Aplicação de MFCCs e CNNs no Reconhecimento Automático de Fala para Produção de Legenda Oculta

Kassius S. Bezerra, Luiz A. Pinto

Resumo Os Autômatos Finitos podem ser utilizados na fase de Análise Léxica (AL) para identificar tokens de uma linguagem de programação, tendo em vista que analisam caractere a caractere da entrada, que é o princípio da AL. Neste artigo, é apresentado um estudo sobre a aplicação de AL na linguagem de programação C, onde são utilizados Autômatos Finitos para determinação de três grupos de *tokens*: palavras-chave, símbolos especiais e operadores. Foi realizado um estudo analítico sobre o uso de Autômato Finito Determinístico (AFD) e Autômato Finito Não-Determinístico (AFND), de modo a embasar a abordagem mais adequado ao problema tratado. A corretude dos resultados alcançados através da estratégia de exploração de AC proposta mostram sua eficácia frente aos trabalhos correlatos apresentados na revisão de escopo.

1 Introdução

Dados do Censo Demográfico do IBGE de 2010 mostram que 5,1% da população brasileira se auto declararam com algum grau de deficiência auditiva. Em valores absolutos, em 2010 existiam no Brasil mais de 9 milhões de deficientes auditivos. O elevado contingente de pessoas com problemas de audição, mostra a necessidade da implantação de mecanismos, legais e tecnológicos, para a promoção da acessibilidade junto a essa parcela da população, dando-lhes acesso ao consumo de conteúdo de audiovisual.

Nesse contexto, as empresas de radiodifusão, emissoras de televisão, estão obrigadas por lei a inserirem legendas ocultas em sua totalidade do conteúdo audiovisual exibida, inclusive comerciais, respeitando os critérios de assertividade e tempo de legendagem. De acordo com a NBR 15290 (**15˙nbr2016˙15290**), para qualquer conteúdo exibido de uma fonte gravada, a taxa de erro da transcrição deveria ser de no máximo 1,5%, enquanto para os conteúdos veiculados ao vivo, uma pequena margem de 5% de erro é aceitável.

Embora o melhor sistema de Reconhecimento Automático de Fala *(Automatic Speech Recognition - ASR)* disponível atualmente no mercado, fornecido pelo *Google*, atinja a taxa de

Kassius S. Bezerra
Programa de Pós-graduação em Engenharia de Controle e Automação, Instituto Federal do Espírito Santo (IFES), Serra, ES, Brasil.
e-mail: ksipolati@gmail.com

Luiz A. Pinto
Programa de Pós-graduação em Engenharia de Controle e Automação, Instituto Federal do Espírito Santo (IFES), Serra, ES, Brasil.
e-mail: pintoluizalberto@gmail.com

Anais do XV Encontro Anual de Ciência da Computação (EnAComp 2020). ISSN: 2178-6992.
Catalão, Goiás, Brasil. 25 a 27 de Novembro de 2020.

erro estabelecida pela NBR, a forma com que a fala é processada interfere no resultado da inteligibilidade do texto. Isso porque, além da extração das informações fonéticas, o *ASR* desenvolvido pelo *Google* também utiliza a análise do modelo de linguagem, que considera os aspectos semânticos e gramaticais, bem como as informações contextuais para a construção das sentenças (**15˙aleksic2015bringing**). A aplicação desse modelo de linguagem permite que palavras em qualquer posição da frase sejam alteradas para se ajustar ao contexto do discurso, o que resulta em um reconhecimento de fala mais próximo do natural.

Contudo, a troca de palavras reduz o tempo de exposição do texto para leitura. Para ilustrar, um telejornal local no estado do Espírito Santo deu a seguinte notícia “Chikungunya está assustando os moradores de Cachoeiro.” No momento inicial o *ASR* legendou como “Chico Cunha está assustando os moradores de Cachoeiro.” Somente após a análise do contexto a notícia foi legendada corretamente. Contudo, o tempo de exposição da primeira frase foi de aproximadamente 3,5 segundos. Após a troca da palavra pelo *ASR* o tempo de leitura para a frase correta foi de aproximadamente 0,5 segundo. Se considerarmos uma aplicação de reconhecimento de fala para inserção de legendas, tendo em vista a acessibilidade de pessoas com deficiência auditiva, a redução do tempo de leitura pode impedir sua completa acessibilidade ao evento.

