CIDI2019BH

9º CIDI
9th Information Design International Conference

9º CONGIC
9th Information Design Student Conference

Tipografia modulada pela fala: avaliação de um algoritmo de geração de prosódia visual em textos

Speech-modulated typography: evaluation of an algorithm for generating visual prosody in texts

Caluã de Lacerda Pataca & Paula D. Paro Costa

prosódia visual, tipografia, computação afetiva, análise de fala, design generativo

Ler um texto, mesmo que silenciosamente, depende não só de estruturas cognitivas de processamento de imagens mas, também, daquelas que tipicamente decodificam sons. A partir dessa constatação, estudos recentes têm explorado a representação gráfica na tipografia de elementos da prosódia, gerando novas ferramentas didáticas, tecnologias assistivas ou mesmo possibilidades expressivas inovadoras. Neste artigo descrevemos um modelo computacional capaz de extrair elementos acústicos do áudio de uma fala e representá-los graficamente na tipografia do texto. O artigo também apresenta os resultados de uma avaliação desse modelo em um exercício de card-sorting com 34 leitores onde medimos quão consistentes entre os participantes foram as interpretações dessas representações gráficas de prosódia, especialmente em relação à dedução de emoções da expressão vocal por meio de sua representação indireta na tipografia. Encontramos indícios estatisticamente significantes de que houve coerência nessas interpretações, mas também que os parâmetros (i.e., associação das features acústicas de amplitude, frequência fundamental e duração de sílaba com os atributos tipográficos de peso, inclinação e largura horizontal, respectivamente), na forma como usados no modelo, têm desempenhos desiguais quando considerados diferentes classes de emoção na voz. Sugerimos que estudos futuros investiguem essa assimetria, explorando maneiras de reduzi-la.

visual prosody, typography, affective computing, speech analysis, generative design

Reading a piece of text, even when done silently, depends not only on the cognitive structures typically involved in the processing of images but also on those that would typically be involved in the decoding of sounds. From this finding, recent studies have explored ways of representing certain elements of prosody graphically in the typographic form, creating innovative teaching tools, assistive technologies or even new possibilities of expression. In this paper, we describe a computational model capable of extracting acoustic features from a recording of a spoken word performance and representing them graphically in its text's typography. The article also presents the results of an evaluation of this same model in a card-sorting exercise with 34 readers in which we measured how consistent the participants interpreted these graphic representations of prosody, especially regarding their use for inferring emotions present in the recorded voice through its indirect representation in typography. We found statistically significant evidence that there was consistency in these interpretations, but also that the parameters (i.e., association of acoustic features of amplitude, fundamental frequency and syllable duration with the typographic attributes of weight, inclination and horizontal width, respectively), as used in the model, have differing performances when considering different classes of emotion in the voice. We suggest that future studies investigate these differences, exploring ways to reduce them.

1 Introdução

Ler é uma habilidade cognitiva de alto nível, que exige longo período de treinamento e que envolve intenso processamento neurológico. Dentre as estruturas cerebrais envolvidas no processo da leitura, surpreende notar que, além daquelas que se encarregam do processamento de imagens, a leitura exige também o emprego das estruturas tipicamente

Anais do 9º CIDI e 9º CONGIC
Luciane Maria Fadel, Carla Spinillo, Anderson Horta, Cristina Portugal (orgs.)
Sociedade Brasileira de Design da Informação – SBDI
Belo Horizonte | Brasil | 2019
ISBN 978-85-212-1728-2

Proceedings of the 9th CIDI and 9th CONGIC
Luciane Maria Fadel, Carla Spinillo, Anderson Horta, Cristina Portugal (orgs.)
Sociedade Brasileira de Design da Informação – SBDI
Belo Horizonte | Brazil | 2019
ISBN 978-85-212-1728-2

relacionados ao processamento de sons (Seidenberg, 2017, cap.7). Isso ocorre porque, mesmo quando lê silenciosamente, cabe ao leitor interpretar em sua voz interna a prosódia do texto, habilidade fundamentalmente relacionada à boa compreensão do mesmo.

Investigar como essa voz emerge a partir do texto não é então uma questão meramente *estética*. Ao contrário da noção vendida por certos cursos de leitura dinâmica de que uma leitura sem subvocalização traria ganhos de velocidade sem perdas na compreensão, o leitor experiente se vale dessa voz interna para, justamente, ler bem um texto — a informação prosódica ajuda a reduzir ambiguidades e, assim, facilitar a compreensão (Seidenberg, 2017, cap.4).

Uma situação que exemplifica a importância da expressividade dessa voz interna se dá em crianças que, na alfabetização, leem de maneira monótona, ou seja, não conseguem extrair do texto as variações naturais da fala. Como discute Bessemans (2017), essas crianças tendem a desenvolver problemas de compreensão que as acompanham em suas vidas enquanto leitoras.

Dada essa situação, Bessemans e seu grupo criaram um tipo de intervenção visual na forma do texto que tem apresentado resultados positivos com essas crianças. O trabalho consiste na produção e avaliação da leitura de textos nos quais certos elementos da prosódia (i.e. intensidade, duração e tom) estão codificados graficamente no desenho e disposição das letras. Assim, uma palavra que deve ser lida com maior intensidade que outras pode estar em negrito; se a sugestão for que ela deve ser lida mais rapidamente, usa-se uma fonte condensada horizontalmente; se a voz que a lê deve ser mais aguda, a palavra poderá ser posicionada acima da linha de base. Essas “dicas” visuais ajudariam os jovens leitores a sair da leitura monótona. Os resultados iniciais têm se mostrado promissores (Bessemans, 2017).

