Comunicação, Mídias e Educação

Vanessa Cristina de Abreu Torres Hrenechen
(Organizadora)

Atena Editora
Ano 2019

Vanessa Cristina de Abreu Torres Hrenechen

(Organizadora)

Comunicação, Mídias e Educação

Atena Editora
2019

2019 *by Atena Editora*

Editora Executiva: Profª Drª Antonella Carvalho de Oliveira
Diagramação: Karine de Lima
Edição de Arte: Lorena Prestes
Revisão: Os Autores

Atena Editora
Ano 2019

Dados Internacionais de Catalogação na Publicação (CIP)
(eDOC BRASIL, Belo Horizonte/MG)

C741 Comunicação, mídias e educação [recurso eletrônico] / Organizadora Vanessa Cristina de Abreu Torres Hrenechen. – Ponta Grossa, PR: Atena Editora, 2019.

Formato: PDF
Requisitos de sistema: Adobe Acrobat Reader.
Modo de acesso: World Wide Web.
Inclui bibliografia
ISBN 978-85-7247-344-6
DOI 10.22533/at.ed.446192205

1. Aprendizagem. 2. Comunicação – Pesquisa – Brasil. 3.Comunicação na educação. I. Hrenechen, Vanessa Cristina de Abreu Torres.

CDD 371.1022

Elaborado por Maurício Amormino Júnior – CRB6/2422

Atena Editora
Ponta Grossa – Paraná - Brasil
www.atenaeditora.com.br
contato@atenaeditora.com.br

Atena Editora
Ano 2019

APRESENTAÇÃO

Essa obra reúne um conjunto de pesquisas sobre as novas tecnologias e técnicas aplicadas à comunicação. O compilado de artigos traz contribuições relevantes para a comunidade científica e profissionais da área.

O e-book, composto por 36 artigos, apresenta diálogos contemporâneos e reflexões sobre o papel da comunicação nos mais diversos âmbitos. Estudos analisam o uso das novas mídias na educação e avaliam a convergência dos meios na partilha de informações e aprendizagem em conjunto. Pesquisas também retratam o consumo midiático, culturas comunicacionais e as manifestações no espaço urbano.

Há artigos sobre o ambiente *comunicacional* digital e o impacto das novas tecnologias na sociedade. Autores também discutem as discrepâncias entre as visões de mundo dos jornalistas e dos usuários de redes sociais e o papel dos meios de comunicação na representação da realidade. O volume traz pesquisadores de peso que compartilham conhecimento e estimulam novos estudos na área da comunicação.

Vanessa Cristina de Abreu Torres Hrenechen

SUMÁRIO

CAPÍTULO 23

UMA ABORDAGEM PARA O DESENVOLVIMENTO E ANÁLISE DE DESEMPENHO DO RECONHECIMENTO OFF-LINE DE VOZ CONTÍNUO

Lucas Debatin
Laboratório de Inteligência Aplicada – Universidade do Vale do Itajaí (UNIVALI)
Itajaí – Santa Catarina

Aluizio Haendchen Filho
Laboratório de Inteligência Aplicada – Universidade do Vale do Itajaí (UNIVALI)
Itajaí – Santa Catarina

Rudimar Luís Scaranto Dazzi
Laboratório de Inteligência Aplicada – Universidade do Vale do Itajaí (UNIVALI)
Itajaí – Santa Catarina

RESUMO: O reconhecimento de voz é uma forma de acessibilidade utilizada para executar tarefas com as mãos e os olhos livres em aparelhos eletrônicos, e isso é vantajoso independente do tipo de usuário. O reconhecimento de voz é realizado por meio de APIs, que apresentam algumas limitações: (*i*) dependem de conexão com a internet; e (*ii*) muitas vezes são softwares proprietários, ou seja, há um custo para a aquisição de licenças de uso. Visando a solução dessas limitações, o presente trabalho propõe o desenvolvimento do reconhecimento off-line de voz contínuo. Inicialmente, realizou-se uma revisão sistemática da literatura para obter o estado da arte da pesquisa. Após a leitura dos artigos selecionados, foram identificadas bibliotecas para facilitar a implementação, tais como CMUSphinx, HTK e Kaldi. Para cada biblioteca foram criados 10 arquivos de configuração de treinamento. As configurações que obtiveram as melhores métricas de avaliação foram implementadas e testadas. Para cada biblioteca, realizou-se a análise de desempenho, no qual foram verificados os percentuais de uso do processador e de memória. A biblioteca Kaldi obteve o melhor resultado, e apresentou uma taxa de erro (WER) de 5,05% no corpus de voz com vários locutores e 1,48% no corpus com apenas um locutor.

PALAVRAS-CHAVE: Reconhecimento de Voz, Contínuo, Off-line.

ABSTRACT: Voice recognition is a form of accessibility used in electronic devices to perform tasks with free hands and eyes, and this is advantageous regardless of the type of user. Voice recognition is performed through APIs, which have some limitations: (i) depend on internet connection; and (ii) are often proprietary software, so there is a cost to purchase usage licenses. In order to solve these limitations, the present work proposes the development of off-line voice recognition. Initially, a systematic literature review was conducted to obtain the state of the art of the research. After reading the selected articles, libraries such as CMUSphinx, HTK and Kaldi were identified and selected to

facilitate implementation. For each library, 10 training configuration files were created. The configurations that obtained the best evaluation metrics were implemented and tested. In order to verify the percentages of processor and memory usage, performance analysis was performed for each library. The Kaldi library obtained the best result, presenting an error rate (WER) of 5.05% in the voice corpus with several speakers and 1.48% in the corpus with only one speaker.
KEYWORDS: Speech Recognition, Continuous, Offline.

1 I INTRODUÇÃO

O reconhecimento de voz é uma importante tecnologia para melhorar a IHC (Interação Homem-Computador), pois pode proporcionar interação mesmo se o usuário estiver com as mãos e olhos ocupados ou se o usuário possuir capacidades limitadas, ou seja, é vantajosa independentemente do tipo de usuário, exceto para pessoas com afonia ou disfemia. Além disso, torna-se mais rápido o acesso às informações nos softwares e aplicações (JURAFSKY; MARTIN, 2008; SILVA, 2010; VEIGA, 2013).

