# *Letras que ecoam a fala: tipografia modulada por prosódia*

*Caluã de Lacerda Pataca, Paula Dornhofer Paro Costa*

## *Objetivos e justificativas do projeto de pesquisa*

De um modo geral, ao representar visualmente um texto a tipografia se ocupa principalmente do que ele diz e não do *como* dizê-lo. Ainda que respeite pausas e mudanças de entoação que diferenciem asserções de interrogações, ao ler em voz alta o leitor tem grande margem para fazer soar as palavras conforme sua própria interpretação do texto, das personagens, das intenções do autor, etc. Essa abertura a diferentes interpretações abre um campo riquíssimo de expressão, pois, dado que o *que se diz* é modificado por *como se diz*, um mesmo texto pode adquirir inúmeros diferentes sentidos a depender das escolhas dramáticas tomadas por seu leitor.

Há situações, no entanto, em que a tipografia não representa um texto abstrato ainda não esculpido por seu leitor mas sim uma fala específica na qual essas escolhas dramáticas *já foram* tomadas. É o caso das legendas em filmes, que fixam visualmente as palavras que as personagens dizem no momento em que são ditas. Aqui há uma equivalência muito próxima entre o que é dito e o que se escreve, mas notem que as escolhas dramáticas se perderam: a tipografia cristaliza o *que se diz*, mas ignora quase que por completo *como foi dito*.

A depender do caso, essa perda terá menor importância: uma legenda que traduza a língua das personagens para a língua dos espectadores cumpre uma função muito específica onde, talvez, a expressão vocal não seja um foco de maior importância. Mas há também as situações em que a legenda serve para dar acesso a essa expressão vocal: sistemas de *closed captions* ajudam pessoas com diferentes graus de deficiências auditivas a ter acesso ao que falam as personagens em um vídeo e, aqui, o fato de que essa tradução visual ignora as escolhas dramáticas que o áudio contempla deve ser enquadrado como um problema.

Estados emocionais, ênfases, ironias, sarcasmos etc se perdem quando representados por uma tipografia que uniformiza toda expressão vocal, igualando o que em sua origem era desigual. De fato, a tipografia em sua origem é exatamente isso: letras construídas a partir de moldes padronizados. Com a tipografia digital, em especial aquela parametrizável, pode-se pensar a questão de maneira distinta.

Partindo da ideia de que os sentidos de um texto se modificam pelas maneiras como ele é dito, nosso trabalho busca desenvolver uma abordagem para modelar a expressividade vocal contida na prosódia para, então, representá-la tipograficamente, no que estamos chamando de *modelo prosódico-tipográfico*, vi-

sando sua aplicação na legendagem, para o que uma abordagem computacional e passível de automatização é fundamental.

Especificamente no escopo deste projeto de mestrado, buscaremos desenvolver um mapeamento entre *features* acústicas relacionadas à prosódia da voz e eixos tipográficos, que implementaremos computacionalmente e cujos resultados avaliaremos em uma série de testes com participantes que investigarão:

1. se nosso modelo produz respostas consistentes nos leitores;
2. quais *features* prosódicas melhor se relacionam a quais eixos tipográficos e em quais situações;
3. se o uso de legendas moduladas pela voz produz uma experiência subjetiva mais rica.

## *Trabalhos relacionados*

### *Prosódia enquanto dimensão emocional*

Nosso estudo é fundado na noção de que a prosódia da fala compreende uma dimensão emocional. Trata-se de uma constatação simples: o mesmo trato vocal que molda a fala é parte de um corpo que, quando sob efeito de diferentes estados emocionais, se tensiona e relaxa de diversas maneiras. A emoção não é a única dimensão que modula os parâmetros acústicos da prosódia.[1] Diferentes autores enquadram sob formas diversas as funções da prosódia (Schötz, 2002), mas para nossos propósitos são três os aspectos que nos interessam: linguístico, paralinguístico e extralinguístico:

> Enquanto o conteúdo verbal – efetivamente o significado das palavras – é considerado como informação linguística, o canal extralinguístico contém informações sobre o estado base do falante, e.g. uma pessoa grande (...) terá uma voz mais grave do que a de uma criança. Alguns parâmetros extralinguísticos são também determinados pela cultura do falante. (...) O canal paralinguístico carrega informações sobre desvios passageiros da linha de base típica (extralinguística), tais como (...) a expressão de emoções. [2]

A presença de aspectos emocionais na prosódia é explorada em da Silva et al. (2016), onde um conjunto de áudios em português foi avaliado por brasileiros e suecos. Ainda que entre os brasileiros tenha havido maior grau de concordância nas respostas, os suecos, para quem o componente linguístico dos áudios era supostamente indiferente, também conseguiram decodificar emoções.

Essas três dimensões prosódicas (linguística, para- e extra-) são articuladas pela produção e recepção de três parâmetros acústico-perceptuais [3], que modificam cada sílaba em relação às demais: intensidade (mais fraca ou forte), *pitch* (mais aguda ou grave) e duração (mais lenta ou rápida).

[1] Ainda que não tenhamos tratado a questão nos dois primeiros experimentos já realizados, lidar com o fato de que variações prosódicas em geral compreendem diversos aspectos além da emoção é um desafio importante em nosso trabalho.

[2] *Whereas the verbal content, the actual meaning of the words, is thought of as linguistic information, the extralinguistic channel contains information about the speaker's basic state, e.g. a big person (...) will usually have a lower voice than a child. Some extralinguistic parameters are also determined by the culture of the speaker. (...) The paralinguistic channel carries information about momentary deviations from the usual (extralinguistic) baseline, such as (...) the expression of emotions.* (Quast, 2001)

[3] Como mostra Barbosa (2019), a relação entre produção e percepção de sons da fala não é linear e os parâmetros que compõe a prosódia são entrelaçados: mudanças em um causam diferenças de percepção em outro.

## *A escrita e a fala*

Parte dessas variações é codificada na linguagem escrita, parte não. Com efeito, a história da escrita acompanha inúmeras mutações em convenções – inicialmente caligráficas, eventualmente tipográficas – que, muitas vezes, estão relacionadas à variações na representação de atributos prosódicos.

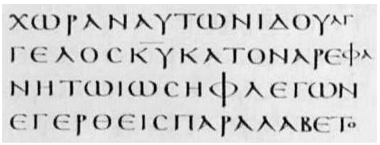

Fig. 1: Trecho do *Codex Vaticanus*, datado do século IV, exemplo de *scriptio continua*. (Commons, 2018)

Um exemplo é o *scriptio continua*: até que fossem introduzidos no século VII, não se usavam espaços ou outros sinais para separar as palavras, o texto um bloco fechado de letras – a partir do que alguns historiadores avançam a teoria[4] de que a literatura na antiguidade era predominantemente acústica (Küster, 2016): a leitura se fazia necessariamente em voz alta pois só ao converter em sons o "bloco" este se tornaria compreensível. O texto então seria "melhor percebido pelos ouvidos do que pelos olhos." (Nünlist, 2016)

A relação entre escrita e fala é complexa. Inovação medieval (ou não), a leitura silenciosa não é, como talvez se possa imaginar, desprovida de prosódia. Dentre as estruturas cerebrais envolvidas no processo de leitura, é surpreendente notar que, além das estruturas encarregadas do processamento semântico e ortográfico, a leitura demanda também aquelas tipicamente relacionadas à produção e processamento de sons (Seidenberg, 2017, cap. 7). Isso ocorre porque, mesmo quando lê silenciosamente, cabe ao leitor deduzir em sua voz interna uma representação sonora do texto, habilidade fundamentalmente relacionada à compreensão e interpretação do mesmo.