Com uma abordagem semelhante situada na intersecção do design com a ciência da computação, Wolfel, Schlippe e Stitz (2015) descrevem um software no qual uma fonte tipográfica construída a partir de modelos matemáticos é modificada a partir de propriedades acústicas da voz. Em uma avaliação com leitores, foram encontrados indícios tanto de que as características da fala conseguiram ser impressas no texto quanto de que uma abordagem nessa linha poderia ser usada para representar emoções presentes na voz.

Segundo discutem Wolfel, Schlippe e Stitz (2015), esse tipo de intervenção poderia vir a apoiar aplicações diversas e importantes: ferramentas para alfabetização; auxílio no aprendizado de línguas estrangeiras; auxílio no tratamento de patologias de fala (e.g. desambiguar as sílabas com ênfase); dicas visuais que ajudem disléxicos a decifrar em sons a linguagem escrita; legendas para filmes nas quais a interpretação dramática que os atores dão a suas vozes esteja representada nas letras; entre outras.

Nesse contexto, nossa pesquisa busca expandir os modelos já existentes de representação da prosódia na tipografia, em particular por meio de sistemas automatizados.

A seção *Metodologia* apresenta nossa proposta de um modelo de extração computacional de características acústicas da voz humana e seu mapeamento em atributos visuais tipográficos, visando imprimir no desenho das letras qualidades perceptivas da fala. Discutimos aspectos da implementação desse algoritmo para, em seguida, apresentar uma avaliação na qual testamos, com um grupo de leitores universitários, efeitos que nosso modelo causou na interpretação de um conjunto de frases geradas a partir da leitura dramática das mesmas por uma atriz profissional.

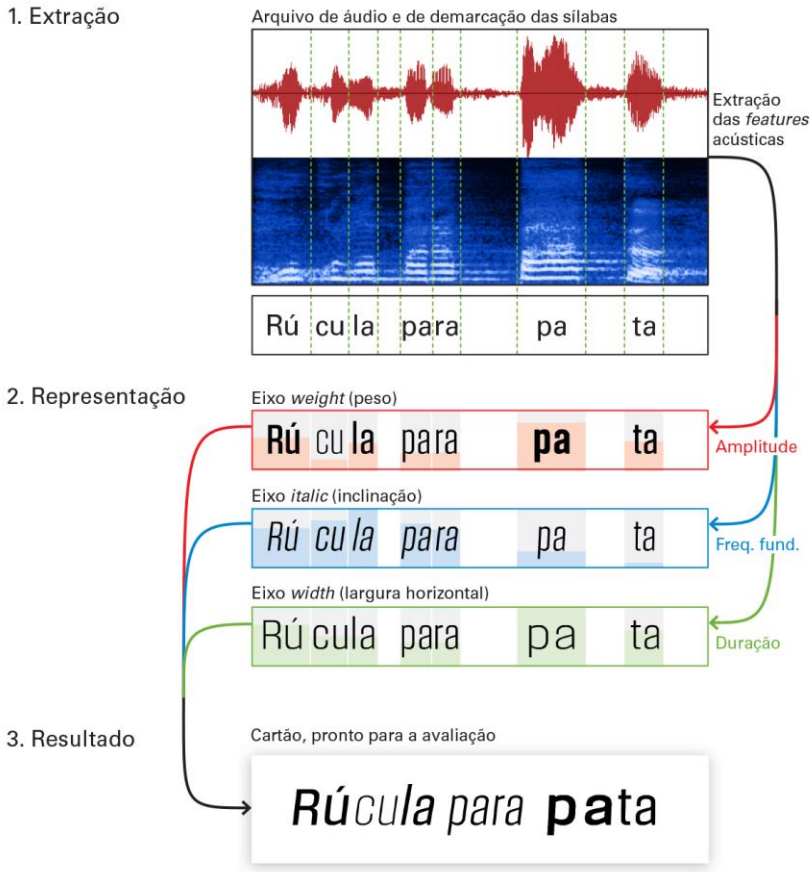
Em *Discussão*, argumentamos que há nos resultados indícios de que por meio da tipografia modificada os participantes conseguiram apreender no texto não só seu conteúdo explícito mas também a emoção presente na voz da atriz.

2 Metodologia

Sobre o algoritmo e modelo subjacente de mapeamento fala-tipografia

A **Figura 1** apresenta as duas etapas principais do algoritmo proposto para a implementação do modelo de mapeamento fala-tipografia.

Figura 1: As duas principais etapas do software e, ao final, a tipografia resultante. Na primeira, o arquivo com as sílabas é usado para recortar o arquivo de áudio, de onde então são extraídas e normalizadas as *features* acústicas de amplitude, frequência fundamental e duração. Na segunda etapa, esses valores são então codificados nos eixos tipográficos de uma *variable font*, recompondo as sílabas (didaticamente demonstramos a influência dos três eixos na tipografia de maneira isolada, mas no software eles são modulados concomitantemente) e, ao final, gerando os cartões usados na avaliação.