Os critérios da NBR para validação de *ASR* para legendagem utiliza duas métricas, a *WER* (1) e a *NER* (2). Enquanto a *WER* (*Word Error Rate*) é uma forma de avaliação da quantidade de erros na legenda, o *NER* (*Number of words, Edition Error* e *Recognition error*) foi proposto para avaliar a qualidade do que está sendo transcrito.

$$WER = \frac{(N-E)}{N} \tag{1}$$

N: Quantidade de palavras ditas pelo interlocutor;
E: Quantidade de palavras transcritas incorretamente.

$$NER = \frac{(N-E-R)}{N} \tag{2}$$

N: Quantidade de palavras ditas pelo interlocutor, pontuações, identificações de interlocutores e pausas;
E: Erros de edição, geralmente causados pelas más escolhas do legendista/sistema na produção do CC (Closed Caption);
R: Erros de reconhecimento causados por pronúncia ou escuta incorreta. Também podem ser causados pela tecnologia usada para produzir as legendas. Esses erros podem ser inserções, exclusões ou substituições.

Quando considerado a *WER*, os melhores *ASRs* disponíveis no mercado atendem os critérios da NBR. Contudo, os valores do *NER* desses sistemas ficam abaixo dos estabelecidos pela norma.

Dessa forma, esse trabalho propõe uma estrutura para *ASR* que faça a transcrição da fala do interlocutor, em tempo real, enfatizando a estrutura fonética do texto, tendo em vista eliminar os erros de reconhecimento e aumentar o tempo de leitura do texto legendado por pessoas portadoras de deficiências auditivas. A avaliação de desempenho do sistema proposto utilizará como métrica a *NER*, tendo em vista que essa métrica já considera o número de palavras classificadas incorretamente, bem como os erros de edição, como as trocas de palavras ao longo da criação da legenda pelos sistemas de *ASR* atuais. Para a implementação do sistema, os espectrogra-

mas obtidos das (*Mel-Frequency Cepstral Coefficients - MFCC*) (**15˙davis1980comparison**) dos sinais de áudio de um *dataset* de teste serão convertidos em imagens que posteriormente serão fornecidas como entradas para uma Rede Neural Convolucional (*Convolutional Neural Networks - CNN*) (**15˙krizhevsky2017imagenet**). A extração de características da fala via *MFCCs* é uma técnica consolidada, visto que suas etapas são modeladas para aproximar o sistema da forma com que o ouvido humano funciona. Sobre essa técnica já bem explorada, trazemos uma nova aplicação, que é a classificação via imagens dos espectrogramas e não o espectrograma em si diretamente, com o objetivo de combinar dois métodos reconhecidamente eficientes, *MFCCs* e *CNNs*.

2 *Mel-Frequency Cepstral Coefficients*

Um sinal de voz é o resultado da convolução entre a sequência de excitação e a resposta ao impulso do sistema vocal. Nesse trabalho decidiu-se pela utilização das *MFCCs* para representar a fala, por ser o que mais se aproxima da forma da escuta do ouvido humano. Para a extração das *MFCCs*, é necessário separar as duas componentes. A análise cepstral, cujo procedimento se encontra sucintamente descrito no que segue (**15˙bogert˙1963**), foi proposta para tornar mais simples a solução desse problema.

Após o sinal estar representado no *cepstrum*, aplica-se um filtro linear para remover os trechos indesejados e selecionar algumas componentes específicas. Às componentes que não foram eliminadas ($x(n)$) aplica-se uma transformação inversa. Este procedimento obedece o princípio da sobreposição, que no caso da convolução pode ser representado por (3), sendo que "$H[.]''$ é um sistema homomórfico e o símbolo $(*)$ representa a operação de convolução.

$$H[x(n)] = H[x_1(n) * x_2(n)] = H[x_1(n)] * H[x_2(n)] \tag{3}$$

Sistemas homomórficos são aqueles que obedecem ao princípio da sobreposição para a convolução. O operador de *cepstrum* complexo $D_*[.]$, cuja representação está mostrada na Figura 1 desempenha um papel importante na teoria de sistemas homomórficos, que é baseado em uma generalização do princípio da superposição.

