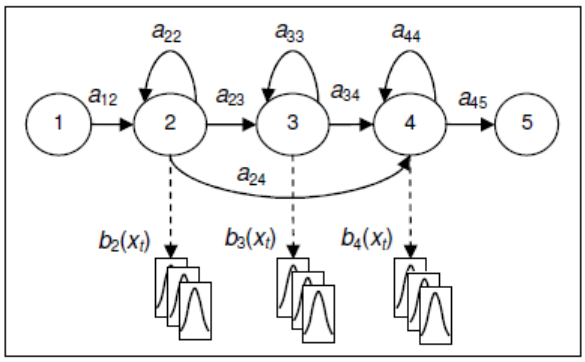

Figura 4: Topologia *left-right* [8].

Por conseguinte, pode-se definir um HMM como um conjunto de M estados que estão conectados por suas transições. À medida que o tempo t avança, existe uma comutação de estados e, nesse processo, símbolos são emitidos com uma dada probabilidade na saída. A sequência de observações (representa a saída do HMM) é gerada a partir da sequência de símbolos emitidos [8] [11] [15] [13] [25].

Dessa forma, o treinamento dos HMMs se dá pelo ajuste dos parâmetros do modelo que busca maximizar a $P(X/W)$ [15] [8].

O algoritmo de *Baum-Welch* mostra-se ser o ideal para ser empregado na etapa de treinamento, visto que por meio de equações recursivas, emprega o critério de maximização da verossimilhança [15] [8] [13].

4) Decodificador: Por fim, a etapa de decodificação transcreve as amostras de voz desconhecidas em sua forma textual. Para isso, o decodificador precisa de um dicionário para indicá-lo as possíveis saídas e uma rede de fonemas que aponta as possíveis transcrições entre fonemas [15] [8] [7].

A decodificação de Viterbi é um algoritmo que, no espaço de estados, busca a sequência de estados que melhor modele a fala; dado que esse espaço de estados é formado a partir da concatenação dos HMMs. Dessa forma, o algoritmo é uma ferramenta que processa cada segmento da fala de forma síncrona [8].

Exemplos de saída do decodificador estão descritos nas Tabelas I e II. Na etapa de avaliação dos resultados, o decodificador faz a análise dos resultados do reconhecimento de fonemas por meio de um alinhamento forçado e gera uma taxa de erro; WER (*Word Error Rate*), que está definida pela Fórmula 5 [17] [11],

$$WER = \frac{S_s + I + D_s}{N_s}, \tag{5}$$

em que N_s é o número total de palavras na sequência de teste e S_s, I e D_s são, respectivamente, o número total de erros por substituição (*substitution*), inserção (*insertion*) e supressão (*deletion*) na sequência reconhecida [17].

Nas etapas de reconhecimento de fala foi utilizado o software livre HTK (*Hidden Markov Models Toolkit*) que consiste em um conjunto de ferramentas que modelam o HMM [17]. No mapeamento dos sinais em Libras foi utilizado a *Python* por também ser uma linguagem de programação livre e multiplataforma.

III. Resultados

O reconhecedor de fala é implementado em duas etapas. A primeira consiste na etapa de treinamento, em que os modelos probabilísticos são treinados a partir de um banco de dados com locuções de diferentes oradores. O banco utilizado é composto de 600 frases (treinamento e teste) foneticamente balanceadas de locuções de homens e mulheres. Esse banco foi obtido do Grupo Fala Brasil que é um grupo de pesquisa criado pelo Laboratório de Processamento de Sinais (LaPS) da UFPA [14]. Na segunda etapa, verifica-se a taxa de reconhecimento desse sistema por meio de testes.

Assim, foi obtido uma taxa de 73,5% de reconhecimento usando modelos acústicos de monofones. A Tabela I ilustra o exemplo da palavra 'pesquisa' obtida na saída do decodificador.