A primeira etapa realiza a extração e o processamento das *features* acústicas da voz. O algoritmo recebe como entrada dois arquivos:

- um arquivo de áudio com a fala cuja expressão vocal será analisada;
- um arquivo de transcrição silábica temporizada, que consiste num arquivo texto contendo as demarcações temporais que, no arquivo de áudio, correspondem a cada sílaba.

A partir das informações de ambos os arquivos, o algoritmo extrai a seguintes informações acústicas da fala:

- Amplitude média, calculada pela *Root Mean Square* (RMS) do trecho;
- Frequência fundamental, calculada usando o método *swipe*, disponibilizado na biblioteca *pysptk*; e
- Duração da sílaba.

Para a implementação do algoritmo, utilizou-se a linguagem de programação Python e as bibliotecas de processamento de áudio *librosa* e *pysptk*.

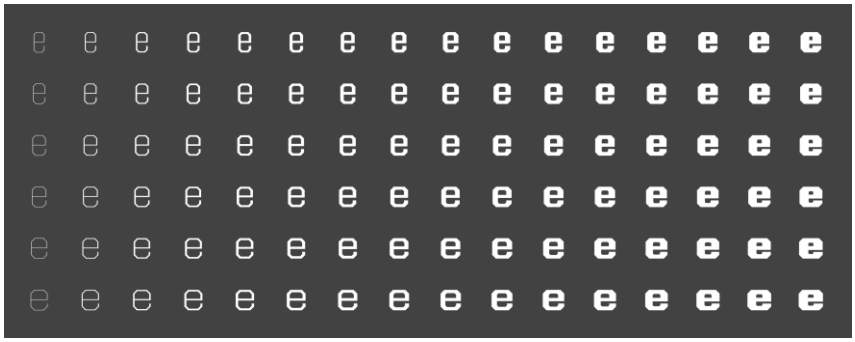
Optamos por aspectos acústicos relacionados à prosódia porque, além dos resultados já citados de Bessemans (2017) e de literatura da computação afetiva apontando para um bom desempenho dessas características na detecção computacional de emoções (Koolagudi; Rao, 2012), essas *features* cumprem dois requisitos importantes em nosso caso de uso:

- São unidimensionais, ou seja, cada *feature* pode ser mapeada diretamente para uma única dimensão visual da tipografia;
- Trazem informações significativas sobre a emoção expressa na fala mesmo quando consideradas em nível **local** (da sílaba) em oposição ao nível **global** (da frase), como demonstram Rao et al. (2010). Assim, a forma com que cada *feature* varia no tempo de uma frase poderá ser mapeada visualmente à tipografia de cada **sílaba**.

Essas *features* são então normalizadas linearmente considerando as medidas máximas e mínimas em cada frase levando em conta o conjunto de áudios que serão processados (o áudio é processado frase por frase, mas para o experimento compilamos um conjunto com 24 arquivos). Esse passo é importante porque, ao processar mais de uma frase ao mesmo tempo, queremos que os valores extremos nos parâmetros tipográficos sejam análogos aos valores extremos nos áudios tomados como um todo. A ideia é que haja consistência de representação visual entre todas as frases.

Esses valores seguem como entrada em um segundo script, no qual cada *feature* normalizada é mapeada para um eixo visual de uma *variable font*. Esta é uma fonte tipográfica onde certos atributos visuais são definidos não como desenhos distintos e completos (e.g. os arquivos separados em uma fonte tradicional para definir negrito, itálico etc) e sim como variações contínuas geradas automaticamente a partir da interpolação do desenho da fonte em suas configurações extremas (Microsoft, 2018). Na **Figura 2** mostramos um exemplo no qual a letra “e” de uma *variable font* tem seu eixo de peso modulado no eixo horizontal ao mesmo tempo em que tem seu eixo de largura horizontal modulado no eixo vertical.

Figura 2: Uma mesma letra de um único arquivo em uma *variable font*.
Na horizontal, modulamos o eixo *weight* (peso); na vertical, o *width* (largura).



Como na **Figura 2**, diferentes eixos podem ser combinados concomitantemente, o que permite que uma *variable font* possa representar graficamente as inúmeras combinações de valor, mesmo quando sutis, presentes nas *features* acústicas da voz. Para ter acesso a funções que permitem modular eixos *variable fonts*, esse segundo script, ainda que também escrito em Python, foi rodado no ambiente *DrawBot*[1].

[1] Software usado na comunidade de *design digital generativo* e que, além de implementar tecnologias tipográficas de ponta, possui uma série de facilidades para gerar arquivos para impressão, como os

Para nossa avaliação usamos o seguinte mapeamento de *features* acústicas para eixos tipográficos:

- *Amplitude média* → *Weight* (peso);
- *Frequência fundamental* → *Slant* (inclinação);
- *Duração* → *Width* (largura horizontal).

Nossa escolha para os dois primeiros mapeamentos deriva do modelo de Wolfel, Schlippe e Stitz (2015) e do de Bessemans (2017), mas optamos por *slant* (inclinação da letra) para representar a *frequência fundamental* por supor que mudanças na linha de base poderiam dificultar a aplicabilidade desse modelo em textos longos.

Avaliação

O propósito da avaliação foi investigar que influências nosso modelo de mapeamento fala-tipografia poderia ter na interpretação de um texto.