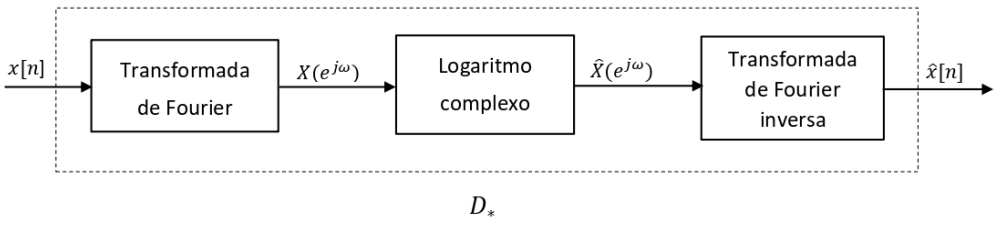

Figura 1: Representação do cálculo do cepstrum complexo. Fonte: Autor.

Na filtragem homomórfica de sinais convoluídos, o operador $D_*[.]$ é denominado sistema característico para convolução, pois tem a propriedade especial de transformar a convolução em adição (**15˙oppenheim˙1968**)]. Consideremos que na transformada Z em (4),

$$X(z) = X_1(z).X_2(z) \tag{4}$$

o logaritmo complexo seja calculado de acordo com a definição do *cepstrum* complexo (5), então,

$$\widehat{X}(z) = log[X(z)] = log[X_1(z)] + log[X_2(z)] = \widehat{X_1}(z) + \widehat{X_2}(z) \tag{5}$$

o que implica que o *cepstrum* complexo dado por

$$\widehat{x}(n) = D_*[x_1[n] * x_2[n]] = \widehat{x}_1(n) + \widehat{x}_2(n) \tag{6}$$

Uma análise similar mostra que, se

$$\widehat{y}[n] = y_1[n] + y_2[n] \tag{7}$$

então segue que

$$D_*^{-1}[\widehat{y}_1[n] + \widehat{y}_2[n]] = \widehat{y}_1[n] * \widehat{y}_2[n] \tag{8}$$

Se os componentes cepstrais $\widehat{x}_1[n]$ e $\widehat{x}_2[n]$ ocuparem diferentes faixas de frequência, a filtragem linear pode ser aplicada ao *cepstrum* complexo para remover ou $\widehat{x}_1[n]$ ou $\widehat{x}_2[n]$. Ainda de acordo com (**15˙oppenheim˙2012**), se esta etapa for seguida da transformação por meio do sistema inverso $D_*^{-1}[.]$, cuja constituição esta ilustrada na Figura 2, o componente correspondente será removido na saída.

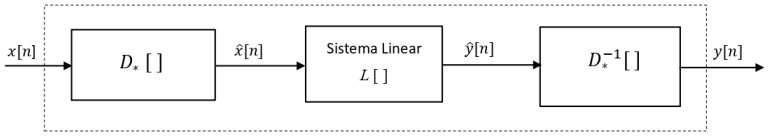

Figura 2: Sistema homomórfico com entrada e saída combinadas por convolução. Fonte: Autor.

O *MEL-Cepstrum* é uma variação do *cepstrum* normal que melhor se ajusta à percepção auditiva humana. A verdadeira frequência de um som e a percepção que o ouvido humano tem dessa frequência não têm relação linear. A frequência percebida pelo ouvido humano, também conhecida como *pitch*, tem como unidade de medição o *MEL*.

3 Redes Neurais Convolucionais

Entre as abordagens para a implementação de algoritmos de *Deep Learning*, as Redes Neurais Convolucionais (*Convolutional Neural Network - CNN)* têm sido utilizadas com sucesso em aplicações de classificação, detecção e reconhecimento de objetos em imagens (**15˙GirshickDDM13**) e em vídeos. Essas arquiteturas de rede são constituídas por camadas de vários tipos: camada convolucional, camada de *pooling*, totalmente conectada, normalização em lote. Cada uma delas com diferentes funções no processamento.

Algumas arquiteturas de redes neurais convolucionais que têm sido propostas no decorrer dos últimos anos, têm se notabilizado pelo desempenho em tarefas envolvendo imagens. Por exemplo, a Alexnet, Zfnet, Vggnet, Googlenet (**15˙christian2014**), Resnet, Resnext e Senet.

