Universidade Estadual de Campinas
Faculdade de Engenharia Elétrica e de Computação

PROJETO DE MESTRADO

A voz na letra: tipografia modulada pela fala

ALUNO
Caluã de Lacerda Pataca
calua.pataca@gmail.com

Aluno de Mestrado

ORIENTADORA
Profª Drª Paula Dornhofer Paro Costa
paula@fee.unicamp.br

Departamento de Engenharia de Computação e Automação Industrial

Faculdade de Engenharia Elétrica e de Computação

Campinas, agosto de 2018

Resumo

A leitura fluente se dá quando o leitor consegue relacionar internamente letras e símbolos com os sons da língua, que ele em geral já traz consigo, em uma espécie de *fala* interna. Em certos contextos esse processo funciona mal — algumas crianças em processo de alfabetização penam para se tornar leitoras fluentes, falantes de língua estrangeira nem sempre mapeiam bem som a letra, surdos não têm as referências sonoras etc. Nosso projeto se insere no rol de pesquisas em tecnologia tipográfica que visam manipular a forma do texto de modo a complementar o conteúdo explícito na escrita com aspectos presentes na fala, investigando os impactos dessas intervenções sobre a leitura. Especificamente, construirá um algoritmo computacional que abstraia características acústicas da fala para mapeá-las visualmente em formas tipográficas, imbuindo no texto elementos expressivos da voz, em especial aqueles relacionados à emoção. Nossa hipótese é que na leitura de um texto composto com essa tipografia modulada pela voz haverá um aumento em medidas como *transporte* (perceber-se "dentro" da obra), *identificação com as personagens* (empatia do espectador para com as personagens) e *realismo percebido* (quão plausível parece ser uma narrativa). Juntas, essas medidas indicam a *imersão* do leitor em uma dada obra. Para testá-la, serão investigadas quais emoções são perceptíveis em uma tipografia modulada pela prosódia na voz. Em seguida — e com essa mesma tipografia usada como legenda — será avaliada sua influência nos espectadores de um filme.

1 Introdução

Ler é uma habilidade cognitiva de alto nível, que exige longo período de treinamento e que envolve intenso processo neurológico. Dentre as estruturas cerebrais envolvidas no processo de leitura, é surpreendente notar que, além daquelas encarregadas do processamento de imagens, a leitura exige também o emprego das estruturas que lidam com processamento de sons (SEIDENBERG, 2017, cap.7). Isso ocorre porque, mesmo quando lê silenciosamente, cabe ao leitor deduzir (ou inventar) em sua voz interna qual é a musicalidade do texto, habilidade fundamentalmente relacionada à compreensão e interpretação do mesmo.

Investigar como essa "voz" emerge a partir do texto não se trata de uma questão meramente *estética*. Certos tipos de dislexia, por exemplo, parecem antes causados por problemas nas estruturas neurológicas que processam sons do que deficiências no processamento de imagens, mesmo que se manifestem sob a forma de dificuldades na leitura (SEIDENBERG, 2017, cap.8). Além disso, crianças em processo de alfabetização que leem de maneira monótona, ou seja, que não conseguem extrair do texto a expressividade da fala, tendem a desenvolver problemas de compreensão (BESSEMANS, 2017). Finalmente — e 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 ter acesso à informação fonológica contida no texto e, assim, reduzir ambiguidades e facilitar a compreensão (SEIDENBERG, 2017, cap.4).

Neste contexto, o presente projeto se insere no âmbito da pesquisa em tecnologia tipográfica. Se debruça sobre algoritmos computacionais de transformação visual de texto almejando transformar o processo de aquisição de informação por meio da leitura mais intuitivo e acessível, com potenciais aplicações no auxílio ao processo de alfabetização, ensino de línguas estrangeiras e, entre ainda outras, como tecnologia assistiva, em especial para a população surda.

Em particular, o presente projeto propõe investigar e avaliar algoritmos de mapeamento automático de parâmetros acústicos da fala expressiva para parâmetros tipográficos do texto, de maneira que um leitor, que tenha tido acesso ao texto mas não à fala, consiga recuperar elementos presentes exclusivamente na expressão sonora.

As seções seguintes apresentam um panorama histórico da evolução de tecnologias e convenções que sustentam a leitura, incluindo trabalhos recentes que servem de referência para este projeto. Além disso, são apresentados os objetivos específicos do projeto, a abordagem metodológica proposta, o plano de trabalho e seu cronograma e, por fim, os resultados esperados para o projeto.

2 Tecnologia do texto: panorama histórico e pesquisas recentes

Por natural que possa parecer, o ato de ler um texto individualmente e em silêncio é uma invenção relativamente recente na cultura ocidental. A história da escrita é cheia de meandros, mas por muito tempo predominou o *scriptio continua*: um texto disposto em um bloco fechado, sem espaços entre as palavras. Para tirar sentido dessa massa de letras, não se esperava do leitor uma leitura silenciosa — em sua essência, o texto servia para ser *falado*.

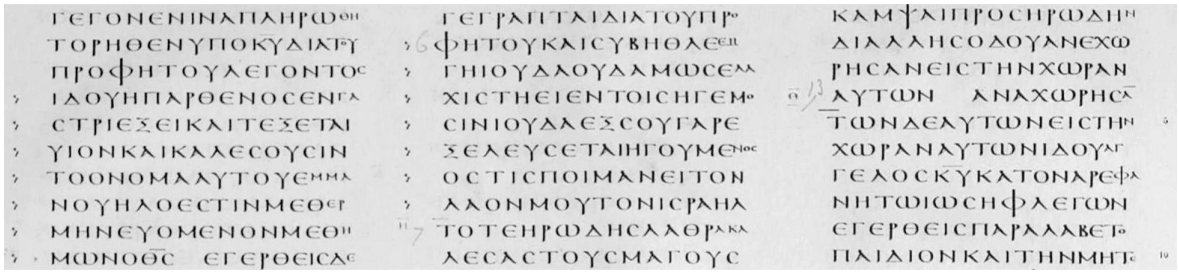

Figura 1: Trecho do *Codex Vaticanus*, datado do século IV. Pontuações e marcações foram adicionadas nos séculos após sua criação. (Adaptado de (COMMONS, 2018).)

Assim era no Grego clássico, por exemplo, onde a literatura era predominantemente considerada em sua dimensão *acústica* (KUSTER, 2016): um texto, quando lido, o era em voz alta. A forma com que se dava essa leitura, ou seja, onde eram colocadas as ênfases, silêncios, diferentes entonações etc, dependia em grande parte de convenções de declamação, que não só variavam por tipo de texto como estavam sujeitas a um grande nível de subjetividade na interpretação. Focada nos aspectos de um texto "melhor percebidos pelos ouvidos do que pelos olhos"(NUNLIST, 2016), era uma leitura centrada em uma relação estreita entre letras e voz que, por impor um processo de treinamento muito exigente, era uma atividade de modo geral restrita a poucos.