Para isolar os efeitos da forma tipográfica, buscamos usar frases cujo conteúdo textual fosse pouco expressivo, ou seja, onde a própria frase não sugerisse nenhuma das emoções. Inversamente, as frases foram lidas com a voz carregada de emoção, ou seja, com uma dramaticidade vocal que maximizasse as variações gráficas no desenho tipográfico.

Trabalhamos a partir das frases da base construída e descrita por Costa (2015). Selecionamos quatro frases de sentido obscuro e/ou ambíguo: Filha, rúcula para a pata!; Passarinho, cuidado com a asa!; Você tem certeza disso?; Lilo, Kika, Luku, puxem o cavalo!.

Os áudios nesta base incluem estas quatro frases lidas diversas vezes por uma mesma atriz. Em cada passada, sua voz se modulava de modo a representar, e com ênfase, as seis emoções básicas (i.e. raiva, nojo, medo, felicidade, tristeza e surpresa) como descritas em Ekman (1970).

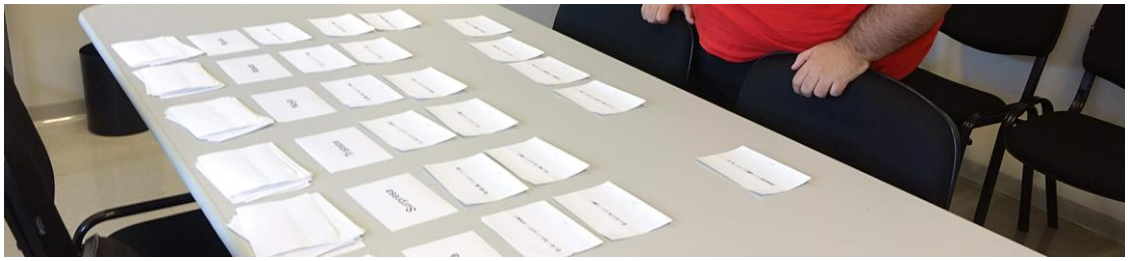
Os seis arquivos de áudio para cada uma das quatro frases foram então processados e os textos resultantes, com a tipografia modificada pela fala da atriz, impressos em cartões de papel tamanho A5. Na **Figura 3**, apresentamos os 24 cartões agrupados por frase e, em cada uma, ordenados por emoção como a seguir: raiva, nojo, medo, felicidade, tristeza e surpresa. Na **Figura 4**, uma fotografia tirada em uma das sessões.

---

cartões A5 usados em nossa avaliação.

Figura 3: As 24 frases como apresentadas aos participantes. (A proporção dos cartões foi aqui adaptada para otimizar uso do espaço na página.)

Figura 4: Um participante organiza os cartões. Na segunda fileira na mesa, os rótulos, e abaixo cada fileira subsequente contém os cartões de cada uma das frases.



Para a avaliação realizamos sessões de *card-sorting*. Foram sessões individuais e divididas, cada uma, em 4 rodadas, uma para cada uma das frases escolhidas.

Optamos pelo *card-sorting* por três principais motivos:

- como discutido em Santos (2006), é um método relativamente rápido e barato para se investigar constructos latentes em avaliações com participantes (em nosso caso, os constructos sendo as próprias classes de emoções);
- sendo um teste presencial, nos permitiu observar os participantes e realizar breves entrevistas semi-estruturadas ao final de cada sessão, complementando os dados quantitativos coletados;
- permite usar duas ferramentas de organização e interpretação de dados (*matrizes de confusão* e *edit-distances*), úteis porque, para além de uma mera taxa de "acerto" para as classificações, ajudam a quantificar a presença (ou não) de coerência na maneira como os participantes interpretaram cada cartão e emoção.

Sobre as sessões de avaliação com os participantes

Antes do início de cada sessão do texto os participantes foram informados de que haveria nos cartões uma correspondência entre a forma tipográfica e certas características da voz de uma atriz que lera previamente os textos (não informamos maiores detalhes). Para cada emoção havia um envelope rotulado e os participantes foram instruídos a depositar cada cartão na "emoção" correspondente a seu desenho tipográfico.

Ao final de cada sessão realizamos breves entrevistas semi-estruturadas. Nossa intenção foi a de buscar levantar possíveis estratégias usadas na organização dos cartões, além de investigar possíveis modelos mentais formulados pelos participantes para explicar o funcionamento do mapeamento fala-tipografia.

Coleta e análise de dados

Para processar as *edit-distances* de cada rodada foi criado um script Python. Para avaliar a independência da distribuição dos cartões obtida em relação a uma distribuição aleatória (hipótese nula) foi usado o teste qui-quadrado, com nível de significância fixado em $\alpha = 0{,}05$.

3 Resultados

Card-sort

O teste de *card-sorting* foi aplicado em novembro de 2018 em três sessões. Contou com 34 participantes, todos estudantes universitários — 29 alunos de cursos de graduação em design e 5 alunos de pós-graduação em engenharia.