O funcionamento dessa "voz" interna é importante. Certos tipos de dislexia, por exemplo, parecem antes causados por problemas nessas estruturas fonológicas do que deficiências nas que processam imagens, mesmo que se manifestem como dificuldades na leitura (Seidenberg, 2017, cap. 8). 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 depende dessa voz interna para reduzir ambiguidades e facilitar a compreensão (Seidenberg, 2017, cap. 4). Finalmente, crianças em processo de alfabetização que leem de maneira monótona tendem a desenvolver problemas de compreensão (Bessemans, 2017).

## *Tipografia modulada pela fala*

Com a hipótese de que a falta de representação prosódica na tipografia seria um empecilho na alfabetização, a pesquisadora Ann Bessemans e seu grupo têm estudado um tipo de intervenção no texto que busca ajudar crianças a ler expressivamente. No estudo, codificaram graficamente prosódia na tipografia: texto *negrito* quando lido com maior volume, *espremido* quando mais rápido, *esticado* quando mais lento e *elevado* quando

[4] Popular, mas bastante controversa. Ver McCutcheon (2015).

agudo. Os resultados iniciais mostram que as crianças conseguiram entender e integrar essas dicas visuais na sua leitura, indicando o potencial da abordagem como uma representação intuitiva da expressividade da voz. (Bessemans, 2017)

Wölfel et al. (2015) criaram o *Design de Tipos Orientado pela Voz* (VDTD, sigla para *Voice Driven Type Design*), uma abordagem semelhante à de Bessemans mas de base computacional. Nela, mapearam atributos acústicos de cada fonema – como intensidade, *pitch* e velocidade – em grafemas de uma fonte geométrica matematicamente modelada e cujos atributos podem ser manipulados. Em uma avaliação com leitores, foram encontrados indícios tanto de que as características da fala conseguiram ser impressas no texto quanto que uma abordagem nessa linha poderia ser usada para representar emoções presentes na voz. [5]

Wölfel et al. (2015) discutem ainda algumas aplicações práticas de seu VDTD. Em volta de uma delas – legendas moduladas pela voz, em especial quando apresentadas na mesma língua em que o vídeo é falado – gravita nosso projeto. Esperamos aqui lançar contribuições para um cenário sobre o qual há benefícios amplamente documentados, como sintetiza Gernsbacher (2015): filmes legendados produzem melhoras na compreensão, atenção e memória em diversos públicos: crianças, adultos ou idosos; leitores experientes ou em fase de aprendizado; falantes ou não da língua em questão; ouvintes ou com deficiências auditivas; etc.

Em especial, Murphy-Berman and Whobrey (1983) levantam um desafio relacionado às legendas que nos parece certeiro: em seu estudo, testaram se, para crianças surdas e alfabetizadas, a presença de *closed captions* se traduzia em um entendimento mais aprofundado de um programa de televisão, especificamente em relação ao seu conteúdo *afetivo*. Os resultados indicaram que sim. Ao final, as autoras se perguntaram sobre quais efeitos gráficos conseguiriam traduzir visualmente no texto a "rica informação tonal que é negada às crianças surdas por não terem acesso à trilha sonora". Nosso projeto, acreditamos, poderá apontar um caminho.

[5] O VDTD serviu como grande inspiração para nosso próprio trabalho, mas buscamos divergir no que consideramos duas falhas em sua abordagem: (1) o uso de uma tecnologia tipográfica apenas com o propósito de representar a expressão vocal traz um grande empecilho para sua eventual adoção nos já existentes ambientes de uso de tipografia digital; (2) a avaliação foi construída principalmente com questionários onde um pequeno número de participantes fez autorrelatos de suas impressões, o que significa que os (pequenos) efeitos medidos são difíceis de generalizar.

## *Metodologia utilizada*

Visamos criar um modelo de mapeamento prosódico-tipográfico aplicável a sistemas de legendagem que tornem mais imersivo o consumo de conteúdos audiovisuais. Descreveremos a seguir nossa abordagem para criação desse modelo e, na sequência, os três experimentos (dois dos quais já realizados) que visam investigar (1) quão consistentemente respondem leitores a essa tipografia modulada pela fala, (2) quais *features* prosódicas melhor se relacionam a quais eixos tipográficos e em quais situações e, finalmente, (3) se o uso de legendas moduladas pela voz produz uma experiência subjetiva mais rica.

### *Modelo de extração e representação de prosódia*

Nossa abordagem pede que se construa um algoritmo que abstraia numericamente certos elementos acústicos que sirvam de bons representantes da expressão vocal para, então, mapeá-los visualmente enquanto tipografia. Pelo que já foi discutido, a prosódia traz um bom conjunto de parâmetros acústicos (*features*), mas considerando que nossa unidade mínima de mapeamento visual será a sílaba, um ponto adicional a favor de se usar *features* prosódicas é que elas podem ser tomadas em nível local (da sílaba) em oposição às de nível global (da frase), como discutem Rao et al. (2010).

Na versão atual do software, usamos as seguintes *features*:

1. Amplitude, calculada em decibéis como a média do Root *Mean Square* (RMS) de cada sílaba;
2. *Pitch*, calculada em Hz a partir da frequência fundamental ($f_0$) média, calculada usando o método SWIPE da biblioteca *pysptk*, aplicado em janelas de 512 amostras com limite de frequências entre 75 e 600 Hz.
3. Duração, em milisegundos.

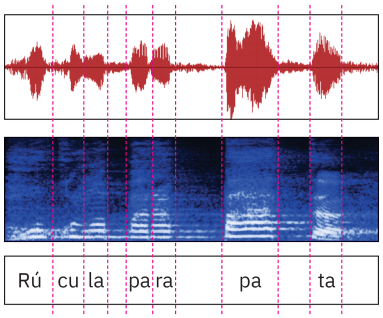


Fig. 2: Visualização esquemática de um arquivo de áudio com a frase "Rúcula para pata", ressaltando três de seus aspectos acústicos, de cima para baixo: amplitude, frequências e duração silábica.

A segunda parte do software envolve o mapeamento visual dessas *features* no texto. Como Wölfel et al. (2015), consideramos que o mapeamento das variáveis contínuas do áudio nas poucas[6] categorias tipicamente disponíveis em uma fonte digital comum (e.g. peso leve, normal e negrito) levaria a resultados que não ecoariam visualmente as sutilezas da fala e que supomos importantes para a apreensão afetiva por parte do ouvinte.

Com isso, podemos considerar dois requisitos a partir dos quais deverá emergir a solução técnica para essas representações: (1) para as modulações tipográficas, não se deve discretizar (ao menos não perceptualmente) as *features* que vierem do áudio, ou seja, a tipografia deverá conseguir ecoar mesmo mudanças sutis na fala; (2) deve ser possível que cada tipo de modulação funcione de maneira independente uma da outra na representação, pois também independentes entre si poderão ser as *features* vindas do áudio.[7]

Dada essa demarcação, excluímos, pela complexidade envolvida, o desenvolvimento de um algoritmo que reconstruísse o desenho da letra a partir de modificações em sua estrutura interna, como o fazem Wölfel et al. (2015); pelos motivos citados por Haralambous (1993), trabalhar com a METAFONT, de Donald Knuth, também seria infrutífero.