Eventualmente, essa leitura recitada deu espaço a outra, *silenciosa*. Dois momentos marcaram essa transição: a popularização do espaço em branco separando palavras, a partir do

século VII, e as convenções de pontuação, como as temos hoje, a partir do século XVI.

Em nosso sistema de escrita, o espaço em branco surgiu na Grã-Bretanha, em especial na Irlanda, em um contexto em que leitores cuja língua materna não era o Latim criaram uma sinalização que codificasse os inícios e fins das palavras em textos religiosos que, de outra forma, eram impenetráveis (KUSTER, 2016).

O segundo momento, a partir do qual se propagaram as convenções de pontuação como as temos ainda hoje, se deu em parte pelo trabalho do editor, tipógrafo e impressor Aldo Manúcio, no período entre os séculos XV e XVI:

> As regras [definidas pela família de Manúcio] ignoravam as antigas marcações que ajudavam a leitura em voz alta. Os livros serviriam, agora, para serem lidos e entendidos, e não entoados. [...] Nos setenta anos entre o período de Aldo Manúcio, o velho, e Aldo Manúcio, o jovem, o cenário de tal maneira se transformou que em 1566 Aldo Manúcio, o jovem, pôde declarar que a função principal da pontuação era a de clarificar a sintaxe. (TRUSS, 2003 apud KUSTER, 2016), tradução livre.

Apoiada nesses desenvolvimentos foi surgindo, então, uma nova leitura, mais simples e mais acessível. Ao invés de um denso bloco de letras como no *scriptio continua*, fechado a ponto de depender que fosse lido em voz alta para se fazer entender, as novas marcações gráficas ajudavam o leitor a atravessar o texto com mais facilidade e agilidade, o que passou, mais e mais, a se dar de maneira silenciosa e individual (KUSTER, 2016).

Ainda assim, muito da informação contida no discurso oral não está representada diretamente na ortografia e pontuação modernas. Talvez essa distância entre texto e fala esteja na origem de certas dificuldades com a leitura que enfrentam algumas populações — perceber a ligação entre imagens e sons depende de conhecimentos prévios nem sempre acessíveis às crianças, aos falantes de outras línguas, aos surdos etc —, mas, de qualquer modo, é fato que há lacunas: para ênfases, tons de voz, dilatações no tempo etc, faltam maneiras, implícitas ou explícitas, de se codificar a informação em um texto.

Há, é claro, sinais que codificam a prosódia, desde os mais comuns, como exclamações, interrogações, vírgulas e pontos[1] àqueles empregados em certas línguas que possuem um

[1] A escrita informal nas mensagens de texto de celular, assim como a publicitária, busca superar alguns desses limites de que falamos através de pequenas transgressões nas convenções de gramática, subvertendo a pontuação comum para aproximar o texto da oralidade. Kuster (2016) traz alguns exemplos de literatura infanto-juvenil na qual os pontos (e.g. "*You. People. Are. Crazy!*") servem como marcações de ritmo.

repertório estendido de sinais, como a interrogação (¿) e exclamação (¡) invertidas que, no Espanhol, indicam com antecedência qual a entonação da frase. Finalmente, existem sinais ainda mais raros, como o *interrobang* (‽), uma espécie de interrogação exclamada. Não obstante, muito da prosódia cabe ao leitor inferir.

Visando buscar soluções para os problemas que decorrem da falta de informações que a pontuação típica não codifica, ou ainda, explorando novos usos do texto no contexto da cultura digital e das artes, certos pesquisadores têm explorado maneiras de se alterar, ou expandir, essas convenções.

A pesquisadora belga Ann Bessemans e seu grupo têm estudado um tipo de intervenção no texto que apresenta potencial para ajudar na alfabetização de 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 representados visualmente no desenho e disposição das letras. Assim, uma palavra que deve-se ler com maior intensidade que as outras pode estar em negrito; se a sugestão for que se leia-a mais lentamente, esticam-se suas letras no sentido horizontal; se a voz deve ser mais aguda, a palavra poderá ser posicionada acima da linha de base. Os resultados iniciais têm se mostrado promissores(BESSEMANS, 2017).

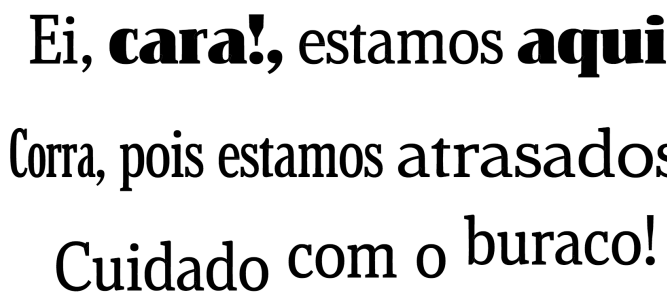

Figura 2: Exemplo demonstrando, respectivamente, modulações na grossura, condensação ou expansão horizontal e posição na linha de base. (Adaptado de Bessemans (2017))

Trabalhando sob um viés distinto, Ondrej Jelinek (2013) explora semelhanças e contrastes entre a linguagem escrita e a falada, buscando criar aproximações visuais entra as duas através de uma tipografia altamente experimental e expressiva. Os resultados, embora tragam questões estéticas interessantes, aqui nos interessam pouco: a ênfase dada à expressividade da forma vem em detrimento de sua legibilidade, o que acaba por tornar as soluções restritas a uma gama de usos muito pequena — dentre as quais não estão as situações que pretendemos explorar.