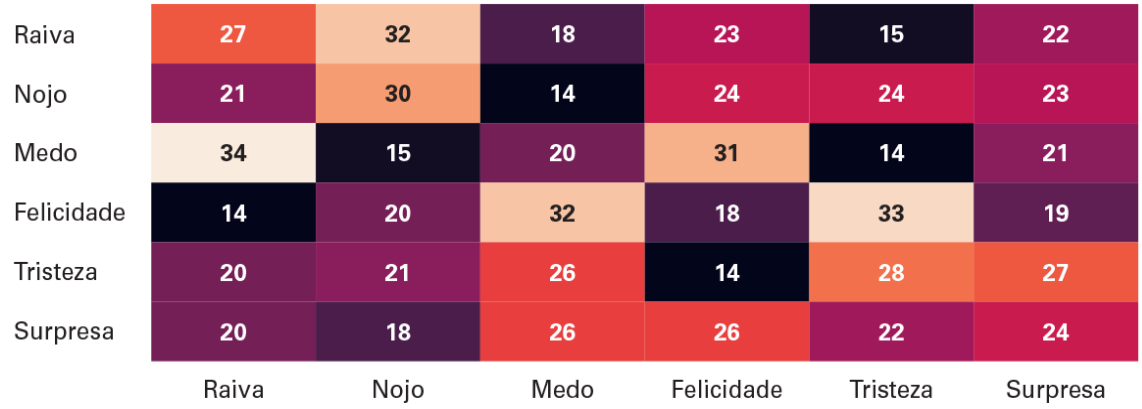
Obtivemos assim 816 pares de dados (34 participantes × 4 frases × 6 emoções), compreendendo cada um a emoção original do cartão, conforme encenada pela atriz que gravou as frases, pareada com a emoção escolhida pelo participante.

A **Tabela 1** apresenta a porcentagem de cartões atribuídos pelos participantes a suas emoções originais, ou seja, aqueles associados à mesma emoção que a atriz buscou interpretar. Por ora, consideraremos esse valor o "acerto" médio de cada emoção em cada frase. A tabela inclui também o acerto médio por frase e por emoção.

Tabela 1: Eficiência na organização dos cartões em cada uma das emoções consideradas. .29 de eficiência na primeira frase quer dizer, por exemplo, que 29% dos cartões classificados como raiva foram gerados a partir do áudio encenando raiva.

Frase	Raiva	Nojo	Medo	Felicidade	Tristeza	Surpresa	média por frase
Filha, rúcula para a pata	.29	.32	.06	.06	.12	.21	.18
Passarinho, cuidado com a asa	.20	.24	.18	.24	.41	.15	.24
Lilo, Kika, Luku, puxem o cavalo	.06	.12	.15	.06	.15	.18	.12
Você tem certeza disso?	.24	.21	.21	.18	.15	.18	.20
média por emoção	.20	.22	.15	.14	.21	.18	

Figura 5: Matriz de confusão compilando as 4 frases



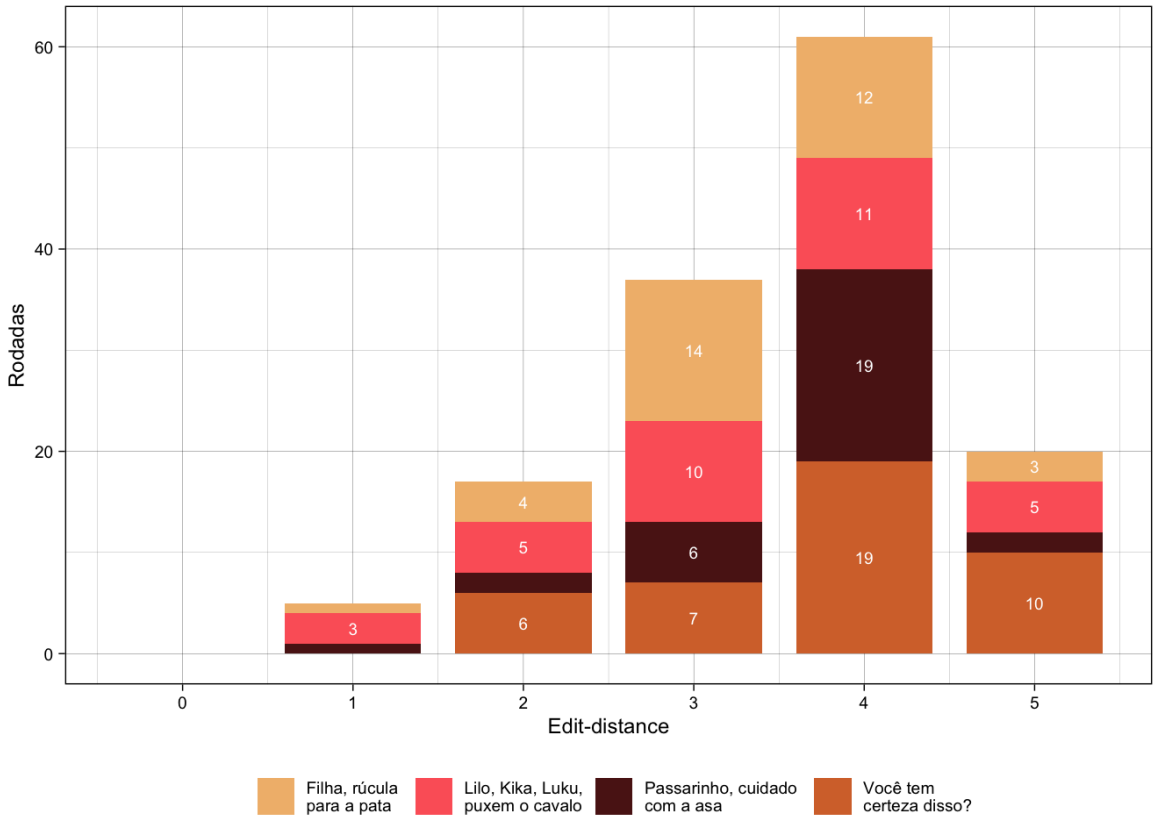
A **Figura 5** contém a *matriz de confusão* da organização dos cartões. Em tal disposição, cada linha representa uma emoção (a), como interpretada pela atriz, e cada coluna representa a classificação dada pelos participantes para esses mesmos cartões (b). O cruzamento em cada célula nos dá, assim, a quantidade absoluta de cartões da emoção (a) classificados como se pertencentes à emoção (b).