Esse reconhecimento pode ser classificado em dois tipos: (*i*) palavras isoladas, que necessita que as sentenças sejam pronunciadas com pausas entre cada palavra; e (*ii*) contínuo, aplicado nesta abordagem, tem como objetivo tornar a comunicação mais eficaz para os seres humanos, visto que reconhecem sentenças pronunciadas de forma natural. (ALENCAR, 2005; JURAFSKY; MARTIN, 2008; RUSSELL; NORVIG, 2004; SILVA, 2010).

Atualmente, as APIs (*Application Programming Interface*), Web Speech, Java Speech, Google Cloud Speech, Bing Speech, dentre outras, facilitam a implementação do reconhecimento de voz contínuo do português brasileiro em softwares e aplicações (DEBATIN; HAENDCHEN FILHO; DAZZI, 2018). Entretanto, as APIs atualmente disponibilizadas não podem ser empregadas em qualquer tipo de aplicação, pois apresentam limitações.

A primeira limitação que as APIs existentes apresentam é que nenhuma delas realiza o reconhecimento em modo off-line. Essa limitação é uma barreira no Brasil, pois aproximadamente 36% da população, com idade acima de 10 anos, não está conectada à internet. Isso afeta diversas pessoas que possuem capacidades limitadas e moram em localidades sem internet, uma vez que o reconhecimento de voz é um importante meio de acessibilidade. Além disso, também afeta empresas que possuem, em seus aplicativos, o reconhecimento de voz via APIs, visto que em muitos casos não é possível distribuir o sinal wireless por toda a empresa (DEBATIN; HAENDCHEN FILHO; DAZZI, 2018; IBGE, 2016).

Outra limitação, é que as APIs são softwares proprietários, e em muitos casos o valor pago pela licença de uso se torna alto, visto que depende diretamente da quantidade de requisições que a API realiza. Essa limitação também afeta as empresas, pois é fundamental que o mesmo seja gratuito, devido ao grande número

de requisições que é necessário (DEBATIN; HAENDCHEN FILHO; DAZZI, 2018).

Embora o termo off-line caracterize uma abordagem antiga, visto que atualmente a sociedade encontra-se na era da computação em nuvem, ela apresenta algumas vantagens: (*i*) não sofre de problemas relacionados à latência e à largura de banda, pois os serviços em nuvem são disponibilizados por servidores remotos; (*ii*) não apresenta problemas relacionados ao compartilhamento do mesmo servidor, visto que os serviços de nuvem atendem a vários clientes, e se as requisições de um usuário comprometerem o servidor, também poderão comprometer aplicativos de outros usuários; e (*iii*) não apresenta problemas de segurança, conformidade e regulamentares, pois os dados na nuvem podem ser acessíveis a terceiros (GROSSMAN, 2009).

Dentro desse contexto, esse trabalho identificou e selecionou as principais técnicas do reconhecimento de voz contínuo. Na implementação, utilizou-se as bibliotecas CMUSphinx, HTK e Kaldi, porém cada biblioteca possui arquivos de configuração que podem ser editados. Realizou-se um estudo comparativo para encontrar a configuração com o melhor custo-benefício entre desempenho e precisão da taxa WER (*Word Error Rate*). Por fim, esse trabalho comparou o processamento e uso de memória das bibliotecas em um computador desktop, pois os mesmos possuem maiores recursos de hardware.

2 I FUNDAMENTAÇÃO

Esta seção apresenta a base teórica dos temas abordados no artigo, que são: reconhecimento de voz, extração de características, decodificador e métricas de avaliação.

2.1 RECONHECIMENTO DE VOZ

O reconhecimento de voz é o processo de converter o sinal de voz analógico em sua representação textual, isto é, o texto gerado é composto pela sequência de palavras que foram identificadas a partir do sinal de entrada (RUSSELL; NORVIG, 2004; SILVA, 2010; VEIGA, 2013).

Para que o reconhecimento tenha um bom desempenho é importante conhecer as características do sinal da voz (ALENCAR, 2005). Além do mais, é fundamental descrever os fatores que aumentam a complexidade do reconhecimento de voz e a sua estrutura básica.

2.1.1 Características da Voz

O sinal de áudio possui diversas informações sobre o locutor, que são classificadas em: (*i*) baixo nível, que são os tons, intensidade, correlações espectrais, entre outros; e (*ii*) alto nível, que são variações na entonação, tais como dialeto, contexto, estilo de falar, estado emocional (por exemplo, dor e alegria), entre outros (MÜLLER, 2006; PATRA, 2007).

A quantidade de dados gerada durante a fala é grande, porém as características essenciais da voz mudam lentamente, isto é, requer uma menor quantidade de dados para representar as características mais importantes. Por isso, em sistemas de voz resumem-se as propriedades do sinal ao longo de intervalos chamados quadros, e cada quadro é representado por um vetor de características (PATRA, 2007; RUSSELL; NORVING, 2004).

2.1.2 Fatores de Complexidade

Segundo Veiga (2013), o reconhecimento de voz simula o sistema de audição humana, e tem seu desempenho influenciado pelas características que afetam o sinal da fala. Além do tipo de reconhecimento de voz (palavras isoladas ou contínuo) existem outros fatores de complexidade, tais como:

I. Tamanho do vocabulário: quanto maior for o tamanho, maior é a probabilidade de erro. O tempo de treinamento e a quantidade de memória utilizada aumenta linearmente com o aumento do vocabulário (ALENCAR, 2005; JURAFSKY; MARTIN, 2008; SILVA, 2010)

II. Variabilidade: está relacionada à fatores: (*i*) internos, que são as diferenças de uma pessoa para outra (diversidade de gênero, idade e sotaques) e em um mesmo indivíduo (estado emocional, ruídos dos ambientes); e (*ii*) externos, que estão relacionados ao modo de transmissão do sinal acústico (diferentes caraterísticas de microfones, linhas de transmissão) (ALENCAR, 2005; SILVA, 2010).

III. Tipo de locutor: os sistemas de reconhecimento de voz podem ser classificados como: (*i*) dependentes, no qual o sistema é treinado somente para um locutor; e (*ii*) independentes, no qual são capazes de reconhecer a fala de qualquer locutor, mesmo aquele que não participou do treinamento (JURAFSKY; MARTIN, 2008; SILVA, 2010).

IV. Presença de ruído: os ruídos do ambiente, tais como vozes de outros locutores, sons de equipamentos, e, até mesmo, os provocados pelo próprio locutor (tosses, espirros), são inevitáveis e influenciam no conteúdo do sinal (ALENCAR, 2005; SILVA, 2010).

