UNIVERSIDADE ESTADUAL DE CAMPINAS
Faculdade de Engenharia Elétrica e de Computação

Caluã de Lacerda Pataca

Speech-Modulated Typography

Tipografia Modulada pela Fala

Campinas

2021

Caluã de Lacerda Pataca

Speech-Modulated Typography

Tipografia Modulada pela Fala

Dissertation presented to the School of Electrical and Computer Engineering of the University of Campinas in partial fulfillment of the requirements for the degree of Master in Electrical Engineering, in the area of Computer Engineering.

Dissertação apresentada à Faculdade de Engenharia Elétrica e de Computação da Universidade Estadual de Campinas como parte dos requisitos exigidos para a obtenção do título de Mestre em Engenharia Elétrica, na area de Engenharia de Computação.

Orientadora: Profª Drª Paula Dornhofer Paro Costa.

Este trabalho corresponde à versão final da dissertação defendida pelo aluno Caluã de Lacerda Pataca, orientado pela Profª Drª Paula Dornhofer Paro Costa.

Campinas

2021

Ficha Catalográfica
Universidade Estadual de Campinas
Biblioteca da Área de Engenharia e Arquitetura
Rose Meire da Silva – CRB 8/5974

Pataca, Caluã de Lacerda, 1984–
P27s Speech-modulated typography / Caluã de Lacerda Pataca. – Campinas, SP: [s.n.], 2021.

Orientador: Paula Dornhofer Paro Costa.
Dissertação (mestrado) – Universidade Estadual de Campinas,
Faculdade de Engenharia Elétrica e de Computação

1. Tipografia. 2. Comunicação audiovisual. 3. Interação humano-computador.
4. Design centrado no usuário. 5. Interfaces de usuário (Sistema de computador).
6. Prosódia (Linguística). 7. Emoções. I. Costa, Paula Dornhofer Paro, 1978–.
II. Universidade Estadual de Campinas. Faculdade de Engenharia Elétrica
e de Computação. III. Título.

INFORMAÇÕES PARA BIBLIOTECA DIGITAL

Título em outro idioma	Tipografia modulada pela fala
Palavras-chave em inglês	Typography
	Audiovisual communication
	Human-computer interaction
	User-centered system design
	User interfaces (Computer systems)
	Prosody (Linguistics)
	Emotion
Área de concentração	Engenharia de Computação
Titulação	Mestre em Engenharia Elétrica
Banca examinadora	Paula Dornhofer Paro Costa [Orientador]
	Priscila Lena Farias
	Tiago Fernandes Tavares
Data de defesa	08-03-2021
Programa de Pós-Graduação	Engenharia Elétrica

Identificação e informações acadêmicas do(a) aluno(a)

ORCID do autor	https://orcid.org/0000-0001-5046-9884
Currículo Lattes do autor	http://lattes.cnpq.br/9013966339541270

COMISSÃO JULGADORA
DISSERTAÇÃO DE MESTRADO

Candidato: Caluã de Lacerda Pataca **RA:** 15605
Data da defesa: 8 de março de 2021

Título da dissertação
Speech-Modulated Typography
(Tipografia Modulada pela Fala)

Prof.ª Dr.ª Paula Dornhofer Paro Costa (Presidente)
Prof.ª Dr.ª Priscila Lena Farias
Prof. Dr. Tiago Fernandes Tavares

A ata de defesa, com as respectivas assinaturas dos membros da Comissão Julgadora, encontra-se no siga (Sistema de Fluxo de Dissertação/Tese) e na Secretaria de Pós Graduação da Faculdade de Engenharia Elétrica e de Computação.

to Ana,
Blai, and
(*as-of-yet-unnamed*)
child number two.

Acknowledgements

Thank you Paula (for trusting this research was worth it when I had so little going for me), Julio Giacomelli and Fabiana Grassiano (for granting me space and students when I ran my first experiment), Marília and José Eduardo Bretas (for so graciously hosting me when I went to present the results of said experiment), Jan-Louis Kruger (for the questionnaires), Tim Mahrt (for helping me with my `praatIO` code), Alexandra Elbakyan (for), Plínio Barbosa (for helping me dip my toes into prosody), my colleagues at the Multimodal AI group (for helping me with the oh-so-many pitfalls of running an experiment), Cláudio Portugal (for the kind comments), Gabriel Edé (for lending me your voice), Lucas Vignoli Reis and Daniel Sarmento Abrahão (for your friendship and patience through all of this), Daniel, Mônica, Caiame, and Iara (for being there), Ana (for the inspiration, the support, and the love), and, especially, thank you Blai.