Optamos, então, por aplicar as modulações em *variable fonts*. Há já implementações em diferentes ambientes – as definições de CSS que definem como usar *variable fonts* já funcionam nos principais navegadores[8], assim como bibliotecas código livre para C e Python.

[6] Há de fato algumas raras famílias tipográficas que possuem tantos pesos a ponto de se poder questionar se a diferença entre cada um é grande o suficiente para ser percebida por si só (e.g. a Lucida Sans, que em sua versão completa tem 18 pesos entre o *UltraThin* e o *UltraBlack*). Mesmo que fosse esse o caso, nossa abordagem ainda se justificaria por, primeiro, manter-se agnóstica em relação a esta ou aquela fonte e, segundo, porque, propondo combinações de variações em eixos que não só o *peso*, acabamos indo além do que se oferece nas mais extensas famílias tipográficas disponíveis.

[7] É igualmente plausível que uma palavra seja dita de maneira forte e rápida *ou* de maneira forte a lenta, por exemplo, e a tipografia deverá conseguir ecoar as duas características de maneira independente uma da outra.

[8] 88% dos usuários já teriam acesso à tecnologia na internet. Ver: https://caniuse.com/#feat=variable-fonts, acesso em 27/11/19.

Publicada em setembro de 2016, a versão 1.8 da especificação OpenType define as *variable fonts*. Nelas, o tipógrafo cria "eixos" – dimensões que guiam diferentes tipos de modulação visual que poderá sofrer cada caractere. Uma fonte pode ter uma quantidade arbitrária de eixos, que operam de maneira independente uns dos outros (Constable and Jacobs, 2017).

Internamente, em uma *variable font* estão definidos os pontos que compõe o desenho de cada letra em sua posição central, *neutra*, além das instruções sobre como cada um desses pontos deve se transformar quando um dado eixo tiver definido seu menor e maior valores possíveis, com as posições intermediárias sendo interpoladas. (Ver exemplo esquemático na Figura 3.)

Por exemplo, uma fonte contendo os eixos *peso* e *largura horizontal* poderá oferecer uma versão estreita e grossa (largura no mínimo, peso no máximo), ou outra estendida e fina (largura no máximo, peso no mínimo), além de todas posições intermediárias possíveis. (Ver exemplo na Figura 4.)

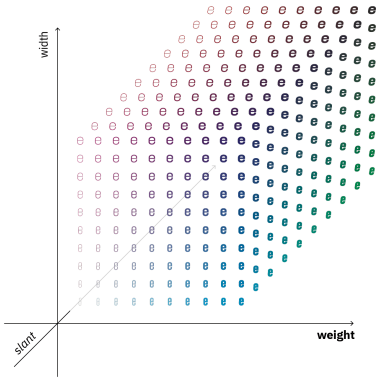


Fig. 3: Como o glifo 'e' se modifica conforme variam os valores de seus três eixos?

**curta, grossa**
longa, fina

Fig. 4: Variação em dois eixos (WGHT, ou peso, e WDTH, ou largura), na fonte *Avenir Next VF*.

### *Experimento #1: Quão consistentes são as interpretações afetivas causadas por nosso modelo?*

O propósito da primeira avaliação era investigar se nosso modelo prosódico-tipográfico produziria interpretações consistentes, especificamente na sua capacidade de representar determinadas emoções.

Para isolar os efeitos da forma tipográfica, buscamos imprimir frases cujo conteúdo textual fosse inexpressivo, ou seja, de mínimo efeito linguístico. Inversamente, as frases deveriam ter sido geradas a partir de áudios com vozes carregada de emoção, ou seja, que produzissem grandes variações visuais no desenho tipográfico. Para tanto, escolhemos a base criada por Costa (2015), onde frases de sentido neutro eram lidas por atores representando as 6 emoções – tristeza, medo, surpresa, repulsa, raiva e alegria – da tipologia "Big Six" (Ekman, 1970). Para minimizar variações entre diferentes leitores usamos a leitura de apenas uma atriz. Para nosso experimento, escolhemos as quatro frases na figura 5 (já com as modulações tipográficas de acordo com a fala).

Aplicamos o seguinte mapeamento de *features* prosódicas para eixos tipográficos:

- Amplitude → *Weight* (peso) – como em Wölfel et al. (2015) e Bessemans (2017);
- Frequência fundamental → *Slant* (inclinação) – Não encontramos exemplos específicos da inclinação na representação de prosódia, mas para evitar as dificuldades teríamos em textos longos caso replicássemos o deslocamento de linha de base representando *pitch* como em (Bessemans, 2017), decidimos por este mapeamento;

- Duração → *Width* (largura horizontal) – que em Van Leeuwen (2006) é associada à ideia de velocidade;

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 tamanho A5.

Para a avaliação realizamos sessões de *card sorting*, uma técnica de exploração e avaliação de taxonomias comum no campo da IHC (Interação Humano-Computador) para arquitetar sites ou softwares, mas que tem aplicabilidade em outras áreas sempre que se queira investigar como diferentes pessoas formulam modelos mentais para estruturar algum conjunto de dados (Soranzo and Cooksey, 2015).

Em suas formas mais comuns, a *open* e *closed card sorting*, faz-se uma atividade na qual cada participante organiza e deposita cartões dentro de diferentes envelopes. Dentro da taxonomia que se quer investigar, cada cartão representa um dado e cada envelope representa uma categoria. Estas podem ser pré-definidas (*closed cart sorting*) ou criadas no ato pelos próprios participantes (*open card sorting*). Nos interessava a primeira opção: um dos usos típicos da *closed card sorting* é testar se uma dada taxonomia é adequada para descrever um conjunto de informações. Em nosso caso, essa taxonomia seria composta pelas seis emoções a partir das quais a atriz encenou a leitura de cada frase. No experimento, esperávamos descobrir se os participantes interpretavam de maneira coerente entre si como os cartões se organizavam nas seis categorias fornecidas.

Para analisar os resultados, decidimos pelo método de comparação de *edit distances* (Nawaz, 2012). Esta é uma medida quantitativa na qual se obtém a soma de operações de troca de posição necessárias para se converter um arranjo de cartões em outro. É, assim, um indicador da divergência entre as diferentes organizações de cartão – útil para medirmos quão convergente cada participante foi em relação às "Big Six."

Se encontrássemos nos dados algum nível razoável de convergência entre as emoções implícitas em cada cartão e a forma como os participantes as interpretaram e organizaram, teríamos bons indícios de que a modulação visual da tipografia conseguiu imprimir um sentido afetivo coerente nos textos.

Também nos interessou usar o *card sorting* pois, além do bom custo-benefício (é relativamente rápido e barato (Goodman and Santos, 2006)), sendo um teste presencial, nos permitiria observar os participantes e realizar breves entrevistas semi-estruturadas ao final de cada sessão, complementando os dados numéricos com impressões, nossas e dos participantes.

