MÉTODO DE TRANSFORMAÇÃO DE UMA SÉRIE TEMPORAL EM REDES DE CLIQUE COM O INTUITO DE ENCONTRAR O SOM DO "S" DE UMA FALANTE DE LÍNGUA PORTUGUESA

Cleônidas Tavares de Souza Junior[1]; Hernane Borges de Barros Pereira[2], Valter de Senna[3]
[1] Doutorando em Modelagem Computacional e Tecnologia Industrial; Bolsista FAPESB; cleonidas@gmail.com
[2] Professor Doutor; Centro Universitário SENAI CIMATEC; Salvador-BA; hbbpereira@gmail.com
[3] Professor Doutor; Centro Universitário SENAI CIMATEC; Salvador-BA; valter.senna@gmail.com

RESUMO

Durante o processo de transcrição de áudios em textos, alguns algoritmos de reconhecimento de voz falham quando têm que codificar algumas variedades da língua portuguesa (e.g. no Brasil a palavra "porta" tem diferente pronúncias). Novas estratégias de codificação da fala contribuirão para a redução ou eliminação desse problema. O objetivo desse artigo é criar uma nova estratégia de codificação e reconhecimento de pronúncias começando com o som "s" de uma pessoa falante de língua portuguesa. O novo método consiste em transformar trechos de uma série temporal (i.e. um discurso oral) em redes de clique e apontar quais delas se assemelham a clique que representa o som do "s". Nos resultados parciais é possível identificar que a estratégia proposta identifica a maior parte dos sons de "s", no entanto, também identifica outros sons cuja produção oral é próxima ao som "s". Em trabalhos futuros, pretende-se aprimorar a identificação do som do "s" e a distinção de sons vizinhos.

PALAVRAS-CHAVE: Série Temporal. Redes Complexas. Cliques. Sons da fala.

1. INTRODUÇÃO

Algoritmos de reconhecimento automático da voz são bem-sucedidos na distinção de alguns sons do português, mas ineficientes em outros. Manfiro[1] demonstra em seu artigo sobre reconhecimento e compreensão da voz que, além das variantes linguísticas regionais, pares mínimos de sons como o "f" e "v", em "*faca*" e "*vaca*", também são um problema para algoritmos interessados em interpretar comandos da voz.

Em experimentos realizados por Duarte[2] com os programas de reconhecimento de voz *IBM ViaVoice*, *Google Web Speech API* e o *software Coruja* foram observadas taxas de erro em torno de 35% (i.e. frases classificadas como pouco ou nada compreendidas pelos algoritmos).

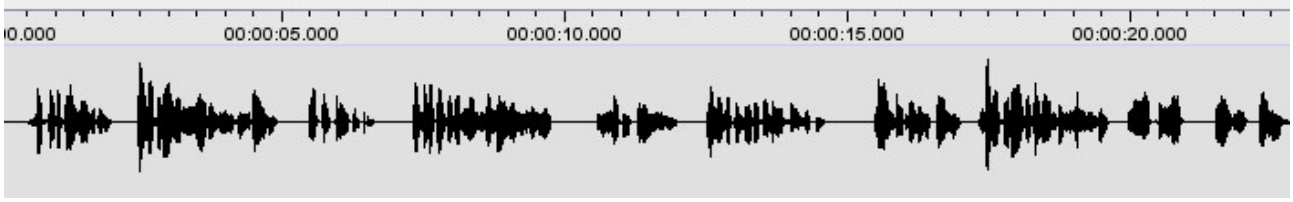
Novas estratégias para codificação da voz humana irão reduzir a falta de precisão na identificação de alguns sons da fala. Nesse aspecto, essa pesquisa buscou relacionar duas teorias: redes complexas e séries temporais. Para Newman[3] a maneira mais simples de definir uma rede é considerá-la como um conjunto de pontos (i.e. vértices) ligados em pares. As redes possibilitam analisar as relações entre objetos sobre outra perspectiva, por exemplo, a combinação entre os objetos de um conjunto gera um grafo completo ou *n* conjuntos de cliques. Antunes & Cardoso[4] definem séries temporais como uma forma de organizar informações quantitativas no tempo.

O objetivo deste artigo é criar uma nova estratégia de reconhecimento dos sons da língua portuguesa começando com o som de "s" de uma pessoa falante de português. A proposta é transformar um discurso oral em uma série temporal, extrair da série estruturas baseadas em redes de cliques, e avaliar se, dentre essas estruturas, é possível distinguir alguns sons típicos da língua portuguesa.