Abstract

The human voice is rich in ways the written word struggles to capture. Yes, speech can be transcribed, lifelessly representing in letters what is expressive while in sound. If in many cases this "simplified" translation between mediums is not an issue for readers, typographic modulation of writing is a technological tool able to support those cases where these gaps between how expressive speech is versus how flat typography renders it indeed present challenges to readers. These cases include children struggling to become fluent readers, immigrants unable to discern the written sounds of a foreign language, deaf viewers for whom emotional nuance in movies is lost by how flatly closed-captions render speech, etc. In this context, our work proposes a new model for modulating typography based on the processing of acoustic measures in speech with the timed transcription of its syllables. We use prosodic features such as syllabic magnitude, pitch, and duration to visually modulate typographic attributes such as font-weight, baseline shift, letter-spacing, etc. In this work, we describe how we investigated and developed this audio-visual translation model, presenting the results of three perceptual experiments. In the first, we evaluated a version of the model inspired by a review of similar works from other authors. Unclear results painted the design-space as more complex than originally envisioned, showing a need for dissecting the model in its constituent parts. We did so in experiment #2, where strong preferences emerged for specific matches between acoustic measures and typographic modulations. We used these as inputs for a novel audio-visual translation model, which we evaluated in a third experiment. The new model allowed participants to discriminate between similarly sounding utterances, which they did in 65% of their attempts. This model showed robust enough results for it to be used in the development and research of real-world applications, and its performance can serve as a baseline with which to compare future models.

KEYWORDS

Typography, Affective Computing, Prosody, Human-Computer Interaction, Audiovisual Communication

Resumo

A voz humana é rica em maneiras que a palavra escrita dificilmente consegue capturar. A fala pode ser transcrita mas, tipicamente, a escrita captura de maneira monótona aquilo que, na voz, é expressivo. Ainda que em muitos casos essa tradução "simplificada" entre meios não aparenta trazer em si problemas para leitores, a modulação tipográfica da escrita é uma ferramenta tecnológica capaz de apoiar os casos onde essa lacuna entre expressão na voz e monotonia na tipografia de fato imponha desafios aos leitores. Esses casos são variados, e podem incluir crianças que penam para tornar-se leitoras fluentes, imigrantes que não conseguem discernir os sons escritos em uma língua estrangeira, espectadores surdos para quem as nuances afetivas de um filme se perdem na maneira com que as *closed-captions* achatam as variações vocais, etc. Nesse contexto, o presente trabalho propõe um modelo inédito de modulação da tipografia a partir do processamento do sinal acústico da fala junto à transcrição temporalizada de suas sílabas. Nisso, são usadas medidas prosódicas como magnitude, *pitch*, e duração silábica para modular visualmente atributos tipográficos como espessura de fonte, deslocamento vertical na linha de base, *letter-spacing*, etc. Este documento descreve o processo de investigação e desenvolvimento desse modelo de tradução audiovisual, apresentando os resultados de três experimentos perceptuais. No primeiro experimento, avaliou-se uma versão do modelo inspirada em uma revisão de trabalhos semelhantes de outros autores. Resultados pouco claros apontaram a complexidade do problema e indicaram a necessidade de dissecar o modelo em suas partes constituintes. O segundo experimento explorou as preferências fortes por associações específicas entre certas medidas acústicas e modulações tipográficas. Os resultados desse experimento foram utilizados para desenvolver um novo modelo de tradução audiovisual, que foi avaliado em um terceiro experimento. Nesse experimento, os participantes foram capazes de identificar a expressividade na fala em 65% de suas tentativas. Esse modelo apresentou resultados robustos o suficiente para que possa ser usado no desenvolvimento e pesquisa de aplicações baseadas em cenários reais, e seu forte desempenho pode servir como uma linha de base contra a qual futuros modelos possam ser comparados.