Por causa do seu desempenho em tarefas de extração de características e de classificação envolvendo imagens, nesse trabalho, decidiu-se pela utilização da rede GoogLeNet. A GoogLeNet é uma *CNN* cuja principal característica é a introdução do módulo *Inception*, que, quando utilizado reduz o número de parâmetros da rede. De acordo com (**15˙busson˙2018**), a introdução do módulo *Inception* pode minimizar os efeitos do problema da localização da informação através da utilização de filtros de diferentes dimensões na mesma camada, deixando a rede mais larga em comparação às arquiteturas de *CNNs* convencionais.

4 Metodologia

Nessa seção são descritas todas as etapas da construção do trabalho, do tratamento dos dados até a avaliação do resultado.

4.1 Base de dados

Na realização do trabalho foi utilizada uma base de dados desbalanceada, com diferentes palavras para comandos de voz (**15˙base1**). Trechos de silêncio antes e/ou depois de cada pronúncia foram removidos (**15˙wvckassius**), de acordo com o procedimento descrito a seguir.

Após o cálculo do histograma, foi aplicado um filtro de mediana com o objetivo de suavizar os ruídos existentes. Os dois maiores valores de histograma foram encontrados e armazenados nas variáveis $M1$ e $M2$. O valor do *threshold* (T) foi obtido por (9), onde W foi definido por (**15˙giannakopoulos2009method**) com valor igual a 5.

$$T = \frac{W.M1 + M2}{W + 1}, \text{onde } \ W\text{=5} \tag{9}$$

O *threshold* calculado foi aplicado em toda a curva, descartando os valores abaixo de T. Com isso, uma nova curva foi obtida, apenas com o trecho com a pronúncia das palavras. Essa nova curva foi salva como um arquivo em formato *wave*, mantendo as especificações do arquivo original.

4.2 Extração das MFCCs

Para a extração das MFCCs, uma etapa de *pre-emphasis* foi aplicada para melhorar a relação sinal-ruído, como representado em (10).

$$y[n] = x[n] - 0,97x[n-1] \tag{10}$$

Em (10), $x[n]$ representa os sinais de entrada e $y[n]$ os sinais de saída do filtro de *pre-emphasis*. Para segmentos sonoros como vogais, há mais energia nas frequências mais baixas do que nas frequências mais altas. Isso é chamado *spectral tilt*, e está relacionado à glote e, portanto, com a maneira como as cordas vocais produzem o som. Destacar as altas frequências melhora o modelo acústico do áudio para a extração das *MFCCs* (**15˙singh˙2014**).

Como a fala é um sinal de natureza essencialmente não-estacionária, sua análise espectral requer o particionamento do mesmo em segmentos de comprimento tal que seja possível a consideração de estacionariedade (**15˙logan2000mel**). Para a determinação do tamanho do *frame*, deve-se estabelecer uma relação de compromisso entre a estacionariedade (segmentos devem ser aproximadamente estacionários) e a preservação de componentes com bandas próximas. Se o *frame* for muito grande, não será possível considerar o sinal estacionário e, portanto, os descritores estarão pouco correlacionados com os sinais de voz. Caso o *frame* seja pequeno, os componentes com bandas próximas serão eliminados, afetando negativamente a resolução na frequência.

A segmentação dos arquivos de áudio, requer que o sinal seja particionado em *frames* com N amostras. A fim de assegurar que as informações que se encontram na fronteira entre dois *frames* não se percam, os sinais adjacentes são sobrepostos por M amostras. Nesse trabalho serão utilizados $N = 256$ e $M = 128$ (**15˙young2002htk**).