Antes do início de cada sessão os participantes seriam 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 (mas não informamos maiores

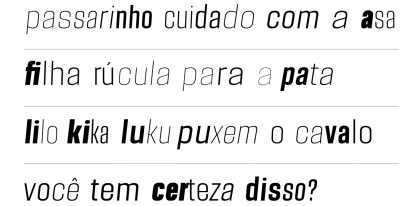


Fig. 5: Exemplo de cada frase usada nos cartões com aplicação do modelo prosódico-tipográfico.

detalhes). Para cada emoção haveria um envelope rotulado e os participantes seriam instruídos a depositar cada cartão na "emoção" correspondente a seu desenho tipográfico.

Como dito, ao final de cada sessão realizaríamos breves entrevistas semi-estruturadas, buscando levantar possíveis estratégias usadas pelos participantes na organização dos cartões, além de investiga possíveis modelos mentais formulados para explicar o funcionamento do mapeamento fala-tipografia.

### *Experimento #2: Preferências nas associações entre features prosódicas, eixos tipográficos e emoções*

O segundo experimento buscou investigar como se relacionavam as *features* prosódicas quando relacionadas a cada um dos eixos tipográficos e considerando cada uma das emoções.

Especificamente, buscamos (1) descobrir se os participantes interpretariam de maneira consistente as relações entre emoções no áudio e modulações tipográficas, (2) explorar relações entre padrões nas *features* e padrões nas preferências dos participantes e (3) ajudar a formular hipóteses sobre o sucesso (ou eventual fracasso) da capacidade de cada eixo tipográfico representar cada *feature* prosódica.

Além dos três eixos tipográficos já testados no primeiro experimento, introduzimos um quarto, que chamamos de *baseline shift*, ou deslocamento de linha de base. Nos inspiramos pela notação informal de melodia em canções brasileiras usada por Tatit (2007) que, supomos, seria especialmente adequada para representação de $f_0$. Na figura 6, um exemplo das quatro modulações tipográficas, geradas à partir do mesmo áudio.

Como são muito numerosas as combinações possíveis entre prosódia, tipografia e emoção, para o segundo experimento migramos para o ambiente online, onde seria possível testá-las sem enormes complicações logísticas. A figura 7 mostra uma captura de tela desse ambiente, mostrando especificamente o teste de preferência entre dois eixos tipográficos como representação de uma *feature* prosódica.

Na imagem, dois cartões contendo a tipografia modulada são exibidos, ambos com a mesma frase e sobre a qual uma das quatro modulações (*weight*, *width*, *slant* ou *baseline shift*) foi aplicada. Em cima das duas frases há um tocador de áudio contendo o áudio da frase a partir da qual foram extraídas as *features* prosódicas. Detalhe importante: a cada rodada só se extraía **uma** *feature*, aplicada em cada cartão a apenas **um** eixo tipográfico. A pergunta que se fazia era "Qual imagem melhor corresponde ao áudio" e, com as respostas, geraramos um conjunto de dados nos dizendo que:

> ...participante [A] escolheu o eixo [B] ao invés do eixo [C], ambos representando a *feature* [D] extraída da frase [E] e na qual fora atuada a emoção [F].

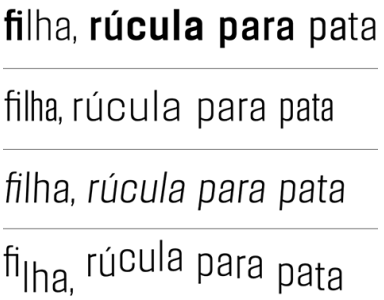


Fig. 6: As quatro modulações tipográficas testadas no segundo experimento. Respectivamente: *weight*, *width*, *slant* e *baseline shift*. Note que, diferentemente do primeiro experimento, elas aqui não estão combinadas entre si.

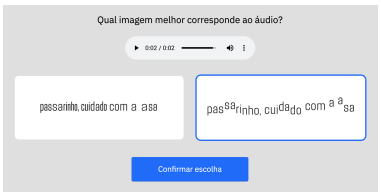


Fig. 7: No segundo experimento, participantes tinham de escolher (30 vezes) qual imagem melhor correspondia a um áudio.

Para reduzir a gama de possibilidades (e, assim, exigir uma quantidade menor de participantes antes de se chegar a resultados estatisticamente significantes), consideramos apenas duas *features* (amplitude e $f_0$), quatro emoções (raiva, alegria, tristeza e surpresa, além do áudio neutro, que introduzimos) e duas frases (*Filha, rúcula para pata* e *Passarinho, cuidado com a asa*). Em relação às *features*, vimos que nos áudios escolhidos *duração* tinha uma correlação negativa forte com *amplitude* e, para maximizar a variação entre cartões ao mesmo tempo em que minimizávamos as possibilidades, a excluímos do teste. Com as emoções, também: *nojo* e *medo* nos áudios escolhidos tinham padrões prosódicos semelhantes aos das outras emoções e, portanto, foram descartados.

No teste em si havia então 5 emoções × 4 eixos tipográficos × 2 frases, equivalendo a 40 cartões, combináveis de 60 maneiras. Em um teste piloto chegamos a 30 rodadas por participante como um bom meio termo entre querer cobrir o máximo número possível de combinações por participante e quão cansativo podia ser o experimento. Dividimos o teste em duas rodadas, cada qual testando uma das *features*. Para compensar efeitos de cansaço e inaptidão no início, além de garantir cobertura uniforme para cada par, a ordem e pares eram sempre sorteados.[9]

De modo a nos permitir coletar impressões subjetivas que escapassem às medidas mais objetivas de preferência na relação entre as três variáveis testadas, ao final das 30 rodadas havia um campo de texto livre no qual os participantes podiam nos deixar comentários quaisquer.

[9] Durante o teste fazíamos algumas medições relacionadas ao uso da plataforma. Duas delas – tempo gasto por rodada e quantidade de vezes que o participante ouviu o áudio antes de fazer a escolha – tem valores consistentemente maiores que a média nas primeiras cinco rodadas, com uma queda forte e subsequente platô nas próximas 20 rodadas e uma queda suave nas últimas rodadas – ou seja, mais para o final os participantes pareciam ir perdendo o interesse, indicando que não adiantaria estender o teste.

### *Experimento #3: Protótipo de legendagem e possíveis efeitos na imersão*

Na etapa final do projeto, pretendemos criar uma prova de conceito da tipografia modulada pela fala aplicada enquanto legendas em um ou mais vídeos para, em seguida, avaliar possíveis efeitos do modelo na fruição desse conteúdo, especificamente no que toca as dimensões de *imersão*.

Pelo estágio atual de desenvolvimento das ferramentas já desenvolvidas e pelo tempo restante do projeto, algumas escolhas foram tomadas para tornar administrável o escopo do protótipo a se criar nesta etapa:

- Em sua versão atual, nosso algoritmo não lida bem com áudios produzidos em condições adversas: ruídos, vozes sobrepostas, trilha sonora simultânea etc. Assim, tomaremos cuidado de, na escolha dos trechos de filmes/séries para a avaliação, escolher aqueles nos quais sabemos que nosso software terá bom comportamento. Ademais, nossa escolha buscará encontrar cenas centradas no diálogo entre personagens, onde supomos que nosso modelo terá maior impacto.

- Um dos desafios de uma aplicação que queira implementar nosso modelo prosódico-tipográfico será o alinhamento automático entre áudio e legendas, supondo que as demarcações temporais em um arquivo de legendas típico demarcam apenas grosseiramente os trechos de áudio correspondendo aos textos. Há abordagens distintas para problemas semelhantes[10] mas a sua exploração e adaptação para nosso contexto é por demais complexa para a prova de conceito. Nela, pretendemos segmentar manualmente os áudios.