Tabela I: Segmentação Automática da palavra 'pesquisa'

Tempo I (μs)	Tempo F (μs)	Transcrição Fonética	Verossimilhança
330000	360000	p	183.064026
360000	410000	e	344.225586
410000	480000	s	473.969818
480000	530000	k	338.399200
530000	600000	i	487.948212
600000	670000	z	436.364166
670000	710000	a	279.523407

O algoritmo de mapeamento segue conforme ilustrado na Figura 5.

Dada a palavra obtida do reconhecimento de fala, verifica-se a possibilidade de haver apenas caracteres que representem letras do alfabeto Português. Quando isso ocorre, garante-se que a palavra não contém erros de transcrição e a mesma é segmentada, letra a letra, e associada a uma imagem do sinal do alfabeto em Libras que a corresponde. Ao final dessa

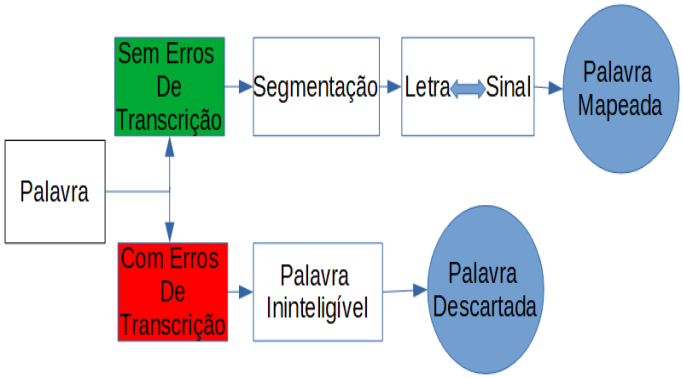

Figura 5: Algoritmo de Mapeamento

etapa, obtém-se um conjunto de letras que, sequencialmente identificadas por seu sinal em Libras, gera a palavra mapeada.

Quando a mesma palavra apresenta caracteres que não correspondem a letras, classifica-se como ininteligível pois a mesma não mapearia uma palavra que fosse facilmente identificada por surdos. Logo, essa palavra é descartada do mapeamento.

De acordo com a Tabela I, considere o exemplo da palavra 'pesquisa'. Essa palavra foi reconhecida e mapeada pelo sistema e observada a sequência de seus sinais em Libras conforme Fig. 6.

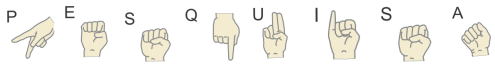

Figura 6: Mapeamento da palavra 'pesquisa'

Assim como apresentado com a palavra 'pesquisa', toda a frase foi mapeada com os sinais em Libras correspondentes. Esse processo foi aplicado em todas as 600 frases do banco utilizado. Considere a palavra 'situações' na saída do decodificador, segundo a Tabela II.

Tabela II: Segmentação Automática da palavra 'situações'

Tempo I (μs)	Tempo F (μs)	Transcrição Fonética	Verossimilhança
560000	590000	s	214.848953
590000	620000	i	208.150284
620000	660000	t	273.837982
660000	690000	u	248.759689
690000	730000	a	313.365295
730000	790000	s	402.475342
790000	840000	o∼	384.243317
840000	890000	i	371.638611
890000	930000	s	275.722168

Nesse exemplo, houve uma falha na identificação dos fonemas 's' e 'o∼'. O erro foi gerado na etapa de reconhecimento de fala, em que não foi possível obter uma conversão grafema/fonema. Por isso os modelos acústicos para esses fonemas devem ser atualizados para corrigir suas transcrições. É interessante salientar que esse mesmo problema ocorreu, principalmente, nas palavras cujas vogais são nasais (a∼ , e∼ , i∼ , o∼ , u∼). Como resultado de um mapeamento falho, obtém-se a seguinte sequência de imagens para a representação da palavra 'situações':

Figura 7: Mapeamento da palavra 'situações'

Por esse motivo, palavras com quaisquer erros em sua decodificação, em um ou mais fonemas, foram excluídas do mapeamento por gerarem palavras ininteligíveis na saída do algoritmo de mapeamento.