A sobreposição entre os sinais adjacentes provoca descontinuidades no início e no final de cada *frame*, o que degrada a qualidade do sinal no domínio da frequência. Para reduzir os efeitos dessas descontinuidades, deve ser aplicada uma função de janelamento, que para essa finalidade, a janela de Hamming é a que provoca menor distorção (**15˙hasan2004speaker**). A função de janelamento de Hamming está mostrada em (11).

$$h(n) = \begin{cases} 0,54 - 0,46cos(\frac{2\pi n}{N}), \text{ se } 0 \leq n \leq N \\ 0, \qquad\qquad\qquad\qquad \text{caso contrário} \end{cases} \tag{11}$$

Uma etapa importante para a extração das *Mel-frequencies*, é a conversão do sinal de áudio original para o domínio da frequência. Para isso, será aplicada a Transformada Discreta de Fourier, representada em (12).

$$[ht]X(w) = \sum_{n=0}^{N-1} x(n)e^{\frac{-j2\pi\omega n}{N}}, 0 \leq w \leq N-1 \tag{12}$$

Em (12), o termo $x(n)$ é a função para o *frame* de entrada no domínio do tempo e $X(w)$ é o correspondente sinal transformado para o domínio da frequência. Uma vez que os coeficientes complexos não são considerados para a obtenção das *Mel-frequecies*, (13) pode ser utilizada para eliminá-los.

$$|X(w)| = \sqrt{(Re(X(w)))^2 + (Im(X(w)))^2} \tag{13}$$

A escala das *Mel-frequencies* foi constituída com base na percepção humana dos sons. Nessa escala, 1 Mel representa um milésimo do tom de $1kHz$. A Figura 3 representa o filtro triangular ao qual cada *frame* é aplicado. Como a maioria dos filtros são vinculados à região de baixa frequência (**15˙molau2001computing**), sua aplicação destaca a região onde o sinal de fala é dominante (**15˙huang2001spoken**).

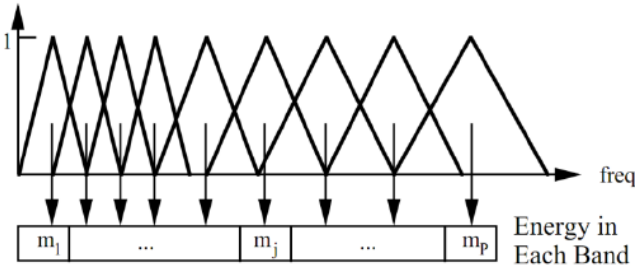

Figura 3: *Mel Scale Filter Bank* (**15˙young2002htk**).

Utilizando (14) pode-se calcular a *Mel-frequency* correspondente a uma frequência específica f em Hz. Nesse trabalho foi utilizado um banco de filtros de *Mel-Frequency* com 20 filtros triangulares.

$$H(Mel) = [2595 * log_{10}(1 + \frac{f}{700})] \tag{14}$$

A audição humana atenua os sons captados aplicando uma escala que se assemelha a uma escala logarítmica. Esse fenômeno pode ser modelado matematicamente de acordo com (15), onde $X(w)$ é o *frame* de entrada no domínio da frequência, $H(w)$ é o *frame* após a aplicação do *Mel Scale Filter Bank* e M é o número de filtros triangulares usados na escala Mel. Nesse caso adotou-se $M = 20$ (**15˙young2002htk**).

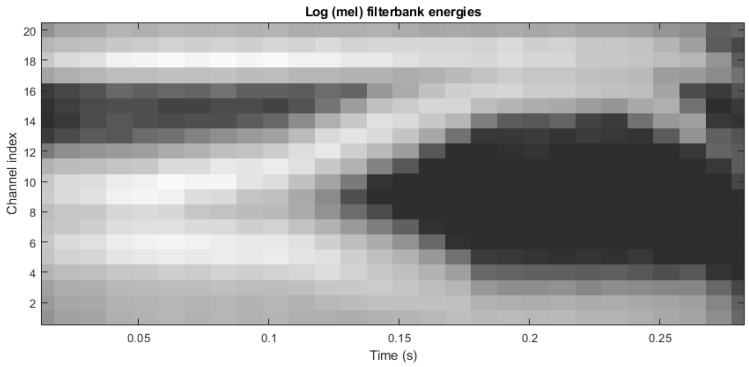

Figura 4: Espetrograma do LFE da palavra *five*. Fonte: Autor.