Na avaliação, faremos um experimento nos moldes do descrito por Kruger et al. (2017) e no qual teremos participantes para os quais inglês é uma língua estrangeira e que terão acesso a vídeos cujo áudio estará em sua versão original em inglês mas apresentado nas seguintes condições: (1) sem legenda alguma, (2) com legendas em inglês em formato tradicional, ou seja, não moduladas pela prosódia e (3) com legendas em inglês geradas à partir de nosso modelo prosódico-tipográfico.

Nossa hipótese é que, como em Kruger et al. (2017), não deve haver entre o cenário (1) e os cenários (2) e (3) diferenças significantes em medidas de *imersão*, mas esperamos encontrar aumentos em medidas de *transporte*, *identificação com as personagens* e *realismo percebido*[11] – com efeitos potencialmente maiores no cenário (3) que no (2), caso nosso modelo consiga colocar em contraste diferenças de afeto na expressão vocal.

Avaliaremos a possibilidade de exibir o(s) vídeo(s) em um auditório, com controle sobre som, iluminação e distrações. Se não for possível, podem se organizar sessões menores com televisores ou, ainda, sessões individuais na internet (onde a perda de controle do ambiente será compensada pelo maior alcance do experimento).

Ao final da exibição, pediremos aos participantes que preencham um questionário de auto-avaliação. Este será composto de questões de escala Likert que, adaptadas do questionário apresentado em Kruger et al. (2016), decomporão *imersão* em alguns sub-componentes: *transporte*,[12] *identificação com as personagens*[13] e *realismo percebido*[14].

Alguns exemplos destas perguntas incluem: Enquanto eu assistia ao drama, me desliguei de atividades que ocorriam à minha volta (*transporte*); Eu entendi os eventos no drama do mesmo modo que XXX os entendeu (*identificação com as personagens*); O que aconteceu no drama é o que acontece com pessoas na vida real (*realismo percebido*).

[10] Testes iniciais nos mostraram promessa na biblioteca Python *aeneas*, que usa um algoritmo DTW (*Dynamic Time Warping*) para correlacionar e criar um mapa de alinhamento entre os MFCCs (Coeficientes Mel-Cepstrais) de uma onda de áudio contendo a fala original, extraída do arquivo de vídeo, e outra, obtida pela conversão em áudio do texto do arquivo de legendas por meio de um sistema TTS (*text-to-speech*, ou texto-fala) (Pettarin, 2017).

[11] Há diferentes definições, mas Lombard and Ditton (2006) as associa ao conceito de *presença*, que seria a "ilusão perceptual da não-mediação" que uma pessoa tem quando, ao interagir com algum conteúdo, deixa de perceber ou reconhecer a existência de um meio que conduz a comunicação, agindo como se este não existisse.

[12] Qualidade de um espectador se perceber transportado para uma realidade fictícia, com consequente suspensão de seu foco na realidade externa imediata.

[13] Afinidade que o espectador cria com as personagens, equivalendo a um entendimento empático de seus sentimentos, desafios etc.

[14] Senso de quão plausível se faz perceber uma obra, ou seja, se ela apresenta consistência narrativa de acordo com as expectativas que o espectador traz consigo.

## *Resultados e conclusões parciais*

### *O primeiro experimento*

Os resultados do primeiro experimento talvez valham mais pelo que as entrevistas revelaram do que propriamente pelas *edit-distances* medidas nas organizações dos cartões. Estas deveriam medir se os participantes conseguiram intuir na tipografia aspectos do estado emocional da voz da atriz que leu as frases impressas nos cartões, mas o efeito capturado foi muito pequeno: a *edit-distance* média no experimento é apenas 2% menor do que a que se poderia esperar caso os participantes simplesmente sorteassem as posições de cada cartão (ver Figura 8).

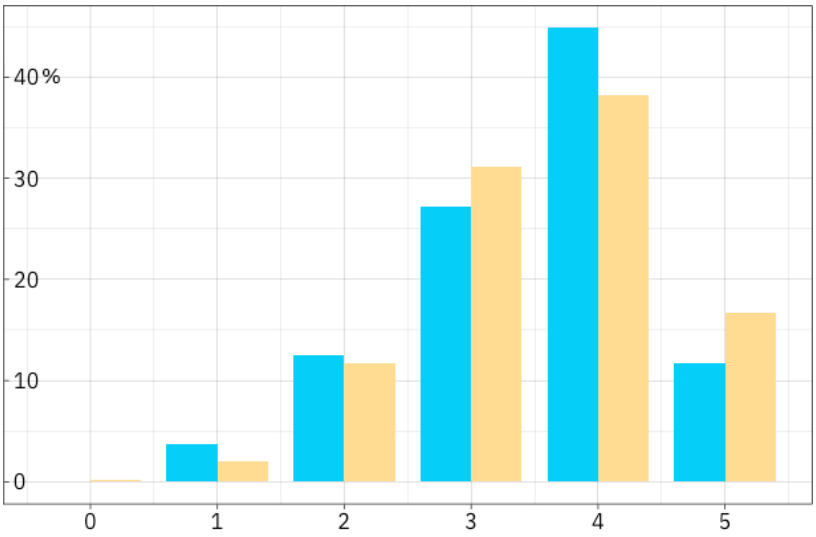


Fig. 8: *Edit-distances* das organizações coletadas dos participantes (em azul) *vs* uma organização "aleatória" (em amarelo).

Aqui, não pudemos intuir se nos dados estávamos vendo que nossa tipografia modulada pela fala era ineficaz ou se estávamos sofrendo com o mau desenho do experimento. Talvez, apostamos, seja mais o segundo caso: o modo como configuramos o *card sort* pode ter conduzido a altas taxas de erro. Como todos os cartões deveriam necessariamente ser atribuídos a uma emoção, estão embaralhados em nossos resultados coletados tanto cartões em que o participante tinha algum grau de certeza sobre a emoção quanto aqueles em que houve um "chute." Temos motivos para crer que esses chutes foram comuns, pois, tanto nas entrevistas quanto espontaneamente durante a atividade, muitos participantes relataram que a atividade era muito difícil.

A essa dificuldade da atividade em si soma-se o fato de que misturamos, e sem controlar seus efeitos, variações de três *features* prosódicas com três eixos tipográficos associados a seis emoções. Dado esse cenário, não é de se espantar que com os dados coletados nos foi muito difícil responder quão bem (ou mal) nosso modelo prosódico-tipográfico representava a voz da atriz em cada uma das emoções presentes.

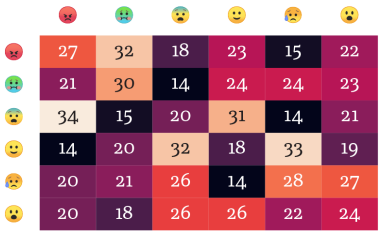


Fig. 9: Matriz de confusão do experimento de *card sort*. Emoção da atriz nas linhas, classificação dos participantes nas colunas.