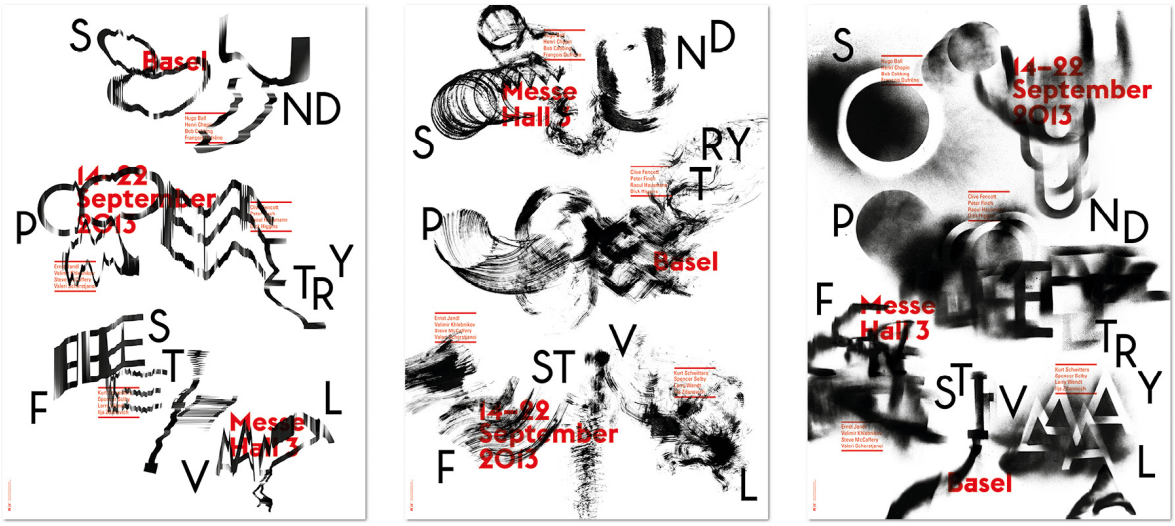

Figura 3: Três pôsteres para um fictício *Sound Poetry Festival*, que exploram, cada um, formas de evocar aspectos da expressividade da voz na tipografia (JELINEK, 2013).

Já na computação, Wolfel, Schlippe e Stitz (2015) descrevem um sistema onde, como em Bessemans (2017), a tipografia é modificada a partir de propriedades da voz, mas no qual se trabalha a partir de uma fonte matematicamente modelada, passível, portanto, de distorções via algoritmo em correspondência às propriedades acústicas de uma gravação de áudio. 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.

Paralelamente, foi publicada em setembro de 2016 a versão 1.8 da especificação OpenType, padrão suportado pela indústria e que define a tipografia digital para sistemas computacionais. Nesta versão, a especificação adicionou o conceito de *variable fonts*. Nelas, o tipógrafo define "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; JACOBS, 2017).

Internamente, no arquivo de uma *variable font* estão definidos os pontos que compõe o desenho de cada letra em sua posição central[2], *neutra*, além das instruções sobre como cada um desses pontos deve se transformar quando um dado eixo tiver definido seu menor valor possível *e* seu maior valor possível, com as posições intermediárias sendo interpoladas linearmente conforme a demanda.

[2]Tipicamente, um desenho vetorial cujos pontos descrevem curvas de Bézier quadráticas ou cúbicas.

Por exemplo, uma fonte contendo os eixos *peso* e *largura horizontal* poderá oferecer uma versão estreita e grossa (eixo da largura no valor mínimo e eixo do peso no valor máximo), ou outra estendida e fina (eixo da largura no valor máximo e eixo do peso no valor mínimo), além de todos os intermediários possíveis.

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

Assim, a manipulação de atributos visuais que em Wolfel, Schlippe e Stitz (2015) foi laboriosamente construída a partir de distorções em uma fonte matematicamente modelada poderá, como proporemos neste projeto, ser mapeada aos eixos de uma *variable font*. Dado que o OpenType é há já muitos anos o formato adotado pelos principais sistemas operacionais, além de mais simples essa solução tem ainda a vantagem de poder ser estendida a inúmeros contextos além de nosso projeto (e.g. sistema web, aplicativos móveis etc).

Ainda assim, independente de empregar uma abordagem talvez já prematuramente datada tecnicamente, o trabalho de Wolfel, Schlippe e Stitz (2015) aponta caminhos promissores. Além da ajuda na alfabetização, são discutidas aplicações como auxílio do 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.

Em torno desta última aplicação — legendas moduladas pela voz, em especial quando apresentadas na mesma língua em que o vídeo é falado — estará centrada nossa pesquisa. Esperamos com ela 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 e Whobrey (1983) levantam um desafio relacionado às legendas, em torno do qual gravita nosso projeto: no estudo, foi testado 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, os autores se perguntam 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.

3 Objetivos

Este projeto tem como objetivo propor, implementar e avaliar um modelo de mapeamento de características acústicas da voz, em especial aquelas relacionadas à expressão da prosódia, para modulações visuais em uma família tipográfica. Uma vez implementado, um segundo objetivo do projeto é avaliar se essas modulações permitem ao leitor inferir a presença de emoções no áudio que não estejam explicitamente presentes no texto para, por fim, apurar que efeitos essa abordagem pode ter na imersão de um espectador que assiste a um filme em cujas legendas foram aplicadas essa modulações, nesse caso calculadas a partir de propriedades acústicas extraídas da fala dos atores.

4 Plano de trabalho e Metodologia

A seguir apresentamos um plano de trabalho de 18 meses, contados a partir da submissão desta proposta. A metodologia proposta é dividida em cinco etapas principais. Além da escrita e pesquisa bibliográfica, incluirá duas avaliações, com profissionais e leigos, intercaladas com duas etapas de desenvolvimento dos softwares que fornecerão os insumos para as avaliações.

4 · 1 ETAPA 1 — Extração de *features* de áudio e sua representação visual em texto

Nesta etapa será desenvolvida aplicação software que gerará as imagens necessárias para a avaliação da decodificação de emoções representadas em textos. A aplicação será composta de duas partes distintas:

Features de áudio

A primeira parte envolverá a extração de *features* de áudio que se possa correlacionar com a expressão de emoções presentes em um registro de fala. A intenção não será, no entanto, deduzir *quais* seriam essas emoções, já que elas não estarão representadas diretamente na tipografia[3], e sim seus *vestígios*, ou seja, as características na fala que, independentes do conteúdo dito propriamente, levam o ouvinte a nela intuir um caráter afetivo.

Escolher um bom conjunto de "vestígios" — as *features* de áudio a se extrair — é, então, um dos desafios nesta etapa. Optamos por aquelas que se relacionam às características supra-segmentais, ou seja, aquelas ligadas à prosódia, que aparecem quando se considera o som a partir da sílaba e/ou da palavra como um todo. Especificamente, trabalharemos com (1) duração, (2) intensidade e (3) entonação (contorno de F_0) da sílaba.

Além de haver literatura que apoie o uso dessas características na detecção de emoções (KOOLAGUDI; RAO, 2012), e além de nosso estudo estar relacionado aos resultados parciais positivos de prosódia representada na tipografia de Bessemans (2017), a escolha se sustenta também na constatação de que, como estarão representadas diretamente no texto, as *features* escolhidas devem cumprir dois requisitos principais:

1. Devem ser unidimensionais, ou seja, cada *feature* deverá ser passível de mapeamento em uma dimensão visual da tipografia. Como veremos a seguir, essas dimensões serão associadas a diferentes eixos em *variable fonts*.