V. Limitações do corpus: para treinar modelos acústicos que abrangem toda a variabilidade do sinal de voz é necessário um corpus com uma grande variedade e quantidade de amostras. Entretanto, para o português brasileiro a disponibilidade de corpora de voz de grande porte é uma das principais limitações (VEIGA, 2013; SILVA, 2010).

Segundo Ferreira e Souza (2017), os sistemas de reconhecimento de voz atuais não possuem precisão absoluta, visto que é praticamente impossível resolver todos os

fatores de complexidade citados acima.

2.2 ESTRUTURA DO RECONHECIMENTO DE VOZ

A estrutura básica do reconhecimento de voz, representada na Figura 1, é dividida em duas etapas: extração de características e decodificador.

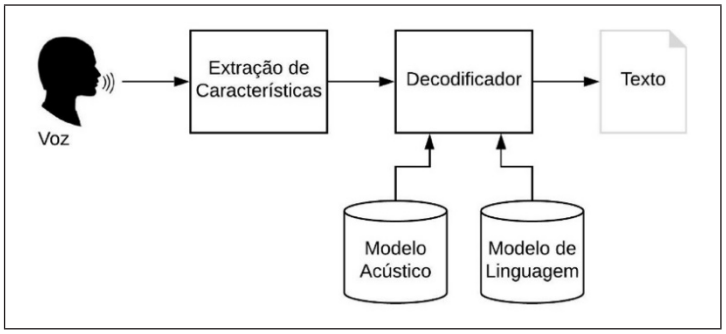

Figura 1 - Estrutura do reconhecimento de voz

Fonte: Adaptado de Silva (2010); Veiga (2013).

2.2.1 Extração de Características do Áudio

A extração e seleção da melhor representação paramétrica dos sinais acústicos é uma tarefa importante do sistema de reconhecimento de voz, visto que afeta significativamente no seu desempenho. Além disso, é importante focar na extração de características, pois um dos problemas do reconhecimento de voz é obter informações úteis do áudio (VEIGA, 2013).

Apesar de existirem muitas representações o MFCC é o mais utilizado para o reconhecimento de voz. O MFCC faz uma análise de características espectrais de curto prazo, baseando-se no uso do espectro da voz convertido em uma escala de frequências denominada Mel. Essa escala visa transcrever as características perceptíveis pelo ouvido humano, ou seja, as de baixo nível, pois são foneticamente mais importantes para a percepção humana do que as de alto nível (JURAFSKY; MARTIN, 2008; PATRA, 2007; VEIGA, 2013).

2.2.2 Decodificador

No decodificador, a sequência textual é concebida pelo modelo acústico e corrigida pelo modelo de linguagem. Esses modelos trabalham em conjunto e um depende do outro, pois, como já visto, existem palavras homófonas e é praticamente impossível para o modelo acústico diferenciá-las, por isso utiliza-se o modelo de linguagem (FERREIRA; SOUZA, 2017; JURAFSKY; MARTIN, 2008; RUSSELL; NORVIG, 2004; SILVA, 2010; VEIGA, 2013).

Segundo Veiga (2013), essa etapa necessita de treinamento, visto que é necessário gerar modelos que forneçam bons resultados e que sejam adequados ao contexto de

aplicação. No treinamento do modelo acústico, os vetores de características do sinal da voz são utilizados para determinar um padrão que melhor represente cada frase do corpus. Já o treinamento de linguagem é utilizado para modelar e compreender as regras gramaticais.

Para solucionar os problemas complexos do reconhecimento de voz são utilizados, nos modelos acústicos e de linguagem, algumas técnicas de IA (Inteligência Artificial), tais como HMM (*Hidden Markov Models*), ANN (*Artificial Neural Networks*) e NLP (*Natural Language Processing*).

2.3 MÉTRICAS DE AVALIAÇÃO

O desempenho do reconhecimento de voz depende da precisão dos modelos acústicos, da complexidade da tarefa definida pelo modelo de linguagem e da qualidade do sinal de áudio adquirido. Para isso, existem na literatura diversas funções matemáticas as quais podem ser chamadas de métricas de avaliação, e nesse trabalho foram utilizadas as seguintes: WER (*Word Error Rate*), SER (*Sentence Error Rate*) e xRT (*Real Time Factor*) (FERREIRA; SOUZA, 2017; VEIGA, 2013).

A WER é uma das métricas mais utilizadas em sistemas de reconhecimento de voz contínuo, ela se baseia na quantidade de palavras que foram inseridas, excluídas e substituídas incorretamente em comparação com a frase de referência. A SER representa quantas frases possuem pelo menos um erro (WER maior que 0%). Já o fator xRT é utilizado para calcular a velocidade do processo de reconhecimento de voz, é calculado dividindo o tempo que o sistema gasta para reconhecer uma sentença pela sua duração (FERREIRA; SOUZA, 2017; JURAFSKY; MARTIN, 2008).

3 I ESTADO DA ARTE

Esta seção apresenta a revisão sistemática da literatura cujo objetivo foi identificar e selecionar as principais técnicas utilizadas na extração de características do áudio e na implementação dos modelos, acústico e de linguagem, para o desenvolvimento do reconhecimento de voz contínuo.

Os artigos obtidos foram selecionados em função das seguintes perguntas de pesquisa: (*i*) quais técnicas estão sendo utilizadas na implementação do modelo acústico do reconhecimento de voz contínuo?; (*ii*) quais técnicas estão sendo utilizadas na implementação do modelo de linguagem para aperfeiçoar o reconhecimento de voz contínuo?; e (*iii*) quais soluções estão sendo estudadas para reduzir as taxas de erros do reconhecimento de voz contínuo?.

Para responder as perguntas, foram selecionados quatro repositórios: ACM, IEEE, Scopus e ScienceDirect. Utilizou-se a seguinte expressão de busca: ("*continuous speech recognition*" E ("*acoustic models*" OU "*neural networks*" OU ann OU "*deep learning*") E ("*language models*" OU lm OU *n-gram* OU "*natural language processing*"

OU nlp)). Ao analisar esta expressão, é possível observar que foram utilizadas palavras chaves que remetem ao problema do estudo. Em seguida, foram definidos os critérios de escolha dos artigos:

I. Critérios de inclusão: (*i*) artigos publicados entre 01/01/2014 até 31/12/2018; (*ii*) expressão de busca filtrando os artigos por meio do título, resumo e palavras-chave; e (*iii*) artigos em inglês e português.