Na tentativa de encontrar explicações alternativas para o baixo ganho no *edit-distance*, nos perguntamos se certos aspectos do modelo prosódico-tipográfico não estariam fazendo os participantes trocarem uma emoção por outra. Esta hipótese nos ocorreu na análise da matriz de confusão da Figura 9, que parece indicar uma possível troca feita pelos participantes especificamente entre os cartões representando medo e alegria (terceira e quarta linhas e colunas). Teria o modelo invertido o sentido de alguma das features ou eixos tipográficos em relação ao que seria intuitivo? Para testar a hipótese, calculamos novamente a *edit-distance*, desta vez considerando que os cartões com medo seriam de alegria e vice-versa. Como mostra a Figura 10, a performance melhora (8% de ganho em relação à distribuição aleatória). Mas o efeito continua pequeno.

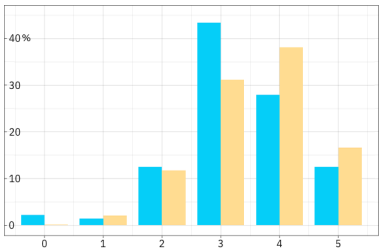

Fig. 10: *Edit-distances* das organizações coletadas, mas com troca alegria–medo (em azul) *vs* uma organização "aleatória" (em amarelo).

O que nos disseram as entrevistas? Além do muito frequente comentário de que a avaliação era muito difícil, emergiram alguns padrões em como os participantes interpretaram as modulações tipográficas. De longe o atributo mais citado, aumento no *peso* foi quase unanimemente associado a maior volume na voz, ainda que foram poucos os que perceberam que pesos leves estavam relacionados a volumes baixos na voz.

Em relação aos outros dois eixos houve grande difusão de comentários. Sobre a *inclinação*, houve quem a interpretasse como velocidade, fraqueza, tristeza, ou mesmo alguns que notaram as modulações nesse eixo mas que não souberam decodificá-las. *Largura* foi citada por apenas um participante, que intuiu corretamente que seu aumento e baixa estavam relacionados a oscilações de duração na pronúncia das sílabas.

Comentando sobre estratégias usadas para classificar cada cartão, emergiram dois principais grupos: no primeiro, mais frequente, o participante olhava para o cartão e buscava "soá-lo" mentalmente[15], tentando interpretar em sons as modulações visuais nas letras. O segundo grupo ia no sentido oposto: tentava fazer soar a frase como que sob o efeito de cada uma das seis emoções para só então buscar nos cartões aquele cuja tipografia se aproximasse do som.

Como muitos participantes sequer perceberam as modulações de *largura* e como houve grande divergência em como foram interpretadas as modulações de *inclinação*, podemos supor que ambas as estratégias relatadas sofrem de um mesmo problema: se apenas os momentos mais gritados foram bem capturados, grande parte da expressividade na voz da atriz se perdeu e, não só isso, essa perda compromete de maneira desigual cada uma das emoções.[16]

[15] Ou, mais raramente, em voz baixa, como notamos em alguns casos.

[16] de Moraes and Rilliard (2016) mostram como variações nas *features* de $f_0$ e duração são sensíveis à expressão de diferentes emoções na voz, ou seja, a prosódia se modifica de maneira desigual a depender da emoção em questão.

### *O segundo experimento*

Os resultados do segundo experimento, sintetizados na tabela 1, são mais esclarecedores. Ao cruzar como interagem as dimensões emoção na voz, *feature* prosódica e eixo tipográfico, é possível perceber que as preferências dos participantes não seguem um padrão único, variando de acordo com as diferentes possíveis combinações dessas dimensões. Pode-se vislumbrar assim que uma representação bem sucedida de qualidades expressivas da prosódia teria que ponderar os atributos dessa mesma prosódia para só então determinar quais eixos tipográficos modular.

| **RMS enquanto *feature* representada.** | | | | |
|---|---|---|---|---|
| emoção na voz | weight | width | slant | baseline shift |
| raiva | 83% | 34% | 43% | 39% |
| alegria | 34% | 54% | 47% | 64% |
| neutra | 47% | 52% | 76% | 25% |
| tristeza | 45% | 52% | 46% | 58% |
| surpresa | 72% | 38% | 49% | 43% |
| **$f_0$ enquanto *feature* representada.** | | | | |
| emoção na voz | weight | width | slant | baseline shift |
| raiva | 87% | 46% | 45% | 26% |
| alegria | 28% | 66% | 40% | 69% |
| neutra | 47% | 55% | 63% | 35% |
| tristeza | 21% | 57% | 39% | 79% |
| surpresa | 71% | 43% | 41% | 44% |

Tabela 1: Preferência dos participantes por cada modulação tipográfica quando usada com cada uma das emoções (considerando a média entre as duas frases testadas).

A preferência de 83% para *peso* quando usado para representar RMS na frase com raiva, por exemplo, indica que ela foi escolhida em 83% das vezes em que foi colocada contra os outros três eixos.

Células em verde indicam que uma dada combinação de eixo tipográfico e *feature* prosódica obteve a maior preferência dentre todas as possibilidades testadas. Células em cinza indica maior preferência apenas para aquela determinada *feature*.

Não elegemos uma preferência para tristeza pois nela a distribuição das preferências não foi estatisticamente significante.

Essa constatação, se verdadeira, carrega uma explicação possível para pelo menos parte do insucesso na abordagem testada no primeiro experimento: ao combinar sempre da mesma maneira as três *features* prosódicas com os mesmos três eixos tipográficos, tínhamos uma tipografia que coincidia com as expectativas dos participantes apenas em parte dos cartões.

Por exemplo: se, como nos mostra a tabela 1, as pessoas não acham que o eixo *peso* reproduz bem a *feature* RMS em uma voz feliz, o fato de que assim o usávamos também nos cartões em que a atriz simulou felicidade pode tê-los aproximado da raiva ou da surpresa, emoções onde o *peso* parece mais apropriado.


Fig. 11: *Inclinação* como representação de RMS na versão neutra da frase.

Outra conclusão interessante é que tanto *largura horizontal* quando *inclinação*, pouco citados nas entrevistas do primeiro experimento, tem em geral uma performance mediana, com a preferência geral tendendo à inclinação apenas quando esta

é associada a uma emoção que é diferente de todas as outras: a voz neutra. De fato, como pode-se notar na figura 11, eleita a representação mais fiel da voz neutra, as modulações tipográficas aqui são praticamente imperceptíveis. Mas esse resultado não estranha: talvez estas tenham sido as escolhas dos participantes justamente porque, em vozes que buscam ser inexpressivas, as melhores representações seriam aquelas nas quais quase não haja modulação tipográfica. Se for mesmo esse o caso, pode-se supor que uma abordagem que pondere os atributos prosódicos para só então determinar quais eixos tipográficos modular deve considerar também a possibilidade de, para determinadas situações, *não modular eixo algum*.

Se quando fizemos o primeiro e segundo experimentos ainda não conhecíamos alternativas mais robustas que o RMS para medir intensidade, na análise dos dados do segundo experimento nos pareceu oportuno nos aproveitarmos da medida obtida pela distância entre a Frequência fundamental ($f_0$) e a Centroide espectral ($C - f_0$), *feature* com a qual havíamos acabado de ter contato e que, como RMS, está relacionada à intensidade, mas que ao contrário desta última captura o esforço vocal do falante com maior independência de particularidades do ambiente de gravação. Apesar de que a *feature* não foi utilizada para modular a tipografia, os participantes tiveram acesso aos áudios, de onde por suposto conseguiram deduzir o esforço vocal da atriz.