Com isso, pode-se distinguir aqueles erros e acertos que sugerem que a tipografia modificada não teve efeito — indicados por uma alta uniformidade na distribuição de cada cartão entre todas as categorias — daqueles que sugerem que houve um efeito consistente. Esse ponto é importante frisar, pois com o *card-sort* almejamos medir se haveria ou não **consistência** nas interpretação e não se os parâmetros que adotamos para o algoritmo conseguiram representar na tipografia exatamente as seis emoções encenadas pela atriz e capturadas nos áudios.

Isso se dá porque, ainda que tenha sido construído à partir de trabalhos semelhantes de outros pesquisadores, o modelo de mapeamento fala-tipografia como apresentado neste artigo não havia, até então, sido levado a campo. “Acertar” os parâmetros demandaria uma configuração experimental de foco muito mais estreito, aplicada a uma população maior e em múltiplas iterações. Não sabendo de antemão se a abordagem seria ou não válida, optamos por buscar confirmação (ou rejeição) do modelo tomado de maneira ampla.

Por fim, a **Figura 6** apresenta a distribuição de *edit-distances* para cada frase:

Figura 6: Distribuição das *edit-distances* para as 4 frases. No eixo X temos a quantidade de operações necessárias para transformar uma dada organização dos cartões na organização "idealizada". No eixo Y, quantas rodadas para cada uma das *edit-distances* possíveis, codificadas por cor por frase.



Como discute Nawaz (2012), *edit-distance* é um índice de divergência entre como estão organizados diferentes *card sorts*. Aqui, nos serve para indicar quão distante estão cada uma das organizações de cartões como feitas pelos participantes do que seriam organizações "perfeitas", ou seja, aquelas nas quais as emoções na voz da atriz estariam perfeitamente alinhadas com aquelas atribuídas pelos participantes.

Esse valor é calculado somando a quantidade de transformações necessárias para trocar as posições desses cartões e transformar uma organização em outra, como indicado na **Tabela 2**:

Tabela 2: Como é calculada a *edit-distance*? Imaginando que a organização "idealizada" aqui fosse "A · B · C", apresentamos em três exemplos como são calculadas suas respectivas *edit-distances*. Em verde, a rodada de trocas em que se chegou à organização "idealizada".

	valor inicial	**1ª troca**	**2ª troca**	***edit-distance***
exemplo 1	A · B · C			0 trocas
exemplo 2	A · C · B	A · **B** ↔ **C**		1 troca
exemplo 3	B · C · A	B · **A** ↔ **C**	**A** ↔ **B** · C	2 trocas

A **Figura 6** nos indica que nenhuma organização foi perfeita e que a maior parte delas demanda entre 3 e 4 trocas de posição. Na comparação entre frases, pode-se notar pior performance em "Passarinho, cuidado…" e "Você tem certeza…" quando comparadas às as frases "Filha, rúcula…" e "Lilo, Kika, Luku…".

Entrevistas

Ao final de cada sessão fizemos breves entrevistas individuais com os participantes.

Um comentário muito frequente foi o de que a avaliação era muito difícil — muitos participantes inclusive o fizeram espontaneamente enquanto tentavam organizar os cartões.

Quando perguntados sobre quais modulações visuais eram percebidas nos cartões, o atributo *peso* foi o mais citado. Muitos participantes o citaram e intuíram que mais peso equivaleria a maior volume na voz. Poucos, no entanto, citaram a relação oposta de que a letra, quando usada em pesos mais leves, corresponderia a uma voz de menor volume.

Para o segundo atributo mais citado — a *inclinação* na letra — houve maior dispersão nas interpretações. Para um participante ela foi entendida como velocidade na fala, enquanto outro a associou à fraqueza na voz da atriz. Já um terceiro relacionou a letra italicizada a uma voz triste. Houve ainda alguns que chegaram a citar a inclinação como atributo percebido, mas sem saber postular qual seria seu sentido.

Por fim, apenas um participante mencionou as variações de largura das letras no eixo horizontal. Curiosamente, conseguiu deduzir o atributo acústico como o mapeamos em nosso modelo, ou seja, a duração das sílabas.

Ao serem questionados sobre quais estratégias tinham utilizado para classificar cada cartão dentro das seis emoções, emergiram dois principais grupos: no primeiro, mais comum, o participante olhava para o cartão e buscava “soá-lo” mentalmente ou, menos frequentemente, em voz alta, tentando interpretar em sons as modulações visuais nas letras. A partir desse som imaginado (ou soado), tentavam deduzir a emoção correspondente para, então, classificar o cartão.