PALAVRAS-CHAVE

Tipografia, Computação Afetiva, Prosódia, Interação Humano-Computador, Comunicação Audiovisual

List of Figures

List of Tables

Contents

Moseby smiled and pointed at the paper. “This paper tells the story of Adam.”

“How can a paper tell a story?”

“It is an art that we Europeans know. When a man speaks, we make marks on the paper. When another man looks at the paper later, he sees the marks and knows what sounds the first man made. In that way the second man can hear what the first man said.”

Jijingi remembered something his father had told him about old Gbegba, who was the most skilled in bushcraft. “Where you or I would see nothing but undisturbed grass, he can see that a leopard had killed a cane rat at that spot and carried it off,” his father said. Gbegba was able to look at the ground and know what had happened even though he had not been present. This art of the Europeans must be similar: those who were skilled in interpreting the marks could hear a story even if they hadn’t been there when it was told.

(...)

The paper version of the story was curiously disappointing. Jijingi remembered that when he had first learned about writing, he’d imagined it would enable him to see a storytelling performance as vividly as if he were there. But writing didn’t do that. When Kokwa told the story, he didn’t merely use words; he used the sound of his voice, the movement of his hands, the light in his eyes. He told you the story with his whole body, and you understood it the same way. None of that was captured on paper; only the bare words could be written down. And reading just the words gave you only a hint of the experience of listening to Kokwa himself, as if one were licking the pot in which okra had been cooked instead of eating the okra itself.

The truth of fact, the truth of feeling. (Chiang, 2019)

Chapter 1

Introduction

You talkin' to me?
You talkin' to me?
You talkin' to me?

Chances are that, as you read these lines, your inner voice inflects each word as Robert de Niro did in his famous *Taxi Driver* monologue, emphasizing a different word at each pass. That we can articulate our voices in many different ways is a mundane observation, but note that despite the many ways we can sound these words, their typographic representation would always be the same.

Sure, we can use italics, bolds, uppercase variants, etc, but typography is, by definition even, composed as a set of letter shapes that remain unchanged throughout a text, as if each glyph was created from the same mold — and, when the printing press was invented, that is exactly what they were. Digital typography does not work the same way as punch cut letters once did but, in the sense that it shapes a flat reading of a text, things have not changed much. Letters shape words, but in so doing give little to no indication of *how* these words could be made to sound.

Not that they necessarily should. Typography is a clear success, especially in its digital incarnations. We have, readily available and at the touch of a finger, sophisticated typesetting engines. They automatically compose text with a finesse that, in a not so distant past, would have demanded whole workshops of skilled artisans to replicate. We casually and without a second thought use these systems daily for communication, information, entertainment, etc. They work, not despite how alphabets simplify spoken language, but maybe precisely because of it.

And yet, a gap exists between how richly expressive speech can be and how flat its textual representation usually is. It is differently felt depending

on the context, and its related issues have inspired many different studies: *Does a lack of prosodic clues hinder learning-to-read children acquire reading fluency?* (Bessemans et al., 2019) *Does the mismatch between the many sounds for the fewer letters in Dutch make reading hard for non-native speakers?* (Verbaenen, 2019) *Do deaf and hard-of-hearing audiences miss out on emotional nuance because closed-caption typography flattens how speech conveys emotions and intentions?* (Murphy-Berman and Whobrey, 1983)

These and other issues stem from the fact that something is left out when we use written text to infer sound (and vice-versa). We can maybe fill these gaps with what we know previously about context, speakers, language, etc, e.g., a competent reader already has a vast repertoire of how phrases and words typically sound like, which they may unconsciously access while reading. Yet sometimes these gaps remain uncrossable, e.g., a deaf person may use closed-captions to read what characters in a movie are saying, but the emotion in their voices is lost.

So, it is in this space between sounds and letters that our research dwells. Approaches to close these gaps vary in objectives and methods but have in common an attempt to augment typography for the cases where this "something" from the speech that is missing in letters is meaningful, and its loss felt.

We are particularly interested in researching the issue from an algorithmic standpoint. This means exploring computer systems that process digital speech and extract from it meaningful bearings with which we can manipulate typography, changing its appearance in such a way as to echo visually what are exclusively acoustic hints.

