Determinação de tamanho mínimo de amostra para estimativa da taxa de produção da fala

Verônica Gomes Lima[1]*; Pablo Arantes*[1]

[1] Universidade Federal de São Carlos

vegomeslima@gmail.com; pabloarantes@gmail.com

Resumo

O objetivo do presente trabalho é investigar quanto tempo deve ter uma amostra de fala para que a taxa de produção de fala possa ser considerada representativa da taxa calculada globalmente, isto é, levando em conta toda a duração da amostra. Um total de 8 falantes do português brasileiro foi gravado lendo a passagem "A menina do narizinho arrebitado", do escritor Monteiro Lobato, que é foneticamente balanceada, isto é, contém pelo menos uma ocorrência dos fonemas do português brasileiro. As gravações foram feitas em três níveis: normal, lenta e rápida. Essas amostras foram analisadas após a segmentação das unidades em vogal a vogal (VV), fones, sílabas e palavras em séries temporais constituídas pelos valores de taxa de articulação e elocução calculadas de forma cumulativa ao longo de amostras de fala com o auxílio da técnica estatística *change point analysis* (KILLICK e ECKLEY 2014) para encontrar a quantidade de tempo necessário para se estabilizar em torno da taxa produzida por toda a amostra. Os resultados obtidos mostram o tempo médio de estabilização da taxa de articulação em 9.35 segundos e de elocução 8. O tempo médio de estabilização entre os três níveis de taxas (lenta, normal e rápida) tiveram uma diferença entre elas, que são: 7.81 segundos para a lenta, 8.41 para a normal e 11.15 para a rápida.

1. Introdução

Dentre as diversas características prosódicas da fala, tais como como entonação, ritmo e volume (*loudness*) (LEHISTE, 1970) trataremos da taxa de produção de fala, um parâmetro de cunho rítmico e temporal, que reflete quão rápido ou devagar a fala é produzida em um determinado enunciado. A taxa de produção pode ser definida pelo número de unidades linguísticas produzidas por unidades de tempo. A maneira como a pausa é incorporada ao cálculo da taxa define dois tipos: a taxa de elocução e taxa de articulação. A primeira inclui as pausas produzidas entre as unidades linguísticas e a segunda não, contabiliza apenas a atividade de fala efetivamente articulada.

Alguns estudos já tiveram como objeto a taxa de elocução, embora tenhamos conseguido encontrar na literatura apenas dois trabalhos que procuraram determinar a duração que uma amostra de fala deveria ter para que fosse suficiente se obter uma estimativa da taxa de elocução de um falante em questão para que possa ser representativa, Arantes e Lima (aceito) e Kendall (2013). Kendall se concentra em usar sílabas por segundo calculadas em "enunciados fonéticos" (que consistem em unidades ininterruptas de material fonético entre pausas silenciosas) como unidade linguística em seu estudo abordando questões relevantes acerca da quantidade de dados em determinadas amostras, como a definição da unidade linguística que será escolhida para a análise, também destacando o problema sobre definir o que será considerado como pausas silenciosas e como a determinação de diferentes limiares em torno de suas durações podem afetar os resultados das análises das amostras.

A preocupação com a determinação do tamanho mínimo de amostra se justifica por mais de uma razão. De modo geral, na pesquisa linguística é importante saber se o corpus a ser estudado é capaz de fornecer resultados que sejam representativos dos padrões linguísticos na população de interesse. Mais especificamente no contexto da prática forense, a duração da gravação questionada (aquela cuja identidade do falante é desconhecida) pode ser extremamente curta, contendo pouca atividade de fala, o que pode impedir certos parâmetros acústicos relevantes de serem analisados nas amostras. Para ser representativa, há a necessidade de determinar qual o tempo mínimo necessário de duração de uma amostra para que se possa usar esse dado de maneira que auxilie na identificação de indivíduos. Esta é a justificativa para o desenvolvimento deste trabalho com o propósito de avaliar e propor um método que permita estimar de maneira objetiva qual deve ser a duração mínima de uma amostra para que se extraia um valor representativo de taxa de produção de fala.

Pela falta de literatura explorando este aspecto da taxa de elocução, Arantes e Lima (aceito) produziram um estudo piloto usando como *corpus* gravações de voz de falantes nativos do português brasileiro lendo a passagem "A menina do narizinho arrebitado", do escritor Monteiro Lobato, que contém pelo menos uma ocorrência de cada um dos fonemas do português brasileiro. As leituras foram realizadas nas taxas de fala normal, lenta e rápida.