II. Critérios de exclusão: (*i*) artigos que não possuem resultados relacionados ao desenvolvimento do reconhecimento de voz contínuo; (*ii*) artigos que não apresentam o desenvolvimento dos modelos; (*iii*) ausência de especificação das técnicas utilizadas no desenvolvimento dos modelos; e (*iv*) artigos curtos (5 páginas ou menos).

Realizou-se a seleção desses artigos com a leitura dos seguintes tópicos: (*i*) título e palavras chaves utilizadas; (*ii*) resumo; e (*iii*) introdução e conclusão. Esses critérios para a leitura e seleção foram úteis para minimizar o esforço.

O Quadro 1 apresenta os artigos selecionados, cada qual com a sua referência e as três perguntas de pesquisa que foram respondidas de maneira sintetizada. Os artigos estão ordenados por ordem alfabética.

Identificação	Modelo Acústico	Modelo de Linguagem	Solução para reduzir a WER
Abushariah (2017)	HMM	1, 2 e 3 grama	-
Georgescu, Cucu e Burileanu (2017)	RNN e HMM	1, 2 e 3 grama	Utilizando modelos acústicos baseados em DNN
Kipyatkova e Karpov (2017)	HMM	RNN-LM e 3-grama	Com um modelo de linguagem baseado em RNN e 3-grama
LAleye et al. (2016)	“monofone” e “trifone” (Kaldi)	3-grama (SRILM)	Removendo diacríticos de tons do modelo de linguagem
Naing et al. (2015)	MLP e HMM	*Word-base*	Utilizando DNN
Pakoci, Popović e Pekar (2017)	MLP e HMM	1, 2 e 3 grama	Combinando o uso de DNN, HMM e modelo de linguagem
Pakoci, Popović e Pekar (2018)	MLP, RNN e HMM	1, 2 e 3 grama	Utilizando DNN de 8 camadas com 625 neurônios cada
Phull e Kumar (2016)	HMM e HMM	2 e 3 grama	-
Tachbelie, Abate e Besacier (2014)	HMM	3-grama (SRILM)	Utilizando unidades acústicas de sílabas baseado em morfema
Zhang, Bao e Gao (2015)	MLP e HMM	2 e 3 grama	Utilizando DNN em conjunto com HMM

Quadro 1 - Respostas das perguntas de pesquisa

As principais técnicas utilizadas no desenvolvimento do modelo acústico foram o HMM e as ANN: MLP (*Multilayer Perceptron*) e RNN (*Recurrent Neural Network*), que também são conhecidos como sendo DNN (*Deep Neural Network*). O modelo 3-grama foi a principal técnica utilizada no modelo de linguagem. Além disso, a grande maioria dos artigos apresentam alguma solução para reduzir as taxas de erros do reconhecimento de voz, e pode-se notar que, em muitos casos, essa solução está associada ao uso de DNN.

Com a leitura dos artigos também foi possível extrair diversas particularidades essenciais para o reconhecimento de voz: (*i*) as bibliotecas mais utilizadas para facilitar a implementação dos modelos (acústico e de linguagem) são CMUSphinx, HTK e Kaldi; e (*ii*) o principal método de extração de características do áudio é por meio do MFCC.

4 I DESENVOLVIMENTO

Esta seção descreve os procedimentos que foram utilizados para o desenvolvimento do reconhecimento off-line de voz contínuo.

4.1 INSTALAÇÃO DAS BIBLIOTECAS

A rapidez no desenvolvimento do reconhecimento de voz contínuo foi um dos principais motivos para a utilização de bibliotecas, porém é necessário que elas estejam em constante atualização e que tenham uma boa documentação. As bibliotecas CMUSphinx, HTK e Kaldi são as mais utilizadas nos artigos selecionados e atendem aos requisitos previamente mencionados. Essas bibliotecas são gratuitas e foram instaladas em um computador com o sistema operacional Antergos (distribuição Linux baseada em Arch Linux) versão 64 bits.

4.1.1 CMUSphinx

O Sphinx é uma biblioteca para reconhecimento de voz contínuo independente de locutor. Essa biblioteca utiliza a técnica HMM no modelo acústico e a técnica n-grama no modelo de linguagem. Ela possui diversos pacotes de bibliotecas para diferentes tarefas e aplicações. Nesse trabalho utilizou-se o pacote PocketSphinx que é ideal para ser utilizada em sistemas embarcados. Esse pacote é escrito na linguagem de programação C (CMUSPHINX, 2019; LEE; HON; REDDY, 1990).

A biblioteca PocketSphinx depende da instalação de outras bibliotecas: (*i*) SphinxBase, que é a base para todos os projetos CMUSphinx; e (*ii*) SphinxTrain, que fornece ferramentas de treinamento de modelo acústico. Nesse trabalho utilizou as bibliotecas “pocketsphinx”, “sphinxbase” e o “sphinxtrain” da versão “5prealpha” (CMUSPHINX, 2019).

4.1.2 HTK

O HTK, disponível na linguagem de programação C, é um kit de ferramentas para construir e manipular HMMs. O HTK é usado principalmente para a pesquisa de reconhecimento de voz. Além disso, as ferramentas do HTK fornecem recursos sofisticados para análise de fala, treinamento de HMM, testes e análise de resultados (HTK, 2019).

Embora a Microsoft mantenha os direitos autorais do código-fonte do HTK, os desenvolvedores podem fazer alterações e contribuí-los para inclusão nas futuras versões. Esse projeto inclui o HTKBook que é uma documentação detalhada sobre cada funcionalidade. Nesse trabalho foi utilizada a versão estável 3.4.1 da biblioteca (HTK, 2019).

4.1.3 Kaldi

É um kit de ferramentas de código aberto para reconhecimento de voz desenvolvido na linguagem de programação C++. A biblioteca está hospedada na plataforma GitHub e neste trabalho utilizou-se a versão 5.5. Essa biblioteca possui uma estrutura baseada em transdutor de estado finito, amplo suporte à álgebra linear e uma licença não restritiva. Além disso, possui suporte a DNN, tais como, MLP e RNN (KALDI, 2019; POVEY et al., 2011).