2. Devem trazer 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). O motivo aqui é que a forma com que cada *feature* varia no tempo de uma frase estará mapeada visualmente à tipografia de cada sílaba.

Tipografia modulada pela voz

A segunda parte do software envolve esse mapeamento visual de *features* no texto.

[3] Sobre a possibilidade de se expressar diretamente na tipografia um conteúdo afetivo, o trabalho de Suksumek (2017) apresenta uma *variable font* experimental na qual há um eixo de valência que determina a modulação do desenho das letras entre formas percebidas como negativas (com um traço anguloso, de formas duras) e positivas (com um traço mais fluído, de formas suaves), com formas neutras no meio.

Como Wolfel, Schlippe e Stitz (2015), consideramos que o mapeamento das variáveis contínuas do áudio nas poucas categorias comumente disponíveis em uma fonte típica (e.g. peso leve, normal e negrito) levaria a resultados de poucas nuances, sem as sutilezas presentes na fala que são importantes na detecção de emoções 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: como é igualmente plausível que uma palavra seja dita de maneira forte e rápida *ou* de maneira forte a lenta, por exemplo, a tipografia deverá conseguir ecoar as duas características de maneira independente uma da outra.

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 Wolfel, Schlippe e Stitz (2015); pelos motivos citados por Haralambous (1993), trabalhar com a METAFONT, de Donald Knuth, também seria infrutífero.

Optamos, então, por aplicar as modulações em fontes OpenType 1.8, especificamente exemplares construídos enquanto *variable fonts*. Há já implementações em diferentes ambientes — as definições de CSS que definem como usar *variable fonts* já funcionam nos principais navegadores[4], assim como bibliotecas código livre para C e Python.

Resumindo, para um dado texto lido em voz alta, extrairemos um conjunto de *features* de áudio e as mapearemos, sílaba a sílaba, nos eixos de uma *variable font*.

4 · 2 ETAPA 2 — Avaliação da decodificação das emoções representadas no texto

O software criado na ETAPA 1 permite muitas combinações de *features* e eixos nos quais representá-las. Nesta segunda fase avaliaremos (1) que sentidos conseguimos imprimir a um texto apenas pelo uso das modulações tipográficas, e (2) sob quais combinações de parâmetros esses sentidos são mais pregnantes.

[4] Cf. https://caniuse.com/#feat=variable-fonts, acesso em 27 de julho de 2018. Note que a composição de navegadores no Brasil é bem adaptada a esse conjunto de funcionalidades.

Quais textos e qual áudio?

Deve-se considerar que a base de áudios utilizada para a extração de *features* precisará cumprir dois requisitos: os textos devem ter conteúdo neutro, minimizando na medida do possível a influência que seu sentido possa ter quando eles, já modificados visualmente, forem lidos e interpretados; de maneira inversa, os *áudios* contendo a leitura desses mesmos textos devem ser o menos neutros possível, isto é, neles é importante que tenha-se buscado na leitura a expressão de uma gama ampla de emoções, tanto no que diz respeito a sua tipologia quanto na própria intensidade com que cada tipo de emoção foi representada. Imagina-se ser possível, por exemplo, que as modulações tipográficas só sejam perceptíveis quando o texto tiver sido "atuado" com grande exuberância expressiva, mas para investigar essa (e outras) possibilidades, é importante que haja riqueza de interpretações na base de áudio.

Assim, mesmo que criada sob motivações diversas às nossas, a base descrita em Costa (2015) nos parece promissora. Nela, e cumprindo o primeiro requisito, foram criados 10 frases sem relação entre si e "neutras, de modo que cada uma permita [pelos atores] a interpretação de diferentes emoções" (COSTA, 2015, p.50). Essas frases foram então, e cumprindo nosso segundo requisito, lidas por 4 atores, cada um buscando representar as 6 emoções da tipologia "Big Six" (tristeza, medo, surpresa, repulsa, raiva e alegria) a partir de 3 níveis de "ativação" (tímido, neutro e expansivo). Há na base, portanto, 720 trechos.

Quais famílias tipográficas?

Do lado da tipografia, a escolha deverá considerar fontes que (1) sejam *variable fonts*; (2) possuam, no mínimo, eixos de largura e grossura da letra[5]; (3) não sejam fontes *display*, ou seja, que não tragam problemas de legibilidade e cujo desenho, antes de qualquer modulação, não seja em si excessivamente expressivo; (4) sejam de livre distribuição[6].

Aqui, ao contrário da escolha da base de áudio, há uma lista muito grande de opções que cumprem todos os requisitos, e a troca de uma fonte por outra é muito simples. Ainda que

[5] Bessemans (2017) representa o *pitch* deslocando verticalmente as palavras. Buscaremos soluções diferentes, pois esse deslocamento dificultaria o uso desses textos visualmente modulados em contextos com mais de uma linha, como pretendemos explorar nas etapas seguintes.

[6] Há atualmente designers — em especial da Adobe, Google e Microsoft, que juntas à Apple propuseram o OpenType 1.8 — disponibilizando gratuitamente *variable fonts*, tanto para explorar a tecnologia quanto promover seu uso. Sites como <axis-praxis.org>, <play.typedetail.com> e <v-fonts.com> etc as têm catalogado.

já tenhamos feito explorado algumas possibilidades (entre elas, Kairos Sans, Avenir Next VF, Compressa etc), maiores explorações serão feitas durante o projeto.

Em cima dos áudios e fontes definidos esperamos, então, construir uma avaliação com leitores que explore quais emoções encenadas na leitura conseguem ser impressas no texto.

Avaliação com Card sorting

Card sorting é uma técnica de exploração e avaliação de taxonomias para quando não há uma taxonomia plenamente aceita e se quer sondar os pontos de vista de diferentes *stakeholders*, técnicos ou leigos. Costuma ser muito usada no campo da IHC (Interação Humano-Computador) para arquitetar sites ou softwares, mas tem aplicabilidade em outras áreas sempre que se queira investigar como diferentes pessoas formulam modelos mentais para estruturar algum conjunto de dados (SORANZO; COOKSEY, 2015).

Em suas duas formas mais comuns, a *open* e *closed card sorting*, a técnica se dá como uma atividade na qual se pede para que um grupo de pessoas[7], uma a uma, organizem um conjunto de cartões de papel em 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*). Em um site, por exemplo, os cartões podem ser páginas e os envelopes seções. Com medidas da convergência e divergência na comparação de como cada participante organizou os cartões, pode-se então definir uma estrutura de navegação mais intuitiva.