Ao fim do mapeamento, conforme a Figura 8, é possível observar a quantidade de palavras que se trabalhou ao longo das 600 frases bem como o número de palavras cujo os fonemas foram bem reconhecidos e portanto bem mapeados e também a quantidade de palavras que houve algum fonema com má identificação e portanto deixou a palavra ininteligível para o mapeamento em Libras. Ao final do mapeamento é possível observar uma taxa de 89,59% de correspondência entre as palavras com suas respectivas representações em imagens que remetem ao alfabeto da Libras.

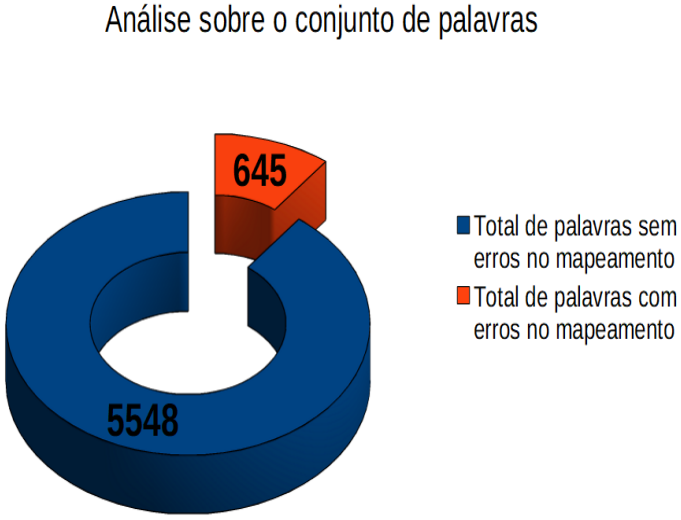

Figura 8: Avaliação das Palavras no Mapeamento

IV. Considerações finais e Trabalhos futuros

Este trabalho propõe uma diminuição nas barreiras de comunicação entre surdos e ouvintes e uma maior integração dos mesmos na acessibilidade de novas tecnologias.

O resultado da taxa de reconhecimento de sentenças por monofones de 73,5% e de mapeamento de 89,59% geram perspectivas promissoras para implementação do sistema de conversão.

Visando a continuação deste trabalho, pretende-se obter os modelos acústicos dos trifones para reconhecimento de palavras bem como o mapeamento de sinais que as representem e, finalmente, a avaliação do sistema por surdos e ouvintes.

Analogamente aos conversores de idiomas orais, é preciso tornar o sistema mais usual para surdos e ouvintes, além de

tornar o conversor em tempo real para as aplicações práticas as quais se destina.

REFERÊNCIAS

[1] G. C. Costa e F. d. G. Daniel, "Google tradutor: Análise de utilização e desempenho da ferramenta," dez. 2013.

[2] Apesar de avanços, surdos ainda enfrentam barreiras de acessibilidade. **Cidadania e Justiça**. set,2016. Disponível em: <http://www.brasil.gov.br/cidadania-e-justica/2016/09/apesar-de-avancos-surdos-ainda-enfrentam-barreiras-de-acessibilidade>. Acesso em: 20 fevereiro 2019.

[3] Honora, M. e Frizanco, M. L. E. **Livro Ilustrado de Língua Brasileira de Sinais: desvendando a comunicação usada pelas pessoas com surdez**. São Paulo: Ciranda Cultural, 2009.

[4] GESSER, Audrei. LIBRAS? que língua é essa?: Crenças e preconceitos em torno da língua de sinais e da realidade surda. São Paulo: Parábola, 2009.

[5] Silva, A. C. da; Camilo, F. M. E.; Ferraz, T. V. D. **Reconhecimento de Fala Dependente de Locutor Utilizando Redes Neurais Artificias**. Trabalho de Conclusão de Curso. Ouro Branco - MG. Universidade Federal de São João Del Rei - Campus Alto Paraopeba - Departamento de Telecomunicações. 2013.

[6] Fechine, J. M. **Reconhecimento Automático de Identidade Vocal Utilizando Modelagem Híbrida: Paramétrica e Estatística**. Doutorado em engenharia elétrica, Universidade Federal da Paraíba, Campina Grande, 2000.