O segundo grupo ia pelo caminho oposto: tentava fazer soar a frase como que sob o efeito de cada uma das seis emoções para, só então, olhar para os cartões e procurar aquele cuja tipografia mais se aproximasse desse som.

Houve relatos de que algumas emoções pareciam mais simples de se classificar que outras, mas nesse ponto as respostas sobre quais seriam as emoções fáceis e quais as difíceis foram muito dispersas.

Alguns participantes mencionaram que as frases não facilitavam a interpretação — como era a intenção —, com exceção de um que nos relatou que a frase “Você tem certeza disso?” não parecia se encaixar igualmente bem em todas as emoções.

4 Discussão

Considerando apenas os resultados como apresentados na **Tabela 1** não seria possível descartar a hipótese nula[2] e afirmar que o desempenho na ordenação dos cartões é significantemente maior do que se esperaria de uma distribuição aleatória.

Voltando à matriz de confusão apresentada na **Figura 5**, é possível perceber que entre os participantes houve um alto grau de confusão entre a representação de cartões de *tristeza* e de *felicidade*. Uma interpretação possível para essa confusão é que o algoritmo teria sim conseguido representar na tipografia características análogas à emoção presente na voz da atriz, mas com alguns pressupostos equivocados. Nesse sentido, os resultados indicariam que as soluções gráficas para *tristeza* seriam melhor adequadas para *felicidade* e vice-versa.

[2] $\chi 2$ (5, N=816) = 6,97, $p > 0,05$

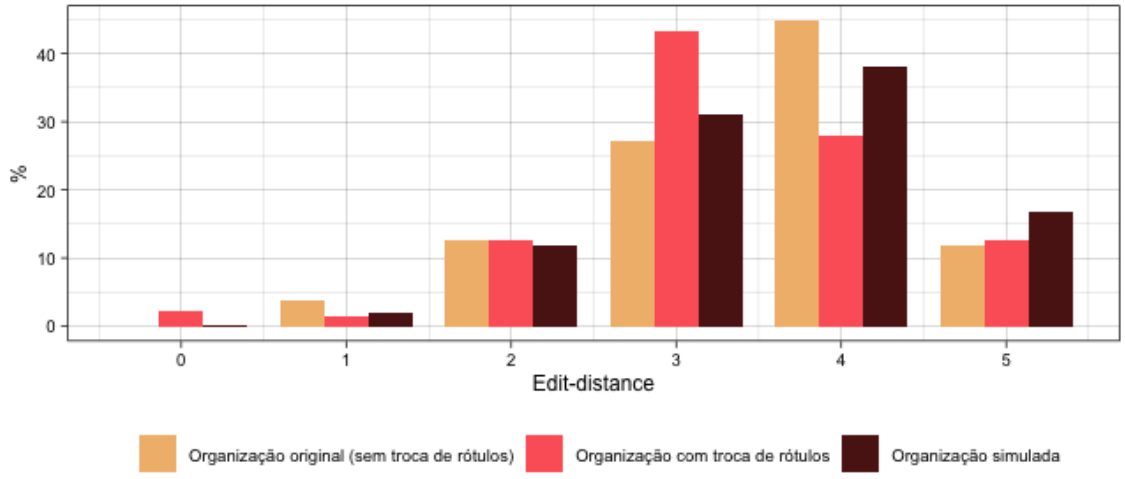
Se invertermos as duas emoções teremos uma configuração como aquela apresentada na **Tabela 3**, na qual é possível descartar a hipótese nula[3] e aceitar, pois, a hipótese alternativa de que os participantes estariam conseguindo deduzir conteúdo emocional presente na voz a partir das modulações tipográficas.

Tabela 3: Eficiência média na organização dos cartões quando se faz a troca Felicidade → Tristeza

	Raiva	**Nojo**	**~~Medo~~ Felicidade**	**~~Felicidade~~ Medo**	**Tristeza**	**Surpresa**	**média por frase**
Ordenação aleatória	.17	.17	.17	.17	.17	.17	.17
Ordenação medida na avaliação	.20	.22	.23	.24	.21	.18	.21

Conclusão semelhante emerge na análise das *edit-distances*. Para poder colocar em relevo os resultados obtidos, fizemos uma simulação que repetiu computacionalmente a avaliação, supondo a situação em que o modelo não tem efeito algum e que os participantes, portanto, organizam aleatoriamente os cartões. Repetimos a simulação 10.000 vezes[4] e somamos os resultados:

Figura 7: Comparação das distribuições de *edit-distances* entre a organização obtida na avaliação, a organização com a troca de rótulos de felicidade e medo e a organização simulada.



O desempenho indicado pela *edit-distance* da organização dos participantes é semelhante àquele medido na organização simulada, como já visto na **Tabela 2**. Mais interessante é notar que a melhora indicada na **Tabela 3**, quando da mudança dos rótulos de *felicidade* e *medo*, também aparece aqui quando essa operação de troca é repetida. Se considerarmos as médias dessas *edit-distances* temos os dados da **Tabela 4**, que evidenciam esta melhora:

[3] $\chi 2$ (5, N=816) = 13,73, $p < 0,05$

[4] Chegamos a esse limite experimentalmente: aumentamos o número de rodadas até chegar a um valor em que não houvesse oscilações significantes na distribuição dos resultados.

Tabela 4: *Edit-distances* para as três situações discutidas e suas médias.