Naquele trabalho, os arquivos foram segmentados em unidades vogal a vogal (VV), unidade linguística constituída pelo intervalo entre dois ataques vocálicos consecutivos. Foram determinadas as taxas de elocução (cuja durações das pausas são incluídas) e articulação (a duração das pausas não são incluídas) nas três taxas de leitura realizadas (lenta, normal e rápida) e calculadas de forma cumulativa ao longo da duração de cada amostra com o intuito de obter suas taxas correspondentes. Posteriormente, foi aplicada a técnica estatística denominada *change point analysis* (KILLICK e ECKLEY, 2014) para encontrar o ponto de estabilização, que os autores definem como o momento no tempo em que a série temporal formada pelos valores consecutivos da taxa atinge um patamar de variabilidade que pode ser considerado estável.

Os resultados iniciais obtidos mostram que o tempo médio de estabilização (desvio-padrão entre parênteses) da taxa de articulação é de 9,35 (3,21) segundos e o da taxa de elocução é de 8,9 (3,19) s, não havendo diferença estatística significativa entre esses dois valores. O tempo médio de estabilização entre os três níveis de taxas (lenta, normal e rápida) tiveram uma diferença que pode ser considerada significativa: 7,81 (2,88) segundos para a lenta, 8,41 (2,72) para a normal e 11,15 (3) para a rápida.

2. Objetivos

O principal objetivo deste trabalho é ampliar os resultados obtidos no estudo de Arantes e Lima (aceito). Essa ampliação se dará pela exploração dos seguintes aspectos ainda não explorados:

1. Comparação dos tempos de estabilização de outras unidades de agrupamento do material linguístico além da unidade VV, a única analisada em Arantes e Lima (aceito), como o fone, a sílaba e a palavra.
2. Comparação entre dois diferentes critérios para a contagem de unidades linguísticas: critério fonético e fonológico.

3. Materiais e Métodos

3.1 Segmentação dos dados acústicos

O mesmo *corpus* constituído pela passagem de Lobato no estudo inicial de Arantes e Lima (aceito) é analisado com a expansão do número de unidades linguísticas usadas para determinar as taxas de elocução e articulação. A primeira etapa consistiu na segmentação de todos os arquivos de áudio em unidades VV, que são unidades que consistem na demarcação de fronteira no início de uma vogal até que se demarque uma fronteira final no momento em que uma nova vogal se iniciará (ou seja, no momento em que a forma de onda indique que uma consoante terminou de ser produzida e deu lugar a uma vogal), fones, sílabas e palavras. Foi necessário reorganizar estas unidades VV pois foram definidos alguns princípios para que o processo de segmentação estivesse completamente consistente e padronizado em todos os arquivos. Os seguintes parâmetros nortearam os segmentos constituídos por fones, sílabas e palavras além das unidades VV:

1. seguir as características típicas que definem o formato de onda das vogais, oclusivas, fricativas, laterais, nasais, vibrantes e tepes observando periodicidade, aperiodicidade (como períodos de obstruções e ruídos transientes), amplitude e outros parâmetros, recorrendo também ao espectrograma para visualizar características que permitem melhorar o efeito visual de identificação dos segmentos da fala como os formantes e suas transições, presença de ruídos, a barra de sonoridade ou não-sonoridade e antirressonâncias. Estas características serão a base das demarcações das fronteiras das unidades VV, fones, sílabas e palavras;
2. a demarcação das fronteiras deve ocorrer sempre no ponto de cruzamento de zero na forma de onda;
3. o trecho de segmentação consiste na duração total do arquivo, não podendo se limitar apenas a trechos;
4. a identificação de cada elemento da passagem deve seguir a notação Sampa-PB (substituta para os símbolos do IPA em contextos nos quais certos símbolos se tornam de difícil acesso, de acordo com seu nível correspondente: VV, fone, sílaba e palavra, tanto na transcrição fonética quanto na fonológica. Os níveis como a sílaba e a palavra compreendem a combinação de mais de um símbolo em sua formação;
5. Demarcar a presença de pausas na camada do objeto TextGrid que compreende a taxa de articulação, na qual se constituirá de um segmento vazio. A camada que corresponde a taxa de elocução conterá segmentos com duração maior que a média de um determinado fone, sílaba ou palavra pois acumulará a duração do mesmo somado com a duração da pausa em questão.
6. Realizar a transcrição fonética e fonológica separadamente para cada amostra. Cada tipo de transcrição resulta em uma quantidade de elementos diferentes no interior dos segmentos mesmo que estejam localizados no mesmo ponto da segmentação fonética para a fonológica.