[7] Ynoguti, C. A. **Reconhecimento de Fala Contínua Usando Modelos Ocultos de Markov**. Doutorado em engenharia elétrica, Faculdade de Engenharia Elétrica e de Computação da Universidade Estadual de Campinas, Campinas, 1999.

[8] Tevah, R. T. **Implementação de um sistema de reconhecimento de fala contínua com amplo vocabulário para o português brasileiro**. Mestrado em ciências em engenharia, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2006.

[9] Selmini, A. M. **Sistema Baseado em Regras para o Refinamento da Segmentação Automática de Fala**. Doutorado em engenharia elétrica, Universidade Estadual de Campinas Faculdade de Engenharia Elétrica e de Computação, Campinas, 2008.

[10] Paranaguá, E. D. S. **Reconhecimento de locutores utilizando modelos de markov escondidos contínuos**. Mestrado em ciências em engenharia, Instituto Militar de Engenharia, Rio de Janeiro, 1997.

[11] Rabiner, L. R. **A tutorial on hidden markov models and selected applications in speech recognition** jan. 1988.

[12] Raphael, L. J.; Borden G. J.;Harris K. S. **Speech Science Primer**. Lippincott Williams Wilkins, 5ªed, 2011.

[13] Silva, Carlos Patrick A. da. **Um software de reconhecimento de voz para português brasileiro**. Mestrado em engenharia elétrica, Universidade Federal do Pará, Belém, 2010.

[14] Laboratório de Processamento de Sinais. FalaBrasil – Reconhecimento de Voz para o Português Brasileiro. http://www.laps.ufpa.br/falabrasil/. Acessado em 23 de janeiro de 2018.

[15] Rocha, Raissa Bezerra. **Desenvolvimento de um codificador de voz pessoal de baixa taxa baseado em modelos de markov escondidos**. Mestrado em engenharia elétrica, Universidade Federal de Campina Grande - UFPB, Campina Grande - PB, 2012.

[16] Picone. J. **Signal Modeling Techniques in Speech Recognition** . Proceedings of the IEEE, 1993.

[17] S. Young et al. **The HTK Book**. Cambridge University Engineering Department, 2009.

[18] Rosa Júnior, Jair da. **Reconhecimento Automático de Emoções Através da Voz**. Universidade Federal de Santa Catarina - UFSC, 2017.

[19] Batista, João Luís F. **Biometria Florestal segundo o Axioma da Verossimilhança - Com Aplicações em Mensuração Florestal**. Universidade de São Paulo - USP, Piracicaba -SP, 2014.

[20] Costa, S. L. do Nascimento Cunha. **Análise Acústica, Baseada no Modelo Linear de Produção da Fala, para Discriminação de Vozes Patológicas**. Doutorado em ciências no domínio da engenharia elétrica. Universidade Federal de Campina Grande, Paraíba, 2008.

[21] Fagundes, R. D. R. e Sanches, Ivandro. **Uma Nova Abordagem Fonético-Fonológica em Sistemas de Reconhecimento de Fala Espontânea**. Revista da Sociedade Brasileira de Telecomunicações; Volume 18, Número 3, Dezembro de 2003.

[22] Landim, T. R. G. **Sistema de Comandos e Identificação da Voz**. Monografia. Universidade de São Paulo. Escola de Engenharia de São Carlos. São Carlos, 2017.

[23] Deller, J.R.; Hansen, J. H. L. e Proakis, J. G. ***Discrete-Time Processing of Speech Signals***

[24] Rabiner, L. e Shafer, R. W. ***Digital Processing of Speech Signals***. *Prentice Hall*, *New Jersey*, 1978.

[25] Coelho, L. F. M. P. **Etiquetagem Automática de Sinais de Fala Segmentação e Classificação Fonética**. Dissertação de Mestrado, Faculdade de Engenharia da Universidade do Porto, Porto, Portugal, Fevereiro de 2005.