O Kaldi possui dependências de ferramentas externas, e nesse trabalho foram utilizadas as seguintes: (*i*) OpenFST, é a ferramenta mais importante para o Kaldi, pois como visto a estrutura da biblioteca é baseada em transdutor de estado finito; (*ii*) OpenBLAS, é responsável pelo suporte à álgebra linear; e (*iii*) SRILM (*SRI Language Modeling Toolkit*), responsável por executar o modelo de linguagem (KALDI, 2019). A biblioteca possui uma documentação detalhada sobre cada funcionalidade.

4.2 PREPARAÇÃO DOS CORPORA DE VOZ

Nesse trabalho, foram utilizados os corpora disponíveis no website do grupo FalaBrasil, porém esses corpora possuem poucas horas de duração. Por isso, o modelo de linguagem é constituído apenas pelas frases que estão presentes em cada corpus de voz, isto é, reconhece apenas as palavras que estão nos corpora.

O corpus LaPS Benchmark é composto por 700 frases e possui 35 locutores com 20 frases cada, sendo 25 homens e 10 mulheres, o que corresponde a aproximadamente 54 minutos de áudio. Todas as gravações foram realizadas utilizando microfones comuns, e o ambiente não é controlado. Já o corpus de voz Constituição Federal é composto por 1.255 frases com aproximadamente 30 segundos de duração cada, totalizando aproximadamente 9 horas de áudio com apenas um locutor do sexo masculino. Além disso, utilizou-se um ambiente de gravação controlado, isto é, com pouca presença de ruído.

Foi necessário dividir os corpora de voz em arquivos de áudio de treinamento e testes. Dividiu-se o corpus LaPS Benchmark em: (*i*) treinamento com 30 locutores (640 arquivos), sendo 23 do sexo masculino e 9 do sexo feminino; e (*ii*) testes com 3 locutores (60 arquivos), sendo 2 do sexo masculino e 1 do sexo feminino. Já o corpus Constituição Federal foi dividido em 1.129 arquivos (90%) para treinamento e 126 arquivos (10%) para testes.

Para utilizar estes corpora nas bibliotecas foi necessário realizar a preparação dos dados. Para isso, desenvolveu-se a ferramenta SCT (*Speech Corpus Treatment*) que foi desenvolvida na linguagem de programação Java, e tem como objetivo realizar a preparação dos dados dos dois corpora de voz de acordo com a documentação das bibliotecas CMUSphinx, HTK e Kaldi. Essa ferramenta, também possibilita a redução da taxa de amostragem dos arquivos de áudio das bibliotecas. Nesse trabalho todos os arquivos de áudio dos dois corpora de voz foram alterados para 16000 Hz de taxa de amostragem.

4.3 IMPLEMENTAÇÃO DO TREINAMENTO

Para cada biblioteca e para cada corpus foi criado um diretório de treinamento que possui as configurações e os arquivos do modelo acústico do corpus, ambos criados de acordo com a documentação das bibliotecas. No total foram executados 30 treinamentos em cada corpus, isto é, 10 arquivos com diferentes configurações para cada biblioteca. Esses 10 arquivos foram divididos em dois conjuntos, modificando a lógica de alteração das configurações, conforme destacados abaixo:

I. Biblioteca CMUSphinx: foram alterados os valores da configuração padrão da biblioteca de modo uniforme: (*i*) reduziu-se os valores em 80%; (*ii*) reduziu-se os valores em 40%; (*iii*) manteve a configuração padrão; (*iv*) aumentou-se os valores em 40%; e (v) aumentou-se os valores em 80%. Além disso, gerou-se mais cinco arquivos de configuração alterando os valores padrão previamente mencionados e alterando as opções textuais dos parâmetros de configuração.

II. Biblioteca HTK: os valores da configuração padrão da biblioteca foram alterados em -80%, -40%, 0%, 40% e 80%. Além disso, gerou-se mais cinco arquivos de configuração alterando as configurações padrão de extração de características do áudio, usando a mesma lógica de alteração uniforme, -80%, -40%, 0%, 40% e 80%.

III. Biblioteca Kaldi: a configuração padrão da biblioteca foi alterada para os valores de -80%, -40%, 0%, 40% e 80%. Entretanto, cinco configurações utilizaram DNN e as outras cinco não.

O treinamento das bibliotecas criou em cada diretório os seguintes arquivos de

resultados para cada configuração: (*i*) a data e hora de início e fim; e (*ii*) a saída da biblioteca com o valor das métricas de avaliação SER e WER.

4.4 IMPLEMENTAÇÃO DOS TESTES

Para os testes das bibliotecas foram: (*i*) implementados apenas os códigos-fonte que são responsáveis por gerar a saída dos modelos já treinados; e (*ii*) utilizados apenas os arquivos de testes dos corpora. Essa aplicação foi responsável por verificar o desempenho das bibliotecas em um computador desktop. Na implementação, utilizou-se a linguagem de programação C++ em conjunto com o Qt SDK (*Software Development Kit*).

A Figura 2 apresenta a tela para desktop, que possui as seguintes opções de configuração: (*i*) biblioteca, que é responsável por selecionar a biblioteca que será utilizada nos testes; e (*ii*) corpus de voz, que é responsável por selecionar o corpus de voz que será utilizado nos testes. O botão “Testar” gera os arquivos de resultados para a biblioteca e corpus selecionados.

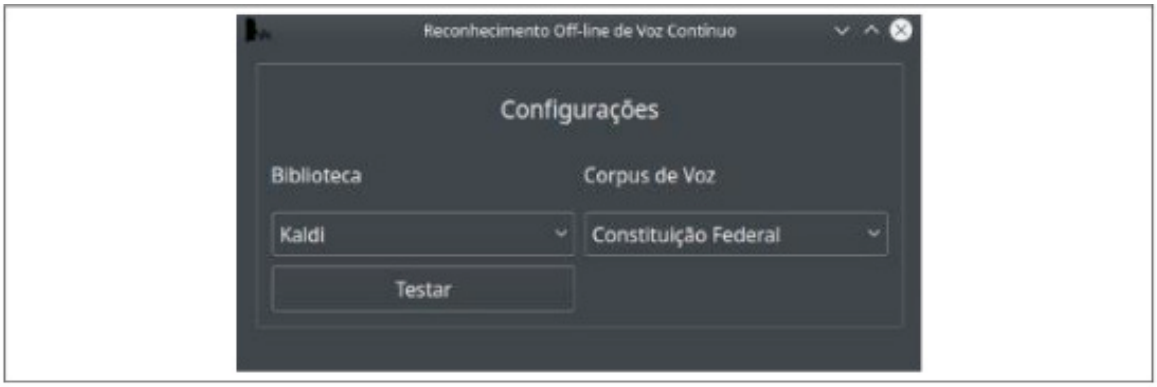

Figura 2 - Tela para testes

Para cada biblioteca e corpus são gerados três arquivos de resultados: (*i*) data e hora de início; (*ii*) uso de processador e memória (capturado utilizando o comando Linux “top”); e (*iii*) data e hora de encerramento.