Aqui, a *closed card sorting* é a variante que nos interessa. Um de seus usos típicos se dá quando se quer testar se uma dada taxonomia é adequada para descrever um conjunto de informações. Em nosso caso, a taxonomia pré-definida são as seis emoções a partir das quais os atores encenaram cada gravação de áudio[8]. Com o *card sorting* nos interessa descobrir se essa taxonomia serve para descrever os textos que foram modulados visualmente de acordo com o áudio — mas que, fora isso, não possuem outras informações a partir das quais os participantes poderiam inferir pertencimento a esta ou aquela emoção.

[7] *Stakeholders* que podem ou não ser especialistas — em algumas situações, busca-se um público leigo para desvendar como ele organiza mentalmente as informações de uma área na qual seu conhecimento é superficial.

[8] Não reproduziremos como categoria para os participantes os três níveis de ativação que também serviram de base para a atuação dos atores — essa informação, no entanto, será preservada para uso em nossas análises.

Para analisar os resultados, usaremos o método de *edit distance*, na qual se obtém uma soma de quantas "operações" de transformação seriam necessárias para se converter um arranjo de cartões em outro, ou seja, um índice da divergência entre os conjuntos — útil tanto para encontrarmos quão convergente cada participante foi em relação à organização de origem dos cartões — as seis emoções —, quanto para medirmos quão divergentes entre si foram os participantes, como discute Nawaz (2012).

Assim, se de fato for detectado algum nível razoável de convergência entre as emoções escondidas em cada cartão e a forma como os participantes as interpretaram e organizaram, teremos bons indícios de que a modulação visual conseguiu de fato indicar implicitamente no texto um nível adicional de informação que não apenas o que está explicitamente escrito.

Quais leitores?

O *card sorting* será aplicado a dois públicos: estudantes de nível universitário em design, de um lado, e estudantes de outras áreas quaisquer, de outro. São três os motivos para essa escolha: (1) sendo alunos de graduação ou acima, esses são públicos para os quais é possível supor uma leitura fluente; (2) estarem, como estamos nós, os pesquisadores, em uma universidade, facilitará o seu recrutamento; (3) se o primeiro público em comparação com o segundo tiver muito mais convergência entre sua organização dos cartões, na média, e a organização esperada, poderemos supor que as soluções visuais precisam de ajuste — afinal, as emoções terão sido detectadas, mas dependendo de um olhar para tanto.

4 · 3 ETAPA 3 — Adaptação do software para o contexto de filmes legendados

A partir dos resultados das etapas anteriores, o software criado até aqui será adaptado, preparando-o para a avaliação final. Para ela, foram pensados os ajustes a seguir.

Definição dos parâmetros de melhor performance

Terá sido possível na avaliação com *card sorting* testar quais *features*, representadas sob quais parâmetros visuais, terão tido melhor compreensão por leitores. Nesta terceira etapa, essas configurações de melhor performance alimentarão os parâmetros do algoritmo de extração e representação na tipografia a serem usados nas legendas de todo um filme — a ideia nesta

etapa não é mais focar na sintaxe visual das modulações e sim medir seus possíveis impactos em uma situação de leitura, então não se quer que dentro de um mesmo filme as legendas apresentem abordagens conflitantes.

Alinhamento som-letra

Teremos nesta etapa dois arquivos de entrada: um arquivo de vídeo e uma arquivo de legenda, transcrevendo as falas contidas no arquivo de áudio contido no vídeo.

Para podermos aplicar o algoritmo criado na ETAPA 1, será preciso subdividir o arquivo de áudio de maneira que cada palavra ou, idealmente, sílaba, tenha seus tempos de início e fim identificados no arquivo de áudio. Para a avaliação com os cartões, essa identificação será feita manualmente, demarcando na onda do áudio os trechos correspondentes aos tempos de cada sílaba. Se, no entanto, a escala desse trabalho parece aceitável na primeira etapa, ela aqui seria excessiva, e portanto exploraremos maneiras de se automatizar o processo.

Partiremos da abordagem usada na biblioteca *aeneas*, em Python, 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).

Apesar de ser uma abordagem pouco comum para solução do problema de *forced alignment*[9], como explica o desenvolvedor da *aeneas*, Pettarin (2016), a DTW/TTS não depende de um corpus extenso de textos lidos para treinar algum modelo qualquer — algo que, como raro em português em condições análogas às nossas, provavelmente inviabilizaria esta etapa. Ademais, como o arquivo de legendas já contém, tipicamente, os tempos de início de fim de cada frase, o processo de alinhamento poderá ser subdivido e aplicado sequencialmente a cada um dos períodos, de modo que eventuais erros em um não afetem outros. Ainda assim, como a extração dos MFCCS é sensível a ruídos de fundo, para evitar que haja uma etapa de tratamento do áudio muito intensa convém que a escolha do filme considere aqueles onde o diálogo esteja relativamente bem isolado de outros sons.

[9] A lista em <github.com/pettarin/forced-alignment-tools> (acesso em 27 de julho de 2018) compara diversos programas e bibliotecas que lidam com a questão. Tirando a *aeneas*, todas são baseadas em algoritmos estatísticos e redes neurais.

4 · 4 ETAPA 4 — Segunda avaliação: que efeitos terão *closed captions* moduladas de acordo com a expressividade nas falas em um filme?

Com o software adaptado conforme o descrito na etapa anterior, chega-se enfim à etapa na qual pretendemos avaliar que impactos podem ser medidos na fruição de um vídeo quando suas legendas tiverem sido compostas nesta tipografia visualmente modulada pela voz.

Voltando à revisão da literatura em Gernsbacher (2015), são muitas as dimensões sob as quais os efeitos que legendas em filmes têm no espectador foram analisadas por diversos autores. Esses efeitos, em geral positivos, foram encontrados em fatores como compreensão, atenção, memória etc. Considerando, no entanto, que até aqui nos pautamos pela extração (na voz) e representação (na tipografia) de características relacionadas à expressão de emoções, nos interessa nesta etapa uma avaliação focada especialmente nos aspectos da experiência de se assistir a um vídeo legendado que sejam de natureza *subjetiva*, por suposto mais sensíveis às modificações que estamos propondo. Destes, a *imersão* nos parece particularmente promissora[10].

Como discutem Kruger, Doherty e Soto-Sanfiel (2017), a imersão está relacionada ao grau com que um espectador é absorvido por uma realidade ficcional e, como tal, é um estado cognitivo sensível, influenciável tanto por fatores externos — por exemplo, a ação do contexto físico (estar em casa ou em um cinema) ou do contexto social (estar sozinho ou acompanhado) — quanto por fatores internos — por exemplo, se o filme é falado na língua do espectador ou, como é o nosso foco, se está legendado e qual a natureza dessas legendas.