2. METODOLOGIA

O áudio usado nessa pesquisa foi um trecho de um discurso da monja Coen[5] no qual ela diz: "Se tudo que existe é a natureza iluminada se manifestando pra que a prática? Essa é uma pergunta de mestre *Dogen Zenji* sama do século XIII e a mesma pergunta que alguém me faz aqui. Qual o sentido de tudo se a natureza da nossa essência já é a pura luz e amor". São quinze palavras que apresentam o som o "s", a saber: *se, existe, se, manifestando, essa, mestre, sama, século, mesma, faz, sentido, se, nossa, essência, luz.* A Figura 01 exibe como ficaram as frequências do trecho extraído.

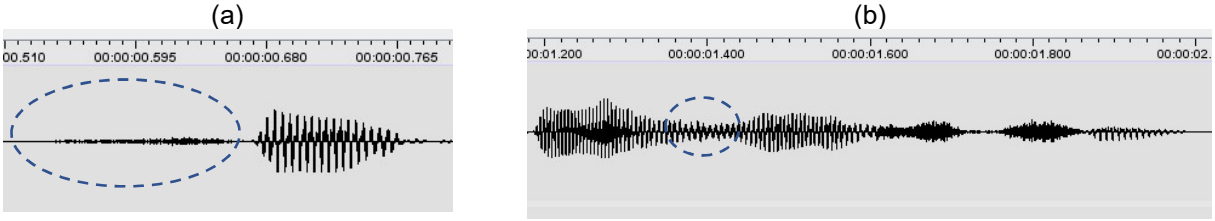
Figura 01: Frequências observadas no discurso oral



Depois de analisar os detalhes, identificou-se que os sons de “s” variam de frequência e amplitude ao longo da gravação. Quando sucedido pela vogal “e”, em “*se*”, a estrutura é uma (Figura 02 (a)); quando sucedido por uma consoante “t”, em “*existe*”, a estrutura é outra (Figura 02 (b)).

Figura 02: O som de “s” destacado em “se” (a) e em “existe” (b)

(a) (b)

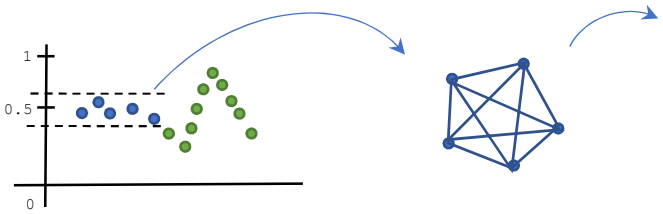


Para diminuir a diferença de amplitudes existentes entre as representações do som “s”, o discurso oral foi transformado em uma série temporal e mudou-se a sequência sonora para uma escala de valores que começa em zero e termina em um; sendo zero a ausência de som e um o som mais alto observado.

Uma amostra do som “s” foi extraída da série temporal (Figura 03 (a)) e, a partir dessa amostra, foi criada uma clique (Figura 03 (b)); devidamente armazenada em uma matriz de adjacência com cento e uma linhas e cento e uma colunas (Figura 03 (c)).

Figura 03: Identificação do som “s” e criação da clique

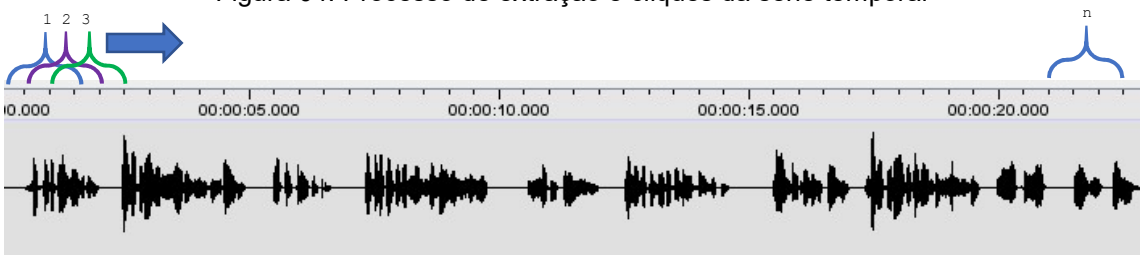
(a) (b) (c)



	0,00	(...)	0,48	0,49	0,5	0,51	0,52	(...)	1,00
0,00	0	0	0	0	0	0	0	0	0
(...)	0	0	0	0	0	0	0	0	0
0,48	0	0	0	**1**	**1**	**1**	**1**	0	0
0,49	0	0	**1**	0	**1**	**1**	**1**	0	0
0,5	0	0	**1**	**1**	0	**1**	**1**	0	0
0,51	0	0	**1**	**1**	**1**	0	**1**	0	0
0,52	0	0	**1**	**1**	**1**	**1**	0	0	0
(...)	0	0	0	0	0	0	0	0	0
1,00	0	0	0	0	0	0	0	0	0