A Figura 4 mostra o espectrograma do áudio da palavra *five* após a aplicação do *Logarithim Filter Energy - LFE*.

$$S(m) = log_{10}\left[\sum_{w=0}^{N-1} \left|X(w)^2\right| .H(w))\right], 0 \leq w \leq M \tag{15}$$

Para obtermos a informação da taxa de variação na banda espectral, ou seja, o *cepstrum*, aplicamos na última etapa uma transformada discreta de cossenos. Como essa transformada será calculada sobre um sinal que já está no domínio da frequência, o espectro resultante não está nem no domínio do tempo e nem no domínio da frequência, e sim em um novo domínio chamado por (**15˙oppenheim2004frequency**) de *quefrency*. Através de (16) obtemos a *MFCC*

dado o *frame* no domínio da frequência, constituindo assim a *Mel-frequency* da informação sobre o sinal em seus coeficientes de ordem mais baixa.

$$C(w) = \sum_{m=0}^{M-1} S(w) cos(\frac{\pi w(m+\frac{1}{2})}{M}), 0 \leq w \leq W \tag{16}$$

O resultado final da etapa de extração de descritores são vetores que constituem as *Mel-frequencies*. As imagens constituídas pelos *cepstrum* da combinação desses vetores serão utilizadas na etapa de classificação.

4.3 Classificação

Para a classificação das palavras pronunciadas uma rede baseada na arquitetura da GoogLeNet foi treinada com as imagens dos *ceptrums* das *MFCCs*. Os arquivos de áudio do *dataset* tiveram sua *MFCCs* calculadas, e os *ceptrum* correspondentes formaram o banco de imagens. Todas as imagens, *ceptrum* têm resolução igual a $224 \times 224 \times 3$. Conjuntos de treino e teste foram obtidos, de forma aleatória, com 70% e 30% do total das imagens, respectivamente.

5 Resultados e Discussão

Para a obtenção dos resultados o classificador recebeu a base de imagens dos *cepstrums* das *MFCCs*. A avaliação de desempenho dos modelos foi realizada, e as acurácias das classes individuais obtidas final das oito épocas de treinamento estão indicadas na Figura 5. Como o

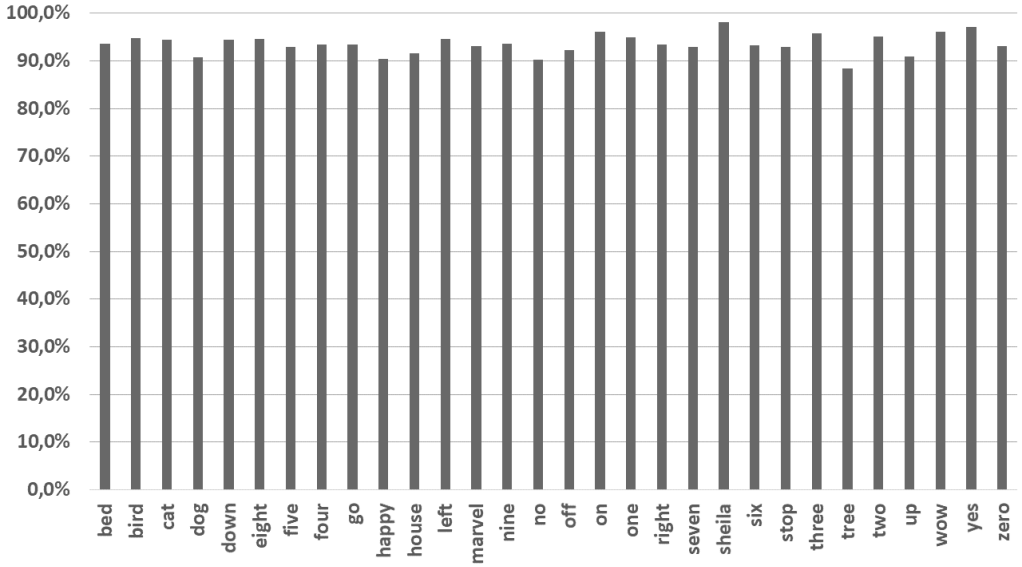

Figura 5: Acurácia por classe. Fonte: Autor.

banco de dados utilizado nesse trabalho apresenta um elevado nível de desbalanceamento, foi

utilizada também a métrica *F1-Score*, além do recall e da precisão. Os resultado obtidos foram: *F1-Score* $93,7\%$; Precisão $93,9\%$; *Recall* $93,6\%$; Acurácia $93,5\%$. Esses resultados se mostram promissores tendo em vista o desenvolvimento de sistemas *ASR* para fins de legendagem de conteúdo de audiovisual.