No próprio artigo de Kruger, Doherty e Soto-Sanfiel (2017) foram feitos experimentos para se investigar se surgiam diferenças significativas em medidas de imersão entre participantes que assistiam a um vídeo com ou sem legendas.

Ao contrário do que talvez suporia o senso comum, os resultados indicaram não haver diferenças relevantes na imersão ao se assistir um filme com ou sem legendas. Assim, acreditamos ser possível fazer uma avaliação que compare essas mesmas medidas de imersão entre um vídeo com legendas tradicionais, assistido por um grupo A, e o mesmo vídeo

[10] Todavia, cabe notar que trabalhando sob a mesma hipótese formulada por Kruger et al. (2016) — de que as legendas ajudariam a direcionar o foco do espectador para o diálogo, conduzindo a um entendimento mais aprofundado do mesmo —, caso tenham bom efeito na imersão é plausível supor que as legendas moduladas pela voz poderiam promover efeitos igualmente positivos em outras dimensões.

com legendas que tenham sofrido manipulações como as que estamos propondo[11], assistido por um grupo B. Eventuais diferenças medidas entre os dois grupos, supomos, poderão ser atribuídas aos efeitos das legendas manipuladas.

Escolha do vídeo

Serão dois os requisitos principais para a escolha: (1) o vídeo deverá ser curto, de modo a não comprometer muito tempo dos participantes — ao invés de um longa metragem, fará mais sentido encontrar um curta ou um episódio de algum seriado; (2) a estrutura narrativa deve ser baseada majoritariamente em cenas de diálogo entre os personagens, e não cenas de ação, pois nelas nossas legendas poderão exercer maiores influências.

Adicionalmente, para se simplificar questões relacionadas ao alinhamento entre sílabas faladas e seus correspondentes escritos, será selecionado um vídeo em língua ortograficamente rasa[12] — fortuitamente, o português se enquadra bem nessa categoria, o que não dificultará a procura por voluntários para a avaliação.

Público

Os testes serão aplicados com estudantes universitários — como no teste da ETAPA 2, é importante que sejam leitores fluentes, o que é razoável supor em um ambiente universitário. Como o vídeo terá áudio e legendas em português, ser fluente na língua será também um requisito para participar da avaliação.

Condições para exibição

Será avaliada a possibilidade de, como o fizeram Kruger et al. (2016), exibir o vídeo 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).

[11] Um ponto importante a se ressaltar é que a legenda e o áudio do filme precisam estar na mesma língua para que se possa alinhar o áudio falado com seu equivalente escrito. Essa foi, aliás, a situação construída no artigo de Kruger, Doherty e Soto-Sanfiel (2017), em cima do que sustentamos algumas de nossas suposições.

[12] A medida de "profundidade ortográfica" de uma língua diz respeito a quão próxima está sua ortografia de uma correspondência de uma letra para cada fonema. Uma língua *rasa*, assim, é uma na qual é fácil prever a pronúncia de acordo com a escrita; em uma língua ortograficamente *profunda*, vale o contrário.

Avaliação da imersão

Ao final da exibição do filme, 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 subcomponentes: (1) *transporte*, ou a qualidade de um espectador se perceber transportado para uma realidade fictícia, com consequente suspensão de seu foco na realidade externa imediata; (2) *identificação com as personagens*, uma medida da afinidade que o espectador cria com as personagens, equivalendo a um entendimento empático de seus sentimentos, desafios etc; (3) *realismo percebido*, por fim, mede um 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 (KRUGER et al., 2016).

5 Cronograma

O plano de trabalho está divido nas seguintes etapas:

ETAPA 1 · *FEATURES* E TIPOGRAFIA Finalização de versão do software (já iniciado anteriormente à submissão deste projeto) que aplica a uma *variable font* modulações em seus eixos de acordo com métricas definidas para cada uma das sílabas de um texto.

ETAPA 2 · AVALIAÇÃO 1: *CARD SORTING* A partir dos textos modulados criados na etapa anterior e com participantes de perfil diverso, aplicação da técnica de *card sorting* para descoberta das configurações de tipografia e métricas de áudio nas quais a emoção contida no áudio tem maior inteligibilidade.

ETAPA EXTRA · AVALIAÇÃO ONLINE DE PARÂMETROS DE MODULAÇÃO VISUO-PROSÓDICA Etapa não prevista na primeira versão deste projeto. Cf. maiores detalhes nas seções 8 e 9.

ETAPA 3 · MODULAÇÃO DE LEGENDAS Adaptações no software criado na etapa 1 de modo que possa ser aplicado ao contexto de filmes com *closed captions*.

ETAPA 4 · AVALIAÇÃO 2: LEGENDAS Avaliação de possíveis consequências na fruição de um filme de ter suas legendas moduladas de acordo com elementos da expressividade na fala de seus atores.

ETAPA 5 · ANÁLISE E REDAÇÃO Análise dos resultados obtidos e confecção da dissertação e artigos para divulgação em congressos.

Dividimos a seguir as atividades acima no período previsto para execução do projeto:

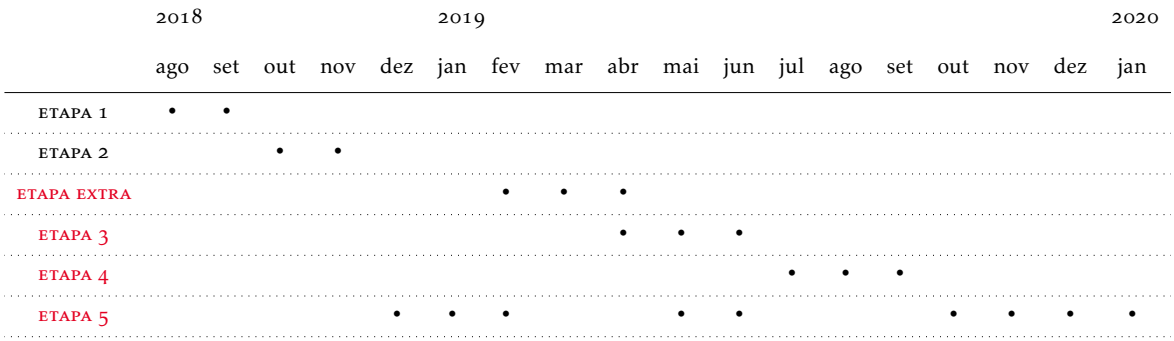

Tabela 1: Cronograma de Atividades *(Em vermelho etapas revisadas)*

6 Resultados esperados

6 · 1 Modelo para tradução de "prosódia em áudio" em "modulação visual de tipografia"

Espera-se obter uma tabela de correspondências entre *features* de um sinal acústico e modulações (ou combinações de modulações) de eixos em famílias tipográficas criadas a partir da especificação OpenType 1.8. Mesmo que criada de acordo com os objetivos de nosso estudo, imagina-se que possa ser expandida futuramente tanto com a exploração de outras *features* de áudio que não estejam necessariamente restritas a elementos da prosódia quanto com a investigação de novas combinações de modulações visuais em eixos tipográficos.