5 I RESULTADOS

Esta seção apresenta e discute os resultados do projeto, permitindo avaliar a sua contribuição e o alcance dos seus objetivos. O Quadro 2 apresenta a configuração do computador desktop utilizado para o treinamento e teste.

Processador	Intel Core i5-7200U 2.50GHz
Memória	16 GB DDR4 2400MHz
Sistema operacional	Linux Antergos de 64 bits
Versão do Java	1.8.0_192
Versão do compilador GCC	8.2.1

Quadro 2 - Configuração do computador desktop utilizado

5.1 MELHORES CONFIGURAÇÕES DE TREINAMENTO

Esta seção apresenta as métricas de avaliação obtidas para cada arquivo de configuração. Com base no percentual das métricas foram escolhidas as melhores para realizar testes de desempenho em um computador desktop. As subseções a seguir apresentam os resultados obtidos para cada biblioteca.

5.1.1 CMUSphinx

O Quadro 3 apresenta os valores obtidos por cada configuração no corpus LaPS Benchmark e no corpus Constituição Federal. A coluna ID é um identificador da configuração e servirá como referência para os resultados da análise de desempenho. Nesse trabalho, a unidade dos valores do tempo de duração foi hora, minuto e segundo (hh:mm:ss).

	LaPS Benchmark			Constituição Federal		
ID	**WER**	**SER**	**Duração**	**WER**	**SER**	**Duração**
1	49,4%	93,3%	00:03:08	14,5%	96,8%	00:22:56
2	9,6%	55%	00:03:01	5%	77,8%	00:22:42
3	9,1%	56,7%	00:03:17	3,2%	75,4%	00:27:17
4	11,4%	61,7%	00:05:29	4,6%	81%	00:30:29
5	22,1%	68,3%	00:08:21	5,6%	81,7%	00:32:18
6	38%	81,7%	00:07:53	10,8%	90,5%	00:40:43
7	8,8%	41,7%	00:04:48	3%	70,6%	00:43:36
8	6,7%	45%	00:05:37	2,2%	59,5%	00:56:03
9	6,2%	35%	00:09:13	2%	62,7%	01:17:39
10	11,9%	50%	00:11:39	3,4%	67,5%	01:24:58

Quadro 3 - Configurações da biblioteca CMUSphinx

De acordo com o Quadro 3 pode-se notar que o as melhores configurações obtidas possuem um WER inferior a 10%. Além disso, pode-se observar que as melhores métricas de avaliação são provenientes de configurações cujo os valores alterados são próximos dos valores de configuração padrão da biblioteca. Além disso, esses resultados demonstram que os testes com apenas um locutor possuem maior precisão (WER) do que com vários locutores.

5.1.2 HTK

A biblioteca HTK apresentou os piores resultados, pois com qualquer configuração e em qualquer corpus não foi possível ter um WER abaixo de 80%. O Quadro 4 apresenta os valores das métricas de avaliação obtidos pela biblioteca no corpus LaPS Benchmark e no corpus Constituição Federal. A classe HVite (2-grama) obteve

os melhores resultados.

	LaPS Benchmark			Constituição Federal		
ID	**WER**	**SER**	**Duração**	**WER**	**SER**	**Duração**
1	94,95%	100%	00:05:19	91,98%	100%	02:08:11
2	92,18%	100%	00:05:33	89,64%	100%	02:12:03
3	93,32%	100%	00:05:25	86,17%	100%	02:09:21
4	92,35%	100%	00:05:21	86,47%	100%	02:06:09
5	93,49%	100%	00:05:25	84,97%	100%	02:06:32
6	95,44%	100%	00:05:21	93,85%	100%	01:58:51
7	93,00%	100%	00:05:25	88,60%	100%	02:03:42
8	92,83%	100%	00:05:19	83,84%	100%	02:08:38
9	92,83%	100%	00:05:18	82,47%	100%	02:11:46
10	92,18%	100%	00:05:17	84,21%	100%	02:19:31

Quadro 4 - Configurações da biblioteca HTK

Ao realizar uma comparação com os resultados obtidos pela biblioteca CMUSphinx (Quadro 3) pode-se observar que os resultados do Quadro 4 foram insignificantes. Além disso, pode-se notar que todas as frases possuíram um ou mais erros.

5.1.3 Kaldi

O Quadro 5 apresenta os valores obtidos por cada configuração da biblioteca Kaldi no corpus LaPS Benchmark e no corpus Constituição Federal. Essa biblioteca também possui diversas classes que são utilizadas para gerar os valores de saída, e as melhores foram representadas na coluna Tipo.

	LaPS Benchmark				Constituição Federal			
ID	**Tipo**	**WER**	**SER**	**Duração**	**Tipo**	**WER**	**SER**	**Duração**
1	MLP	5,05%	41,67%	00:49:17	MLP	1,44%	46,03%	07:45:17
2	RNN	2,61%	21,67%	02:24:06	RNN	0,98%	41,27%	26:38:27
3	RNN	6,19%	41,67%	14:18:37	RNN	0,93%	38,89%	166:47:14
4	RNN	8,14%	51,67%	92:54:37	N/A	N/A	N/A	N/A
5	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
6	tri3b	5,05%	40%	00:23:41	tri3b	1,48%	46,03%	01:50:37
7	mono0a	7,17%	38,33%	00:45:13	tri1	1,61%	49,21%	01:54:36
8	tri1	6,68%	43,33%	00:42:41	tri1	1,34%	38,89%	02:02:11
9	tri1	5,37%	31,67%	00:38:46	tri1	1,75%	46,03%	05:32:35
10	tri1	6,51%	41,67%	00:40:57	tri1	1,80%	48,41%	07:09:14

Quadro 5 - Configurações da biblioteca Kaldi

De acordo com o Quadro 5 pode-se perceber que o uso de ANNs apresentaram os melhores resultados, entretanto o único problema é o elevado tempo gasto para o treinamento. Por isso, as configurações de ID 4 (Constituição Federal) e ID 5 (ambos)

não foram treinadas, pois levaria mais de 100 horas. Os melhores resultados obtidos foram utilizando RNN e MLP. Além disso, as classes tri3b e tri1 também apresentaram bons resultados.