Os parâmetros de segmentação, uma vez definidos, se aliam à decisão de transcrever não somente foneticamente, mas também fonologicamente, pois através da definição do que abrange cada tipo de transcrição, torna-se inevitável que seja incluído na análise. Sabe-se que a transcrição fonética se define por representar a fala realmente pronunciada e a fonológica representa todos os elementos que compõem a fala uma vez que mesmo não sendo propriamente pronunciados estão presentes na memória do falante.

A segmentação foi feita através do *software* de análise acústica Praat (2001) que se une a estes parâmetros na criação das camadas que representam a elocução e a articulação, nas quais um elemento quando for delimitado, em ambas as camadas haverá sua respectiva notação que será transcrita de forma fonética em um objeto TextGrid e de forma fonológica em um segundo objeto, sendo mantidos separadamente.

Künzel (1997) inicia uma discussão acerca da contabilização de material linguístico na taxa de articulação pertinentes ao critério fonético, na qual expõe que pode ocorrer a redução de unidades como a sílaba ou até mesmo a palavra na fala espontânea. Jessen (2007) segue o argumento exposto por Künzel, percorrendo questões de variação linguística que são relevantes ao considerar a transcrição fonética.

A representação de uma certa palavra no léxico mental pode variar entre os falantes de uma mesma comunidade linguística. Para alguns falantes a palavra "fósforo" possui 3 sílabas fonológicas, e espera-se encontrar na sua produção pronúncias em que as três ocorram, embora em algumas situações ela possa aparecer em forma reduzida, com duas sílabas, "fosfro". Para outros falantes, no entanto, é possível que em sua produção nunca apareça a forma com três sílabas, apenas a versão com duas sílabas, o que permitiria supor que em seu léxico mental a palavra esteja representada apenas com 2 sílabas. Se assumirmos a mesma forma fonológica para todos os falantes e usarmos o mesmo número de sílabas para calcular a taxa de elocução ou articulação, o resultado superestimaria o valor das taxas dos falantes que produzem a versão “reduzida” em todos os contextos. Künzel (1997 p. 50) opta pela contagem do número de sílabas efetivamente pronunciadas, justificando que o tema central de sua investigação é a articulação, não a intenção do falante. Assim como Künzel, Jessen (2007 p. 57) também opta por determinar a taxa de produção a partir das unidades fonéticas

efetivamente presentes no sinal de fala, uma vez que no contexto forense nem sempre o material de fala de que o perito dispõe permite traçar um perfil linguístico do falante que permitisse determinar com segurança qual seria a forma fonológica de cada palavra presente na amostra.

Neste estudo optamos por não privilegiar uma das formas de transcrição e estudar ambas. Assim, podemos comparar os resultados das duas formas e determinar se a adoção de um critério ou outro tem algum efeito para a determinação do tempo de estabilização da taxa de produção da fala.

Como era possível prever, observou-se que a quantidade de unidades fônica em um trecho de fala segmentado pode diferir de uma modalidade de transcrição para a outra. Na figura 1, pode-se ver um exemplo das diferenças no número de fones identificados em cada intervalo segmentado.

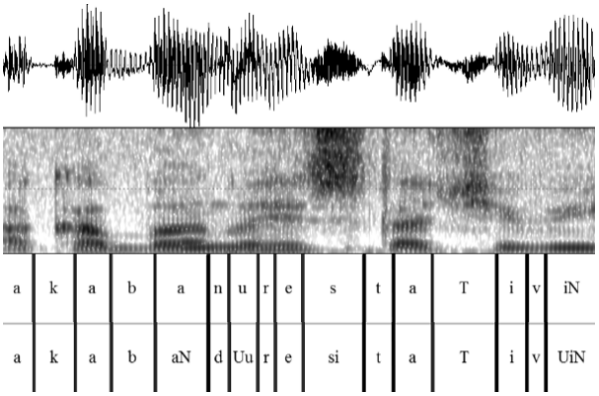

FIGURA 1 Forma de onda e espectrograma de trecho da passagem Lobato "[...] acabando o recitativo em [...]" segmentada em fones de amostra na taxa normal do falante em questão, na transcrição fonética (acima) e fonológica (abaixo).