Para verificar a adequação da legendagem aos critérios da NBR 15290, foi avaliado também o tempo máximo de transcrição de programas ao vivo, ou seja, o atraso entre a entrada do sinal de voz e a escrita do texto. De acordo com a norma esse tempo não deve exceder a 4 segundos. Para avaliar esse requisito, um teste de carga foi realizado com 100 entradas. Nesse teste 100 arquivos de áudio foram classificados e seus tempos de transcrição foram medidos. Em média, o tempo dispendido para transcrever os áudios de entrada ficou estabelecido em $6,46ms$. A Tabela 1 mostra, para cada etapa, o tempo mínimo, máximo e a média para o protótipo baseado nas *MFCCs*.

	Remover silêncio	Extração MFCC	Classificar	TOTAL
Máximo (ms)	0,22	0,70	8,26	9,19
Mínimo (ms)	0,134	0,31	5,34	5,78
Média (ms)	0,16	0,38	5,92	6,46

Tabela 1: Tempo dispendido para transcrição dos áudios por etapa.

6 Conclusão

Nesse trabalho buscou-se uma solução para melhorar a qualidade da acessibilidade das legendas ocultas. Hoje, apesar dos melhores sistemas disponíveis no mercado atenderem ao critério de *WER* para a tarefa de transcrição de fala para texto, eles não conseguem obter resultados satisfatórios para atender a metodologia de medição *NER*, sendo essa a que está em vigor na legislação brasileira. O fato dos sistemas atuais utilizarem-se de modelos de linguagem e de análise de contexto para a tarefa de transcrição, permite que algumas palavras ao longo da pronúncia de uma sentença seja modificada, o que dificulta a inteligibilidade. Essas trocas de palavras são fatores que impactam no resultado *NER*.

Sobre os objetivos propostos, podemos afirmar que a abordagem de reconhecimento automático de fala a partir de imagens das *MFCCs*, com o uso das *CNNs* mostrou-se apta para a tarefa de transcrição de fala em texto, para fins de legendagem, respeitando as definições da NBR 15.290. O valor de *NER* se aproximou do valor estabelecido pela norma, $93,54\%$ contra 95%. Em relação ao tempo de transcrição o resultado superou o determinado pela normativa, $6,46ms$ contra 4 segundos.

Embora a acurácia (NER) obtida não tenha atingido o valor estabelecido pela norma os resultados mostraram que a abordagem do problema de reconhecimento de fala utilizando-se da capacidade das CNNs em classificar imagens, é um caminho de pesquisa promissor para esse campo da ciência(**15˙krizhevsky2012imagenet**).

Referências

GEEKSFORGEEKS. **C/C++ Tokens**. [S.l.: s.n.], 2019. Disponível em: https://www.geeksforgeeks.org/cc-tokens/. Acesso em 03 de Maio de 2019.

GOMES, J. Development of an ATLAS test language to automatic test markup language translator. In: IEEE. PROCEEDINGS AUTOTESTCON 2004. San Antonio, TX, USA: [s.n.], 2004. p. 191–195.

KLEIN, G.; ROWE, S.; DÉCAMPS, R. **Jflex User's Manual**. [S.l.: s.n.], 2010. Disponível em: https://jflex.de/manual.html. Acesso em 30 de Setembro de 2020.

MARKOVIĆ, I. M. An Application for Visual Representation of Deterministic Finite Automaton Generated by JFlex. In: IEEE. 2018 26th Telecommunications Forum (TELFOR). Belgrade, Sérvia: [s.n.], 2018. p. 420–425.

MEMETI, S.; PLLANA, S. PaREM: A Novel Approach for Parallel Regular Expression Matching. In: 2014 IEEE 17th International Conference on Computational Science and Engineering. Chengdu, Sichuan, China: [s.n.], 2014. p. 690–697.

SANJU, V et al. An exploration on lexical analysis. In: IEEE. 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT). [S.l.: s.n.], 2016. p. 253–258.

YANG, W. On the look-ahead problem in lexical analysis. **Acta Informatica**, Springer, v. 32, n. 5, p. 459–476, 1995.