5.2 ANÁLISE DE DESEMPENHO

Os testes foram realizados em apenas um computador desktop, conforme configuração já destacada no Quadro 2. Esses testes foram realizados sem conexão com a internet, sem nenhum software sendo executado em paralelo e utilizando apenas os arquivos de testes dos corpora. O Quadro 6 apresenta os resultados obtidos no corpus LaPS Benchmark. Para o cálculo do fator xRT utilizou-se a duração total dos arquivos de teste do corpus, que é 00:04:36.

Biblioteca	ID	Duração	xRT	Processador	Memória
CMUSphinx	9	00:01:01	0,221	Máx.: 106,70% Méd.: 99,20% Mín.: 93,30%	Máx.: 0,50% Méd.: 0,50% Mín.: 0,50%
HTK	10	00:02:41	0,583	Máx.: 106,70% Méd.: 98,79% Mín.: 87,50%	Máx.: 0,60% Méd.: 0,60% Mín.: 0,50%
Kaldi	2	00:03:31	0,764	Máx.: 100,00% Méd.: 97,79% Mín.: 66,70%	Máx.: 0,40% Méd.: 0,39% Mín.: 0,20%

Quadro 6 - Desempenho das bibliotecas no corpus LaPS Benchmark

Para cada biblioteca foi selecionada a melhor configuração. No Quadro 6 pode-se observar que a biblioteca Kaldi se destacou, pois obteve: (*i*) o menor percentual WER; (*ii*) a menor média de uso do processador; e (*iii*) a menor média de uso de memória. Já o Quadro 7 apresenta os resultados de desempenho obtidos utilizando o corpus Constituição Federal. Para o cálculo do fator xRT utilizou-se a duração total dos arquivos de teste do corpus, que é 00:53:06.

Biblioteca	ID	Duração	xRT	Processador	Memória
CMUSphinx	9	00:07:45	0,146	Máx.: 106,70% Méd.: 98,92% Mín.: 81,20%	Máx.: 0,50% Méd.: 0,50% Mín.: 0,50%
HTK	9	01:26:07	1,622	Máx.: 113,30% Méd.: 98,52% Mín.: 75,00%	Máx.: 0,70% Méd.: 0,70% Mín.: 0,50%
Kaldi	3	02:39:45	3,008	Máx.: 100,00% Méd.: 98,11% Mín.: 81,20%	Máx.: 1,00% Méd.: 0,95% Mín.: 0,70%

Quadro 7 - Desempenho das bibliotecas no corpus Constituição Federal

Para esse corpus também foi selecionada a melhor configuração de cada biblioteca. No Quadro 7 pode-se observar que as bibliotecas CMUSphinx e Kaldi

possuíram os melhores resultados. Nesse corpus a biblioteca Kaldi também se destacou, pois obteve: (*i*) o menor percentual WER; (*ii*) o menor valor xRT; (*iii*) a menor média de uso do processador; e (*iv*) a menor média de uso de memória.

De acordo com o Quadro 6 e o Quadro 7, as bibliotecas utilizaram mais recursos de processamento do que de memória RAM. Além disso, o uso de ANNs apresentou os melhores resultados (2,61% corpus LaPS Benchmark e 0,93% corpus Constituição Brasileira), entretanto, requer um grande custo computacional, em alguns casos ultrapassando duas horas de processamento.

6 I CONCLUSÃO E TRABALHOS FUTUROS

Essa pesquisa teve como objetivo principal o desenvolvimento do reconhecimento off-line de voz contínuo do português brasileiro e a análise do seu desempenho em um computador desktop.

As melhores configurações foram selecionadas com base nos valores das métricas de avaliação WER obtidos nos treinamentos. Essas configurações foram testadas em um computador desktop visando encontrar a biblioteca que possuísse os melhores resultados de desempenho. A biblioteca Kaldi apresentou os melhores resultados, já a biblioteca HTK apresentou os piores resultados, pois o menor valor WER obtido foi maior que 80%. Os testes das bibliotecas foram realizados sem conexão com a internet, buscando comprovar que este reconhecimento de voz contínuo desenvolvido funciona com êxito em ambientes que não possuem internet.

A principal contribuição desse trabalho foi o desenvolvimento do reconhecimento off-line de voz contínuo para o português brasileiro e sua implementação em computadores desktop. Esse reconhecimento de voz desenvolvido poderá ser utilizado em softwares e aplicativos: (*i*) que auxiliam na comunicação de pessoas com deficiência; (*ii*) empresariais que agilizam o trabalho dos funcionários; e (*iii*) que necessitam desta função em áreas sem conexão com a internet.

Ao longo do desenvolvimento deste trabalho, puderam ser identificadas algumas possibilidades de melhoria e de continuação a partir de futuras pesquisas, as quais incluem:

- A criação de um novo corpus de voz com vários locutores, para o português brasileiro e com no mínimo dez horas de duração, a fim de se obter resultados de desempenho fidedignos a um cenário real de aplicação, isto é, sem um vocabulário restrito;
- Implementação das melhores bibliotecas em dispositivos móveis;
- A redução do custo computacional (processamento) exigido pelas ANNs que realizam o reconhecimento de voz contínuo;
- A comparação das métricas de avaliação do reconhecimento off-line de voz contínuo desenvolvido com as APIs existentes no mercado;

- O teste do reconhecimento off-line de voz contínuo em sistemas embarcados, a fim de obter o desempenho em situações com recursos de processamento e memória limitados.

REFERÊNCIAS

ABUSHARIAH, M. A. TAMEEM V1.0: speakers and text independent Arabic automatic continuous speech recognizer. **International Journal of Speech Technology**, Nova Iorque, v. 20, n. 2, p. 261-280, jun. 2018.

ALENCAR, V. F. S. **Atributos e Domínios de Interpolação Eficientes em Reconhecimento de Voz Distribuído**. 2005. Dissertação (Engenharia Elétrica) – Departamento de Engenharia Elétrica, PUC-Rio, Rio de Janeiro.

CMUSPHINX. **Open source speech recognition toolkit**. 2019. Disponível em: <https://cmusphinx.github.io>. Acesso em: 06 jan. 2019.