Através da Figuras 1 vê-se que alguns segmentos como o quinto, sexto, sétimo, décimo e os seguintes contem símbolos a mais na transcrição fonológica para espelhar todos os componentes do trecho da passagem. Logo, com a diferença entre eles, haverá divergência na média de fones por segundo ao analisar os segmentos de cada transcrição. O propósito é comparar os resultados de ambas transcrições para saber se haveria um valor significativo de diferença entre ambas as transcrições. Essa decisão foi mantida em todos os níveis de segmentação conforme determinado no sexto parâmetro de segmentação.

No interior de cada trecho segmentado, a notação apresentada corresponde a símbolos da proposta de notação fonética Sampa-PB.

3.2 Procedimentos de extração, manipulação e análise dos dados

Ao fim da etapa de segmentação e etiquetação dos segmentos fônicos, cujos critérios foram discutidos em 3.1, são produzidos objetos TextGrid para cada áudio do corpus, que armazenam as informações sobre as taxas de fala lenta, normal e rápida em ambas as transcrições fonética e fonológica para processamento dos dados obtidos através da segmentação.

Em cada TextGrid há duas camadas de anotação, uma em que as pausas silenciosas são incorporadas à duração da unidade segmentada imediatamente anterior à pausa e outra em que a duração das pausas silenciosas não é incorporada à segmentação das unidades relevantes, como é possível observar na Figura 2. Os dados da primeira camada são usados para o cálculo da taxa de elocução e os dados da segunda para o cálculo da taxa de articulação.

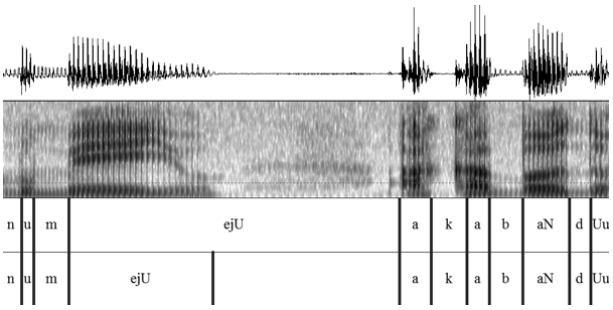

FIGURA 2 Forma de onda e espectrograma de trecho da passagem Lobato "[...] no meio, acabando o [...]" segmentada em fones de amostra na taxa normal do falante em questão, na transcrição fonológica. A primeira camada remete a taxa de elocução que inclui as pausas e a segunda, a taxa de articulação que não as inclui.

Uma situação em especial como mostra a Figura 3, são os casos em que ocorrem consoantes oclusivas desvozeadas que sucedem pausas silenciosas, na qual se adotou a delimitação da duração dessas consoantes em 80 milissegundos (limiar escolhido em função de 80 ms ser o valor de duração típico das plosivas desvozeadas no português brasileiro) [cf. Tabela de descritores estatísticos das durações dos fones do PB apresentados em Barbosa 2006 p. 489].

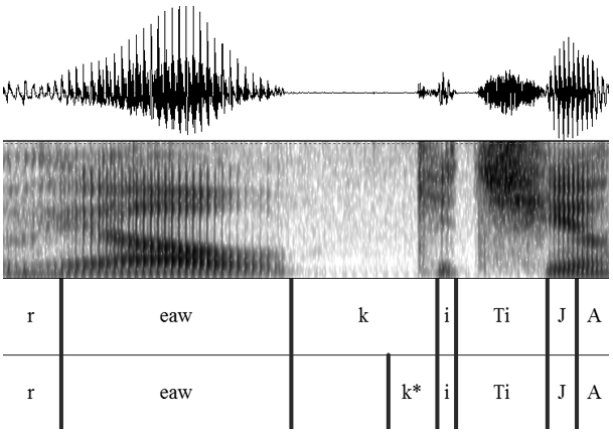

FIGURA 3 Forma de onda e espectrograma de trecho da passagem Lobato "[...] real que tinha [...]" segmentada em fones de amostra na taxa normal do falante em questão, na transcrição fonológica. As oclusivas posteriores às pausas silenciosas são marcadas com o símbolo * para indicar a duração em 80ms.

Scripts do Praat foram usados para a extração de duração das unidades segmentadas nas diferentes camadas de cada arquivo TextGrid. No total foram produzidos 288 arquivos de texto, resultado da combinação das variáveis FALANTES, NÍVEIS de taxa (lenta, normal e rápida), UNIDADES de agrupamento (fone, sílaba e unidade VV), TIPO de taxa (elocução e articulação) e CRITÉRIO de segmentação (fonética e fonológica): 8 FALANTES x 3 níveis de taxa x 3 unidades de agrupamento x 2 tipos de taxa x 2 tipos critérios de segmentação.