O algoritmo proposto irá, a cada conjunto de *n* milissegundos (Figura 04), construir uma clique com os dados observados e compará-la com a amostra extraída inicialmente do som do “s”. A comparação consiste em subtrair uma matriz da outra; se a resposta for diferente de zero, o algoritmo avança uma unidade de tempo e faz uma nova extração/comparação; se a resposta for zero, o algoritmo registro em que período de tempo o suposto som foi encontrado. Ao final desse processo, uma lista é gerada com os pontos onde supostamente o som de “s” foi localizado.

Figura 04: Processo de extração e cliques da série temporal



Na primeira parte da pesquisa, para analisar as frequências e elaborar a transcrição do áudio, o programa ELAN[6] foi usado; optou-se por esse programa porque ele permite observar, controlar e fazer anotações dos áudios. Para desenvolver o algoritmo usou-se, além do programa R[7], a biblioteca tuneR[8] para transformar o discurso oral em uma série temporal.

3. RESULTADOS E DISCUSSÃO

Analisando intervalos de tempos de 0,04 segundos, o algoritmo identificou vinte e dois sons; treze sons de “s” e nove outros sons. Desses nove, três foram ruídos e os outros seis foram palavras que tem sons pronunciados em pontos de articulação na fala próximos ao som do “s” (e.g. “t” com chiado /txi/ da palavra “*prática*”; “d” com chiado /dgi/ da palavra “*de*”; “j” de “*já*”; “k” de “*aqui*” e “qual”).

Considerando intervalos de tempos maiores, outros sons começam a interferir no processo de identificação e reduzem, com isso, o número de acertos. Com intervalos menores, o mesmo som é identificado mais vezes e, consequentemente, o algoritmo retorna uma lista com os pontos de sons duplicados. O processo de identificação do som precisa ser ajustado com intervalos de tempo que acomodem o som a ser buscado.

4. CONSIDERAÇÕES FINAIS

O algoritmo proposto distinguiu do discurso oral a maior parte das palavras com som de “s”, porém, não distinguiu o som do “s” de outros sons próximos como /dgi/, em “de”, e /txi/ em “existe”. Isso aconteceu porque na boca o ponto de articulação do som do “s” é muito próximo dos pontos de articulações de outros sons e o parâmetro definido para detecção do “s” generalizou esses sons próximos.

O próximo passo da pesquisa é investigar a formação de clique nesses sons próximos de “s” e sugerir uma solução que possa distingui-los. À medida que os sons forem sendo distinguidos um dos outros, ficasse mais próximo de uma solução para o problema de identificação automática das variantes linguísticas na língua portuguesa brasileira.

Agradecimentos

Agradeço à FAPESB (BOL0241/2018) e aos colegas e professores do Senai pelo apoio.

5. REFERÊNCIAS

[1] MANFIO, Edio Roberto. Relação entre reconhecimento e compreensão de voz: experimento para análise linguística. **Entretextos**, Londrina, v. 17, n. 1, p. 281 - 300, jan./jun. 2017.

[2] DUARTE, Tiago da Silveira. **Máquinas de tradução aplicada à comunicação em tempo real para desenvolvimento distribuído de software**. 2014. Dissertação (Ciência da Computação) – Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre, 2014.

[3] NEWMAN, Mark E. J. **Networks: An Introduction**. New York: Oxford press, 2010.

[4] ANTUNES, José Leopoldo Ferreira; CARDOSO, Maria Regina Alves. Uso da análise de séries temporais em estudos epidemiológicos. **Epidemiol. Serv. Saúde**, Brasília, 24(3):565-576, jul-set 2015.

[5] Por que buscar a evolução se nossa essência é iluminada? In: Monja Coen Responde. Base de dados MOVA. 2019. Disponível em: <https://www.youtube.com/watch?v=_XfK8zniutg>. Acessado em: 12/03/2019.

[6] ELAN (Version 5.2) [Computer software]. (2018, April 04). Nijmegen: Max Planck Institute for Psycholinguistics. Retrieved from https://tla.mpi.nl/tools/tla-tools/elan/

[7] R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[8] LIGGES, Uwe; KREY, Sebastian; MERSMANN, Olaf; SCHNACKENBERG, Sarah. (2018). tuneR: Analysis of Music and Speech. URL: https://CRAN.R-project.org/package=tuneR