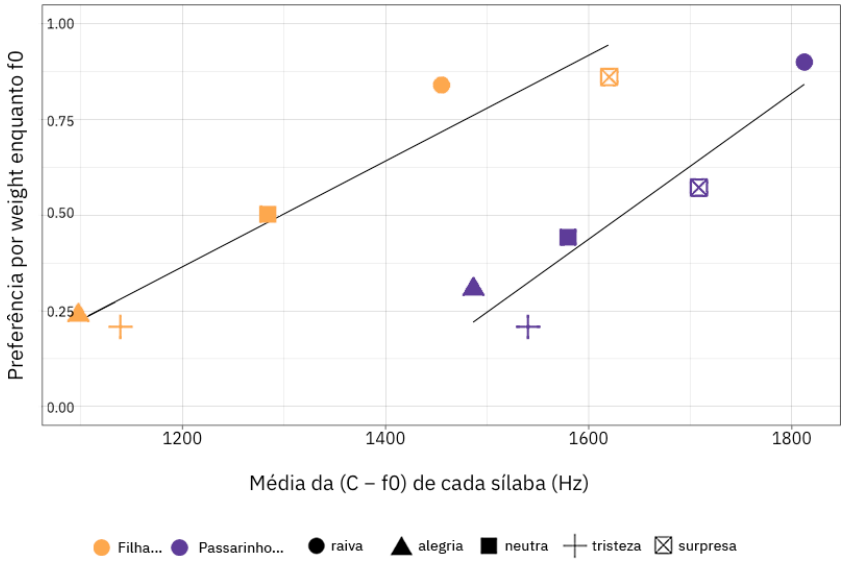


Fig. 12: Relação entre índice de preferência de *weight* enquanto representação de $f_0$ e a média de $C - f_0$ para cada emoção.

Na figura 12, mostramos como a performance do eixo *weight*, quando usado para representar a $f_0$, está relacionado ao valor $C - f_0$ médio de cada frase. Para a frase "Filha", encontramos uma equação de regressão significante[17] com um $R^2$ de 0,93 e fórmula de performance de $(C - f_0) * 10^{-3} - 1,29$. Para a frase "Passarinho", também encontramos uma equação de regressão

[17] "Filha": (F(1,3)=39.27, p<0.05), RSE=0.09; "Passarinho": (F(1,3)=24.07, p<0.05), RSE=0.10

significante com um $R^2$ de 0,89 e fórmula de performance de $(C - f_0) * 10^{-3} - 2,61$.

Já na figura 13, tomamos a performance do eixo *baseline shift* representando $f_0$ e a relacionamos às mudanças de $C - f_0$ nas duas frases. As retas aqui se invertem, e as preferências dos participantes estão inversamente correlacionadas aos valores de $C - f_0$, ainda que neste caso as equações de regressão não tenham sido significantes[18] – são, portanto, mostradas como linhas tracejadas.

[18] "Filha": (F(1,3)=1.36, p=0.33), RSE=0.23; "Passarinho": (F(1,3)=5.26, p=0.11), RSE=0.16

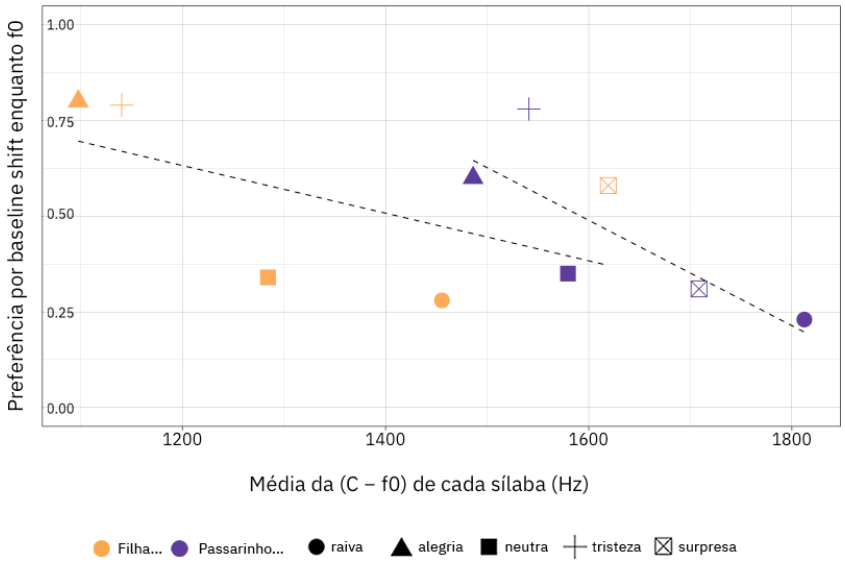


Fig. 13: Relação entre índice de preferência de *baseline shift* enquanto representação de $f_0$ e a média de $C - f_0$ para cada emoção.

Sintetizando os dois achados, vemos que conforme aumenta o esforço vocal no áudio aumenta também a preferência pelo uso do eixo *weight*. Em frases ditas com menor intensidade os participantes preferem, ainda que de maneira mais difusa, o uso do *baseline shift*.

Uma última discussão se dá com os comentários deixados pelos participantes. Muitos desses textos discorrem sobre como foram interpretadas as modulações tipográficas. Notamos que cada eixo foi associado a diversas qualidades – agressividade e firmeza para o *peso*; uma voz melódica, quase cantada, ou divertida, ou ainda bêbada para o *deslocamento vertical*; uma voz tímida para a *largura horizontal*; etc.

Surge que, nesse compêndio de qualidades, mesmo que de maneira difusa, de um modo geral os participantes parecem não estar descrevendo os atributos *acústicos* da prosódia que os eixos tipográficos representavam e tampouco as seis emoções que a atriz buscou representar. Antes, as descrições falam de vozes *modificadas*, como se os participantes não tenham deduzido nem nosso modelo prosódico-tipográfico nem, como esperávamos no primeiro experimento, as emoções que a atriz buscou representar, mas sim uma interpretação própria que assimila emoção *e* acústica.

Esse resultado condiz com o que discutem Boehner et al. quando apontam que "a comunicação de afetos em um modelo interacional (...) é mais do que a mera transmissão, (...) requerindo das partes uma interpretação ativa." [19] A partir desta perspectiva, pode-se enquadrar melhor o propósito de um sistema construído a partir de nosso modelo prosódico-tipográfico: não querê-lo como uma maneira de informar ao leitor que em dada frase a voz representava esta ou aquela emoção, categórica e de contornos bem definidos, mas sim usar a tipografia como maneira de ressaltar aspectos expressivos na voz, ambíguos como a própria voz e, como ela, abertos a múltiplas interpretações. Ainda segundo Boehner et al., a "medida de sucesso de tais sistemas [como o nosso] está então não no fato de que deduzem ou não a emoção 'correta', mas em se provocam o reconhecimento e reflexão sobre emoções em seus usuários." [20]

[19] *[C]ommunication of affect in an interactional model (...) is more than transmission, [and] requires active interpretation.*(Boehner et al., 2005)

[20] *Measures of success for such systems are therefore not whether the systems themselves deduce the 'right' emotion but whether the systems encourage awareness of and reflection on emotions in users.*(Boehner et al., 2005)

## *Publicações e apoios*

### RESUMOS EXPANDIDOS PUBLICADOS EM WORKSHOPS

Pataca, C. L.; Costa, P. D. P. (2019). Tipografia modulada pela fala. *Décimo segundo Encontro dos Alunos e Docentes do Departamento de Engenharia de Computação e Automação Industrial*- XII EADCA. FEEC/*Unicamp*. 17-18 de outubro de 2019.