Cada um dos 288 arquivos é uma tabela que lista todas as unidades segmentadas e informações como a duração de cada uma, o valor cumulativo da taxa de produção (elocução ou articulação) e o número de segmentos cumulativos. Os valores

das taxas de produção cumulativos são calculados por meio da fórmula (1):

$$cSR_i = \frac{i}{\sum_{j=1}^{i} dur_j} \quad (1)$$

Essas tabelas foram então processadas por um *script* do ambiente de computação estatística R (2016) para a aplicação da técnica estatística *change point analysis*, que identifica na série temporal formada pelos valores da taxa cumulativa o ponto no tempo em que a variância sofre uma redução significativa do número de unidades compreendidos entre o início da amostra e o ponto de estabilização

Foram calculados também a estimativa de erro (diferença percentual entre o valor da taxa no ponto de estabilização e o valor global da taxa, isto é, aquele calculado com base em todas a duração da amostra de fala) determinada pela fórmula (2), em que os termos r_{st} se refere ao valor da taxa no momento em que o ponto de estabilização é definido e r_g refere-se a taxa global

$$e = \frac{r_{st} - r_g}{r_g} \cdot 100 \quad (2)$$

4. Resultados obtidos

Nesta seção, apresentaremos resultados preliminares das seguintes variáveis: tempo de estabilização, número de segmentos presentes entre o início da amostra e o ponto de estabilização e o erro médio. A etapa seguinte consistirá na análise estatística completa dos dados.

Para efeito de ilustração, as Figuras 4 e 5 mostram dois exemplos de série temporal formada pelos valores das taxas calculadas de forma cumulativa. Em ambos os gráficos, a linha vertical tracejada indica o ponto no tempo, contado a partir do início da amostra de fala, em que a técnica *change point analysis* encontrou o ponto de estabilização. Na Figura 4, o tempo de estabilização ocorre mais precocemente devido a baixa variabilidade no nível da sílaba. Na Figura 5, o tempo de estabilização da série temporal ocorre mais a frente considerando o tipo e o nível das taxas.

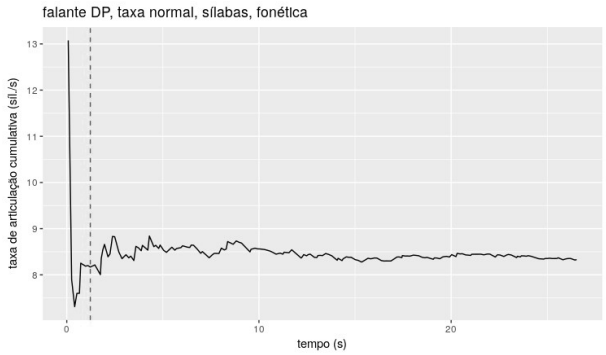

FIGURA 4 O traçado vertical indica o ponto de estabilização da amostra em questão.

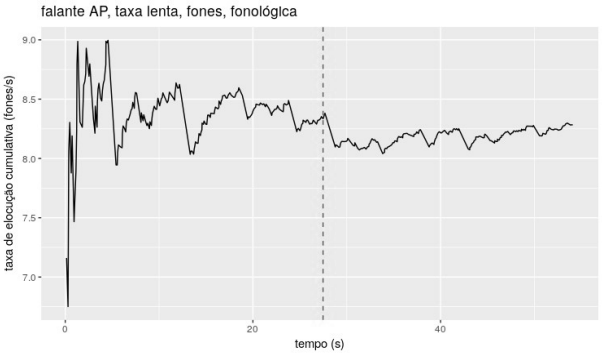

FIGURA 5 O traçado vertical indica o ponto de estabilização da amostra em questão.

As tabelas 1 e 2 apresentam os valores médios das variáveis: tempo de estabilização, número de segmentos presentes entre o início da amostra e o ponto de estabilização e o erro médio em função das unidades de segmentação, do tipo e do nível das taxas de acordo com o nível de transcrição (fonética ou fonológica):

Unidade	Tipo de taxa	Nível da taxa	Estabilização	Nº segmentos	Erro (%)
fone	articulação	rápido	9.92	145.62	0.64
		normal	12.50	144	-0.31
		lento	9.73	106.93	-1.64
	elocução	rápido	12.34	185.75	3.10
		normal	13.40	162.81	1.64
		lento	16.90	185.70	3.20
sílaba	articulação	rápido	8.50	67.31	3.94
		normal	9.83	62.93	2.02
		lento	11.35	68.25	4.05
	elocução	rápido	10.41	81.43	7,01
		normal	11.30	72.5	7.5
		lento	13.13	79.31	7.90
unidade VV	articulação	rápido	8.45	49.12	5.80
		normal	8.52	41.90	3.44
		lento	11.0	50.18	4.00
	elocução	rápido	7.06	42	9.6
		normal	8.10	40.40	6.71
		lento	10.70	49.5	8.24

TABELA 1 A transcrição fonética apresenta os respectivos dados sobre o ponto de estabilização em segundos, o número de segmentos total da amostra analisada e a taxa de erro estimada em porcentagem.