6 · 2 Parâmetros para representação de emoções em tipografia

Associada à tabela citada acima teremos outra, na qual estarão listados parâmetros que, ao juntar *features*, famílias tipográficas e regras de manipulação de seus eixos, produzem representações visuais de emoções contidas na fala que sejam perceptíveis. Como esta tabela terá sido validada empiricamente, acreditamos que além do uso que dela faremos em nosso próprio estudo poderá auxiliar ainda outros projetos de computação afetiva.

6 · 3 Software de extração de features e sua representação na tipografia

Além dos dados e parâmetros gerados, espera-se disponibilizar em repositórios públicos o código fonte que faz a extração das *features* e sua representação enquanto tipografia, ademais já ele próprio baseado em outras bibliotecas abertas.

6 · 4 Software de aplicação em vídeos de legendas moduladas visualmente pela voz

Tratamos este módulo como um software à parte a partir do entendimento que a nossa abordagem específica na escolha de *features*, famílias tipográficas e eixos é apenas uma configuração possível dentre outras em um categoria mais ampla de aplicações construídas à partir da modulação da forma tipográfica baseada em parâmetros acústicos. Portanto, o mesmo software que em nossa pesquisa produzirá legendas de um tipo específico poderia servir de base para outros estudos com abordagens distintas e, assim sendo, será disponibilizado em um repositório distinto do tópico anterior.

6 · 5 Modelo da influência de legendas moduladas visualmente pela voz nos subcomponentes da *imersão* no assistir a um vídeo

A partir da avaliação realizada na ETAPA 4 espera-se obter dados que indiquem quais componentes da *imersão* são mais ou menos afetados pelas legendas construídas conforme nossa abordagem, em cima do que imagina-se ser possível propor tanto estudos que refinem essa mesma abordagem ou, dependendo do que se descobrir, que proponham abordagens outras.

7 Considerações finais

Neste documento descrevemos uma proposta para relacionar texto e fala na tipografia, estabelecendo assim novas possibilidades para a leitura. Como vimos, há na história exemplos variados de como mudanças nas estruturas da escrita precipitaram impactos sociais e culturais profundos. Mesmo que partindo de uma abordagem um tanto mais modesta, acreditamos que nosso projeto atua em um novo campo muito fértil para experimentação. Não excluídas as atribuições da educação, considerando que há diversos contextos em que leitores dos mais variados enfrentam dificuldades com a alfabetização e a leitura fluente — crianças, falantes de línguas estrangeiras, surdos etc —, acreditamos ser possível, através da combinação de conhecimentos e tecnologias de campos distintos como este projeto propõe, desenvolver aplicações que possam ajudar essas pessoas.

Abaixo, adicionamos três novas seções relatando os avanços que ocorreram entre a escrita do projeto, submetido para a Fapesp em meados de setembro de 2018, e seu ponto atual, em meados de março de 2019.

8 Sobre a primeira avaliação

Como previsto no projeto original, em meados de novembro de 2018 realizamos uma avaliação do tipo *Card Sort* com frases visualmente modificadas de acordo com a prosódia na voz de uma atriz que as lera previamente.

O teste foi feito em três etapas: alunos de design da Unicamp, alunos de engenharia da Unicamp e alunos de design da FAAL Limeira. No total, testamos 34 alunos, cada qual ordenando 24 cartões (6 emoções por cada uma das 4 frases escolhidas).

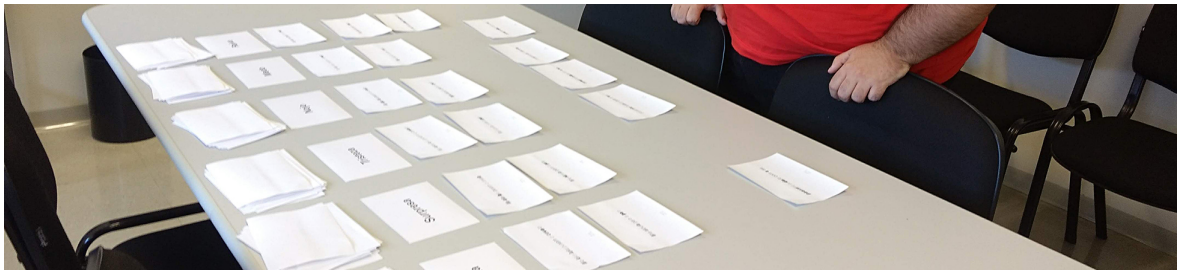

Figura 5: Avaliação na Faculdade de Engenharia Elétrica e de Computação da Unicamp realizada em 13 de novembro de 2018.

Os testes foram feitos individualmente. Para alunos de design, ao final havia uma breve entrevista não-estruturada na qual buscávamos estimular os participantes a compartilhar suas hipóteses sobre quais características tipográficas estariam sendo manipuladas e a quais características acústicas estas poderiam estar relacionadas, além de uma discussão curta sobre quais estratégias eles adotaram para fazer a correspondência entre cartão e emoção.

9 Sobre a segunda avaliação

Os resultados desse primeiro experimento ainda estão em discussão, mas em especial sua parte quantitativa de modo geral nos pareceu muito difusa, motivo pelo qual vimos a necessidade de realizar um novo experimento, não antevisto originalmente e com previsão de aplicação em meados de abril, desta vez em plataforma web.

O foco será muito mais estreito: apresentaremos o áudio da frase lida e duas imagens, uma com modulações tipográficas correspondentes ao som, outra com modulações tipográficas **não** correspondentes. Usaremos duas features — *pitch* e amplitude média[13] —, aplicadas, uma **ou** a outra sempre no eixo *weight*, ou grossura da letra[14].

Imaginamos com isso conseguir medir quão efetiva é nossa proposta de codificação visual da prosódia, mas também índices relativos de performance entre as duas features consideradas e entre cada uma das emoções consideradas — afinal, é possível que certas emoções estejam melhor representadas em nossa abordagem do que outras.