DEBATIN, L.; HAENDCHEN FILHO, A.; DAZZI, R. L. S. Offline Speech Recognition Development: A Systematic Review of the Literature. In: ICEIS, 20., 2018, Funchal. **Proceedings...** Setúbal: SciTePress, 2018. p. 551-558.

FERREIRA, M. V. G.; SOUZA, J. F. Use of Automatic Speech Recognition Systems for Multimedia Applications. In: WEBMEDIA, 23., 2017, Gramado. **Anais dos Workshops e Pôsteres do Webmedia**. Porto Alegre: SBC, 2017. p. 139-176.

GEORGESCU, A.; CUCU, H.; BURILEANU, C. SpeeD's DNN approach to Romanian speech recognition. In: SPED, 9., 2017, Bucharest. **Proceedings...** Piscataway: IEEE, 2017. p. 1-8.

GROSSMAN, R. L. The Case for Cloud Computing. **IT Professional**, Piscataway, v. 11, n. 2, p. 23-27, mar. 2009.

HTK. **HTK Speech Recognition Toolkit**. 2019. Disponível em: <http://htk.eng.cam.ac.uk>. Acesso em: 10 jan. 2019.

IBGE. **Pesquisa nacional por amostra de domicílios**. 2016. Disponível em: <http://www.ibge.gov.br/>. Acesso em: 11 mai. 2018.

JURAFSKY, D.; MARTIN, J. H. **Speech and Language Processing:** An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. 2. ed. Upper Saddle River: Prentice-Hall, 2008.

KALDI. **Kaldi ASR**. 2019. Disponível em: <http://kaldi-asr.org>. Acesso em: 08 jan. 2019.

KIPYATKOVA, I.S.; KARPOV, A.A. A study of neural network Russian language models for automatic continuous speech recognition systems. **Automation and Remote Control**, Nova Iorque, v. 78, n. 5, p. 858-867, mai. 2017.

LALEYE, F. A. A; BESACIER, L.; EZIN, E. C.; MOTAMED, C. First automatic fongbe continuous speech recognition system: Development of acoustic models and language models. In: FEDCSIS, 5., 2016, Gdansk. **Proceedings...** Piscataway: IEEE, 2016. p. 477-482.

LEE, K.; HON, H.; REDDY, R. An Overview of the SPHINX Speech Recognition System. **IEEE Transactions on Acoustic Speech, and Signal Processing**, Piscataway, v. 38, n. 1, p. 35-45, jan. 1990.

MÜLLER, D. N. **COMFALA - Modelo Computacional do Processo de Compreensão da Fala**. 2006. Tese (Ciência da Computação) – Instituto de Informática, UFRGS, Porto Alegre.

NAING, H. M. S.; HLAING, A. M.; PA, W. P.; HU, X.; THU, Y. K.; HORI, C.; KAWAI, H. A Myanmar large vocabulary continuous speech recognition system. In: APSIPA, 3., 2015, Hong Kong. **Proceedings...** Piscataway: IEEE, 2015. p. 320-327.

PAKOCI, E.; POPOVIĆ, B.; PEKAR, D. Improvements in Serbian Speech Recognition Using Sequence-Trained Deep neural Networks. In: ARTIFICIAL INTELLIGENCE, KNOWEDGE AND DATA ENGINEERING, 2018, St. Petersburg. **SPIIRAS Proceedings.** St. Petersburg: SPIIRAS, 2018. p. 53-76.

PAKOCI, E.; POPOVIĆ, B.; PEKAR, D. Language model optimization for a deep neural network based speech recognition system for Serbian. In: SPECOM, 19., 2017, Hatfield. **Proceedings…** Nova Iorque: Springer, 2017. p. 483-492.

PATRA, S. **Robust Speaker Identification System**. 2007. Disponível em: <http://www.serc.iisc.ernet.in/graduation-theses/spatra_dec07.pdf >. Acesso em: 26 mai. 2018.

PHULL, D. K.; KUMAR, G. B. Investigation of Indian English Speech Recognition using CMU Sphinx. **International Journal of Applied Engineering Research**, Delhi, v. 11, n. 6, p. 4167-4174, 2016.

POVEY, D.; GHOSHAL, A.; BOULIANNE, G.; BURGET, L.; GLEMBEK, O.; GOEL, N.; HANNEMANN, M.; MOTLICEK, P.; QIAN, Y.; SCHWARZ, P.; SILOVSKY, J.; STEMMER, G.; VESELY, K. The Kaldi Speech Recognition Toolkit. In: ASRU, 12., 2011, Hawaii. **Proceedings…** Piscataway: IEEE, 2011.

RUSSELL, S.; NORVIG, P. **Inteligência artificial**. 2. ed. Rio de Janeiro: Elsevier, 2004.

SILVA, C. P. A. **Um software de reconhecimento de voz para português brasileiro**. 2010. Dissertação (Engenharia Elétrica) – Instituto de Tecnologia, UFPA, Pará.

TACHBELIE, M. Y.; ABATE, S. T.; BESACIER, L. Using different acoustic, lexical and language modeling units for ASR of an under-resourced language - Amharic. **Speech Communication**, Amsterdam, v. 56, n. 1 p. 181-194, jan. 2014.

VEIGA, A. O. **Treino não supervisionado de modelos acústicos para reconhecimento de fala**. 2013. Tese (Engenharia Eletrotécnica e de Computadores) – Departamento de Engenharia Eletrotécnica e de Computadores, Universidade de Coimbra, Coimbra.

ZHANG, H.; BAO, F.; GAO, G. Mongolian speech recognition based on deep neural networks. In: CHINESE COMPUTATIONAL LINGUISTICS AND NATURAL LANGUAGE PROCESSING BASED ON NATURALLY ANNOTATED BIG DATA, 14., 2015, Guangzhou. **Proceedings…** Nova Iorque: Springer, 2015. p. 180-188.

SOBRE A ORGANIZADORA

Vanessa Cristina de Abreu Torres Hrenechen: Graduada em Comunicação Social/Jornalismo (UEPG); mestre em Crítica de Mídia (UEPG). Tem 10 anos de experiência em assessoria de imprensa.

Atualmente é proprietária de agência de publicidade que presta serviços na área de marketing e comunicação empresarial.

Agência Brasileira do ISBN
ISBN 978-85-7247-344-6