Unidade	Tipo de taxa	Nível da taxa	Estabilização	Nº segmentos	Erro (%)
fone	articulação	rápido	9.7	160	0.6
		normal	12.4	165	-0.7
		lento	10.6	136	-0.2
	elocução	rápido	13.2	204	3.1
		normal	12.8	159	1.3
		lento	15.5	176	3
sílaba	articulação	rápido	8.3	72	3.8
		normal	8.3	61	1.4
		lento	9.6	66	4
	elocução	rápido	8.8	70	7.4
		normal	10.5	67	8.3
		lento	12.6	76	8.2
unidade VV	articulação	rápido	7.7	49	5.7
		normal	7.5	42	3.4
		lento	10	52	3.9
	elocução	rápido	7.1	42	9.6
		normal	8.1	40	6.7
		lento	10.8	51	7.8

TABELA 2 A transcrição fonológica apresenta os respectivos dados sobre o ponto de estabilização em segundos, o número de segmentos total da amostra analisada e a taxa de erro estimada em porcentagem.

O tipo de transcrição (fonética ou fonológica) não parece ter um efeito importante sobre os tempos de estabilização. A diferença absoluta entre os tempos de estabilização para as 18 categorias mostradas nas Tabelas 1 e 2 varia entre 0 e 3,2 segundos, com média de 1,2 s (desvio-padrão de 1,2 s). Em termos do número de segmentos presentes no intervalo de estabilização o efeito mostra-se de maior magnitude. As diferenças variam entre 0 e 57 segmentos com média de 14 segmentos. A maior diferença se dá na taxa de articulação em fones/s, nível rápido. Essa diferença não é inesperada, uma vez que a taxa rápida é aquela em que há mais casos de omissão de fones no sinal acústico ou coarticulação extrema entre fones. A influência do tipo de transcrição sobre a taxa de erro não é de grande magnitude. A diferença absoluta entre as taxas varia de 0 a 2,8, com média de 0,5 (desvio-padrão de 0,7). As maiores diferenças nas três medidas acontecem na taxa de articulação em fones/s, nível rápido.

5. Referências

[1] ARANTES, P.; LIMA, V. G. Estimates of minimum sample length for speaking rate. Revista do GEL. (aceito).

[2] BARBOSA, A. P. Incursões em torno do ritmo da fala. Campinas: Pontes, 2006.

[3] BARBOSA, A. P.; MADUREIRA, S. Manual de Fonética Acústica Experimental: aplicações a dados do português. São Paulo: Cortez, 2015.

[4] BOERSMA, P. Praat, a system for doing phonetics by computer. Glot International, v. 5, n. 9/10, p. 341-345, 2001.

[5] JESSEN, M. Forensic reference data on articulation rate in German. Science and Justice, v.47, p. 50-67, 2007.

[6] KENDALL, T. Speech Rate, Pause, and Sociolinguistic Variation: Studies in Corpus Sociophonetics. Oregon: University of Oregon, 2013. p. 121-157

[7] KILLICK, R.; ECKLEY, I. A. changepoint: An R Package for Changepoint Analysis. Journal of Statistical Software, v. 58, n. 3, p. 1-19, 2014.

[8] KÜNZEL, H. Some general phonetic and forensic aspects of speaking tempo. Forensic Linguistics, v. 4, n. 1, p. 48-83, 1997.

[9] LEHISTE, I. Suprasegmentals**.** Cambridge, MA: MIT Press, 1970.

[10] R CORE TEAM. R: A Language and Environment for Statistical Computing. Vienna. 2016.

[11] TURK, A.; NAKAI, S. Acoustic segment durations in prosodic research: a practical guide. In: SUDHOFF, S. Methods in Empirical Prosody Research. Berlim: Walter de Gruyter, 2006. p. 1-28.

[12] *Communication Association, August 20–24, Stockholm, Sweden, Proceedings*, 2017, pp. 100–104