### TRABALHOS PUBLICADOS EM ANAIS DE CONGRESSOS

Pataca, C. L.; Costa, P. D. P. (2019). Tipografia modulada pela fala: avaliação de um algoritmo de geração de prosódia visual em textos, p. 1882-1890. In *Anais do 9º* CIDI | *Congresso Internacional de Design da Informação, edição* 2019. São Paulo: Blucher, 2019. DOI 10.5151/9cidi-congic-4.0314

### TRABALHOS SUBMETIDOS E EM PROCESSO DE APRECIAÇÃO

Pataca, C. L.; Costa, P. D. P. Speech Modulated Typography: Towards an Affective Representation Model. *Submetido ao* ACM IUI 2020 / *International Conference on Intelligent User Interfaces*.

### APOIOS RECEBIDOS


- SIGCHI Student Travel Grant for IUI 2020 (aplicável caso o artigo submetido seja aprovado).
- Apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superio – Brasil (CAPES) – Código de Financiamento 001

## *Cronograma*


Na tabela acima apresentamos as atividades realizadas desde o início da concepção deste projeto, em março de 2017, até os últimos passos antes da defesa do mestrado, prevista para meados de maio de 2020. Etapas não realizadas estão indicadas por linhas tracejadas. A realização ou não de atividades é indicada por cores saturadas *vs* cores apagadas.

## *Considerações finais*

Neste documento apresentamos uma proposta de desenvolvimento e avaliação de uma abordagem que mapeia parâmetros da prosódia em eixos tipográficos, gerando uma tipografia modulada pela fala que, ao ecoar a expressão dramática nas vozes das personagens, poderá tornar mais imersiva a experiência de assistir a filmes nos quais ela foi usada como legenda. Nossos resultados parciais indicam que há consistência em como os leitores interpretam certas configurações do modelo, mas diferentes emoções pedem diferentes combinações de parâmetros prosódicos com eixos tipográficos.

Nos próximos meses teremos uma prova de conceito na qual será possível avaliarmos se legendas sob a forma de tipografia modulada pela fala alteram a experiência subjetiva de espectadores.

## *Bibliografia*

Barbosa, P. A. (2019). *Prosódia*. Parábola Editorial.

Bessemans, A. (2017). Expressive typography to improve communication. ATypI MONTRÉAL 2017. http://youtu.be/JfsixaAmNOw.

Boehner, K., DePaula, R., Dourish, P., and Sengers, P. (2005). Affect: from information to interaction. In *Proceedings of the 4th decennial conference on Critical computing: between sense and sensibility*, pages 59–68. ACM.

Commons, W. (2018). File:codex vaticanus matthew 1,22-2,18.jpg — wikimedia commons, the free media repository. Acesso em 28/7/2018.

Constable, P. and Jacobs, M. (2017). Opentype font variations overview. `https://docs.microsoft.com/en-us/typography/opentype/spec/otvaroverview`. Acesso em 7/4/2018.

Costa, P. D. P. (2015). *Two-Dimensional Expressive Speech Animation*. PhD thesis, Faculdade de Engenharia Elétrica, Universidade Estadual de Campinas, Campinas, SP, BRA.

da Silva, W., Barbos, P. A., and Abelin, Å. (2016). Cross-cultural and cross-linguistic perception of authentic emotions through speech: An acoustic-phonetic study with brazilian and swedish listeners. *DELTA: Documentação de Estudos em Lingüística Teórica e Aplicada*, 32(2):449–480.

de Moraes, J. A. and Rilliard, A. (2016). Prosody and emotion in brazilian portuguese. In *Intonational Grammar in Ibero-Romance*, pages 135–152. John Benjamins Publishing Company.

Ekman, P. (1970). Universal facial expressions of emotions. *California mental health research digest*, 8(4):151–158.

Gernsbacher, M. A. (2015). Video Captions Benefit Everyone. *Policy Insights from the Behavioral and Brain Sciences*, 2(1):195–202.

Goodman, M. B. and Santos, G. J. (2006). Card sort technique as a qualitative substitute for quantitative exploratory factor analysis. *Corporate Communications: An International Journal*.

Haralambous, Y. (1993). Parametrization of PostScript fonts through METAFONT — an alternative to Adobe Multiple Master fonts. *Electronic Publishing*, 6:145–157.

Kruger, J.-L., Doherty, S., and Soto-Sanfiel, M.-T. (2017). Original language subtitles: Their effects on the native and foreign viewer. *Comunicar: Media Education Research Journal*, 25(50):23–32.

Kruger, J.-L., Soto-Sanfiel, M. T., Doherty, S., and Ibrahim, R. (2016). Towards a cognitive audiovisual translatology. In *Reembedding Translation Process Research*, pages 171–194. John Benjamins Publishing Company.

Küster, M. W. (2016). Writing beyond the letter. *Tijdschrift voor Mediageschiedenis*, 19(2).

Lombard, M. and Ditton, T. (2006). At the heart of it all: The concept of presence. *Journal of Computer-Mediated Communication*, 3(2):0–0.

McCutcheon, R. W. (2015). Silent reading in antiquity and the future history of the book. *Book History*, 18(1):1–32.

Murphy-Berman, V. and Whobrey, L. (1983). The impact of captions on hearing-impaired children's affective reactions to television. *The Journal of Special Education*, 17(1):47–62.

Nawaz, A. (2012). A comparison of card-sorting analysis methods. In *The 10th Asia Pacific conference on computer human interaction* (APCHI2012).

Nünlist, R. (2016). Users of literature. In Hose, M. and Schenker, D., editors, *A companion to Greek Literature*, chapter 19, pages 296–297. John Wiley & Sons, West Sussex.

Pettarin, A. (2017). Aeneas: How does this thing work? `https://github.com/readbeyond/aeneas/blob/master/wiki/HOWITWORKS.md`. Acessado em: 24/7/2018.

Quast, H. (2001). Automatic recognition of nonverbal speech: An approach to model the perception of para-and extralinguistic vocal communication with neural networks. Master's thesis, University of Gottingen.

Rao, K. S., Reddy, R., Maity, S., and Koolagudi, S. G. (2010). Characterization of emotions using the dynamics of prosodic features. In *Speech Prosody 2010-Fifth International Conference*.

Schötz, S. (2002). Linguistic & paralinguistic phonetic variation in speaker recognition & text-to-speech synthesis. In *In Speech Technology, GSLT*. Citeseer.

Seidenberg, M. (2017). *Language at the Speed of Sight: How We Read, Why So Many Can't, and What Can Be Done About It*. Basic Books, New York, 1st edition. Kindle version.

Soranzo, A. and Cooksey, D. (2015). Testing taxonomies beyond card sorting. `https://www.slideshare.net/atrebla/testing-taxonomies-beyond-card-sorting`. Acessado em: 22/7/2018.

Tatit, L. (2007). *Todos Entoam — Ensaios, Conversas e Canções*. Publifolha, São Paulo, Brazil, 1st edition.

Van Leeuwen, T. (2006). Towards a semiotics of typography. *Information design journal*, 14(2):139–155.

Wölfel, M., Schlippe, T., and Stitz, A. (2015). Voice driven type design. In *2015 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*.