	Edit-distances						
Situação	**0**	**1**	**2**	**3**	**4**	**5**	***Edit-distance* médio**
Organização original (sem troca de rótulos)	0	5	17	37	61	16	3,49
Organização com troca de rótulos	3	2	17	59	38	17	3,31
Organização simulada	0	3	16	42	52	23	3,55

Aceitando assim a confirmação de que a interpretação dos cartões foi influenciada de maneira consistente pelo mapeamento fala-tipografia, cabe mensurar esse efeito. E este foi baixo: apesar de termos detectado um aumento de 26% no desempenho médio versus a ordenação aleatória, como indicado na **Tabela 3**, cerca de 79% dos cartões foram ainda assim colocados em categorias “erradas”. Essa taxa encontra eco na declaração de muitos dos participantes que disseram que a avaliação era *muito difícil* — ou seja, os efeitos do modelo são ainda muito sutis ou ambíguos.

Por termos misturado variações em diversas parâmetros ao mesmo tempo (as três *features* combinadas com os três eixos tipográficos), os resultados quantitativos nos fornecem pistas muito difusas sobre por onde iniciar essas melhorias. Já nas entrevistas, ao contrário, há alguns caminhos mais nítidos: o peso parece ser um eixo visual bastante pregnante e sua associação à amplitude na voz é bastante plausível. Inversamente, as múltiplas interpretações nas variações na inclinação da letra e na largura horizontal indicou possíveis problemas na maneira como esses dois eixos foram utilizados e/ou como estiveram associados aos atributos acústicos de frequência fundamental e duração de sílaba, respectivamente.

Em relação ao público considerado, há indícios conflitantes. Por um lado, como os participantes eram todos estudantes universitários, pode-se supor dúvida sobre a aplicabilidade dessa abordagem para públicos mais amplos, em especial aqueles tipicamente menos habituados à linguagem escrita. Por outro lado, o fato de que os participantes não receberam treinamento prévio ou instruções detalhadas sobre o funcionamento do algoritmo pode indicar que este tem potencial de produzir resultados suficientemente intuitivos de modo a ter uso em contextos mais gerais.

5 Conclusão

Apresentamos evidências neste artigo que o mapeamento fala-tipografia, como proposto por nosso modelo, é plausível e que, mesmo em seu estágio inicial, consegue traduzir de maneira consistente parte da expressividade da voz humana.

Por outro lado, estudos futuros devem desmembrar e investigar em profundidade a efetividade de cada um dos parâmetros de extração de *features* acústicas e de sua representação enquanto eixos tipográficos em *variable fonts*. Para tanto, deve-se avaliar como cada eixo tipográfico é percebido quando associado a cada *feature*, investigando inclusive se essa percepção é constante independente da emoção na voz da atriz.

Restam dúvidas sobre como diferentes públicos reagirão a essa tipografia modificada. Conforme o modelo for iterado deve-se explorar como reagem a ele diferentes públicos, especialmente conforme forem exploradas aplicações práticas

Agradecimento

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nivel Superior - Brasil (CAPES) - Código de Financiamento 001.

Referências

Bessemans, A. (2017). Expressive typography to improve communication. Em *ATypI Montreal, 2017*. Disponível em: <http://youtu.be/JfsixaAmNOw>.

Costa, P. D. P. (2015). Two-Dimensional Expressive Speech Animation. *Tese (Doutorado em Engenharia Elétrica e de Computação)* — Faculdade de Engenharia Elétrica e de Computação, Universidade Estadual de Campinas. Campinas.

Ekman, P. (1970). Universal facial expressions of emotion. Em *California Mental Health Research Digest*, 8 (4),151-158.

Koolagudi, S. G.; Rao, K. S. (2012). Emotion recognition from speech: a review. Em *International Journal of Speech Technology*, v. 15, n. 2, 6.

Microsoft. (2018). OpenType Font Variations Overview. [S. l.], 15 ago. 2018. Disponível em: https://docs.microsoft.com/en-us/typography/opentype/spec/otvaroverview. Acesso em: 29 jun. 2019.

Nawaz, A. (2012). A comparison of card-sorting analysis methods. Em *The 10th Asia Pacific conference on computer human interaction (APCHI2012).* [s.n.]. Disponível em: <http://openarchive.cbs.dk/handle/10398/8587>.

Rao, K. S. et al. (2010). Characterization of emotions using the dynamics of prosodic features. Em *International conference on speech prosody*. Chicago, EUA.

Santos, G. (2006). Card sort technique as a qualitative substitute for quantitative exploratory factor analysis. *Corporate Communications: An International Journal*, Vol. 11 Iss 3 pp. 288 - 302, 2006.

Seidenberg, M. (2017). Language at the Speed of Sight: How We Read, Why So Many Can't, and What Can Be Done About It. 1st. ed. Nova Iorque: Basic Books. Versão Kindle.

Wolfel, M.; Schlippe, T.; Stitz, A. (2015). Voice driven type design. Em *2015 International Conference on Speech Technology and Human-Computer Dialogue (SpeD).* [S.l.: s.n.].

Sobre os autores

Caluã de Lacerda Pataca, Unicamp, Brasil <calua.pataca@gmail.com>

Paula Dornhofer Paro Costa, PhD, Unicamp, Brasil <paulad@unicamp.