O sistema on-line de avaliação já está pronto. No momento em que escrevemos, aguardamos aprovação de nosso projeto no Comitê de Ética em Pesquisa (CEP) da Unicamp. Em paralelo, está sendo feita a preparação dos dados e textos explicativos, além da lista de pessoas a quem serão enviados os convites para participar — além do sistema da DAC Unicamp, que distribui testes como o nosso para alunos de pós-graduação, convidaremos outros profissionais e alunos do design.

10 Sobre o estágio atual do desenvolvimento do software

Atualmente, estão prontos os scripts Python que fazem a extração e processamento das features acústicas a partir de uma base de arquivos de áudio previamente preparada, assim como os scripts que, em seguida, traduzem essas características em modulações nos eixos de *Variable Fonts*, aplicando-as às frases que geraram os áudios.

As três features consideradas são (1) frequência fundamental, calculada usando o método SWIPE disponibilizado na biblioteca *pysptk*; (2) amplitude média, usando o *root mean square* de cada sílaba; e (3) duração por sílaba, extraída manualmente[15].

Com as features extraídas, o software normaliza os valores a partir de máximos e mínimos considerando o contexto da frase em si e da frase em relação a todas as demais, a partir

[13] Percebemos que, em nossa base de áudio, duração tem uma forte correlação negativa com *pitch* e, portanto, tende a criar padrões visuais que podem gerar à confusão de uma feature com a outra.

[14] De longe o mais citado como percebido nas entrevistas.

[15] Como previsto no projeto, a extração automática será explorada na última etapa através do algoritmo de alinhamento disponibilizado pela biblioteca *aeneas*. Já foi feito um protótipo que, através dela, extraiu os inícios e fins de cada **frase** nos áudios usados na primeira avaliação, mas ainda é preciso refinar a abordagem para descer ao nível da palavra ou, idealmente, sílaba.

Figura 6: Simulação com os 24 cartões, ordenados por frase e emoção (raiva, nojo, medo, alegria, tristeza e surpresa).

do que são gerados os valores a se aplicar nos eixos tipográficos. Estes, assim com as features em seus valores brutos e já processados, são salvos em formato JSON, permitindo futuros usos na web[16], mas também importação no software que escrevemos que gera os PDFs para impressão dos cartões para aplicação do *Card Sort*.

Este último foi escrito em Python no ambiente DrawBot, software muito 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 — em nosso caso, cartões em tamanho A5, cada qual com uma frase.

[16]Como já mencionado, as *Variable Fonts* já estão implementadas nos principais navegadores em uso atualmente.

Referências

BESSEMANS, A. *Expressive typography to improve communication*. 2017. Http://youtu.be/JfsixaAmNOw.

COMMONS, W. *File:Codex Vaticanus Matthew 1,22-2,18.jpg — Wikimedia Commons, the free media repository*. 2018. Acessado em: 28/7/2017. Disponível em: <https://commons.wikimedia.org/w/index.php?title=File:Codex_Vaticanus_Matthew_1,22-2,18.jpg>.

CONSTABLE, P.; JACOBS, M. *OpenType Font Variations Overview*. 2017. <https://docs.microsoft.com/en-us/typography/opentype/spec/otvaroverview>. Acessado em: 4/7/2018.

COSTA, P. D. P. *Two-Dimensional Expressive Speech Animation*. 2015.

GERNSBACHER, M. A. Video Captions Benefit Everyone. *Policy Insights from the Behavioral and Brain Sciences*, v. 2, n. 1, p. 195–202, 10 2015. ISSN 2372-7322, 2372-7330. Disponível em: <http://journals.sagepub.com/doi/10.1177/2372732215602130>.

HARALAMBOUS, Y. Parametrization of PostScript fonts through METAFONT — an alternative to Adobe Multiple Master fonts. *Electronic Publishing*, v. 6, p. 145–157, 1993.

JELINEK, O. *The Spoken Word in Typography*. Dissertação (Mestrado) — FHNW HGK – Visual Communication Institute, The Basel School of Design, 2013.

KOOLAGUDI, S. G.; RAO, K. S. Emotion recognition from speech: a review. *International Journal of Speech Technology*, v. 15, n. 2, 6 2012.

KRUGER, J.-L.; DOHERTY, S.; SOTO-SANFIEL, M. T. Journal article (Paginated), *Original Language Subtitles: Their Effects on the Native and Foreign Viewer*. 2017. Disponível em: <http://eprints.rclis.org/30595/>.

KRUGER, J.-L. et al. Towards a cognitive audiovisual translatology: Subtitles and embodied cognition. In: MUñOZ, R. (Ed.). *Reembedding Translation Process Research*. Amsterdã: John Benjamins, 2016. cap. 8, p. 171–194. ISBN 9789027258748.

KUSTER, M. W. Writing beyond the letter. *Tijdschrift voor Mediageschiedenis*, v. 19, n. 2, 12 2016.

MURPHY-BERMAN, V.; WHOBREY, L. The Impact of Captions On Hearing-Impaired Children's Affective Reactions To Television. *The Journal of Special Education*, v. 17, n. 1, p. 47–62, abr. 1983. ISSN 0022-4669, 1538-4764. Disponível em: <http://journals.sagepub.com/doi/10.1177/002246698301700107>.

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

NUNLIST, R. Users of literature. In: HOSE, M.; SCHENKER, D. (Ed.). *A companion to Greek Literature*. West Sussex: John Wiley & Sons, 2016. cap. 19, p. 296–297. ISBN 78-1-4443-3942-0.

PETTARIN, A. *Inside aeneas Part 1: Motivation And Design Principles*. 2016. <http://www.albertopettarin.it/blog/2016/08/03/inside-aeneas-part-1-motivation-and-design-principles.html>. Acessado em: 27/7/2018.

PETTARIN, A. *Aeneas: How Does This Thing Work?* 2017. <https://github.com/readbeyond/aeneas/blob/master/wiki/HOWITWORKS.md>. Acessado em: 24/7/2018.

RAO, K. S. et al. Characterization of emotions using the dynamics of prosodic features. In: . [S.l.: s.n.], 2010.

SEIDENBERG, M. *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, 2017. Kindle version.

SORANZO, A.; COOKSEY, D. *Testing Taxonomies Beyond Card Sorting*. 2015. <https://www.slideshare.net/atrebla/testing-taxonomies-beyond-card-sorting>. Acessado em: 22/7/2018.

SUKSUMEK, P. *Emotional Type*. Dissertação (Mestrado) — FHNW HGK – Visual Communication Institute, The Basel School of Design, 2017.

TRUSS, L. *Eats, Shoots & Leaves*. Londres: Profile Books, 2003.

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