Balancing the disadvantages and advantages of this computational approach is a relevant challenge in our work. The algorithmic translation of subjectively interpreted complex acoustic signals is of course, a challenging process, as non-verbal communication is hard (if at all possible) to quantify objectively. Conversely, working with digital systems allows us to tackle problem spaces that would otherwise be out of reach, be it because of their scope, scale, or otherwise.

Notably, we think closed-captions would benefit from speech-modulated typography. As has been extensively shown, video captions benefit everyone, with an array of benefits in comprehension, attention, memory, etc, for viewers in different contexts, such as children, adults or elderly, experienced readers or beginners, native or non-native speakers, d/Deaf and Hard-of-hearing audiences (Gernsbacher, 2015).

Particularly for the latter group, closed-captions help to better understand emotional nuance, and we were particularly inspired by a provocation set by Murphy-Berman and Whobrey (1983), around which our research gravitates. Their paper describes an experiment that showed that alphabetized deaf children had a deeper understanding of emotional innuendo in a TV show when it was captioned. In their conclusions, Murphy-Berman and Whobrey imagine a closed-captioning system that would be able to visually translate the "rich tonal information denied [to] deaf children who lack access to the audio soundtrack[, influencing their] interpretation of a program's affective content."

Ours is an exploratory research. Considering algorithmically-based approaches to create speech-modulated typography, it is one of few. Ultimately, we hope it will subsidize the development of applications — such as that futuristic closed-captioning system dreamed up by Murphy-Berman and Whobrey in the 80s — where speech is used to modulate typography so that it becomes a visual embodiment of prosody.

1.1 *Objectives*

This research has as its primary objective the design of a model to map acoustic features of the human voice as visual modulations in typographic shapes. Its specific objectives were:

- Investigating the perceptual interpretations elicited by different mappings of acoustic features and typographic modulations.
- Proposing a model derived from the results of said investigation.
- Implementing a pilot system of said model.
- Evaluating said pilot system.

1.1.1 *Research questions*

Related to the primary objective, we mean to answer the following research questions:

- Can readers of speech-modulated typography recognize prosodic features of its originating audio but *not* present in its textual content?
- What typographic modulations can be generally recognized as visual proxies for the prosodic features of amplitude, pitch, and duration?

1.1.2 *Hypotheses*

The following hypotheses guided our experiments and design choices:

- It is possible to map acoustic features to typographic modulations in such a way that readers will be able to intuitively derive from typography its generating prosody.
- There will be some greater-than-random degree of coherence between readers' interpretations of typographic modulations, i.e., how they decode from the typography its acoustic underpinnings.
- Readers will be able to infer emotions originally present in speech from speech-modulated typography.

1.2 *Publications*

1.2.1 *Peer-reviewed conference papers*

As part of an emerging field of study, our work sits between different areas, drawing from linguistics, graphic design, speech technology, etc. As such, its results found natural fits in conferences with very different focuses. We had papers accepted for two conferences, the first focused on information design, the second on intelligent user-interfaces (of which speech-modulated typography can be an important element). We also presented a poster in a broader-focused computer engineering academic meeting.

- Pataca, C. de L.. and Costa, P. D. P. (2019). Tipografia modulada pela fala: avaliação de um algoritmo de geração de prosódia visual em textos. In *Anais do 9º CIDI | Congresso Internacional de Design da Informação, edição 2019 e do 9º CONGIC | Congresso Nacional de Iniciação Científica em Design da Informação*. Editora Blucher
- de Lacerda Pataca, C. and Costa, P. D. P. (2020). Speech modulated typography: towards an affective representation model. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, pages 139–143

1.2.2 *Conference posters*

- de Lacerda Pataca, C. and Costa, P. D. P. (2019). Tipografia modulada pela fala. In *Proceedings of the 12th edition of the EADCA, Encontro de Alunos e Docentes do DCA*

1.3 *How this work is organized*

The structure of this thesis roughly echoes the chronology for the research itself. Our first experiment, ran at the end of 2018 and described in Chapter 3 (page 36), measured how well (a first version of) our audio-visual translation model allowed participants to infer emotions that were present in speech by only looking at their typographic representations. Results were inconclusive, probably because of a mix between bad choices in the model and the experimental setup, both of which we discuss in-depth in that chapter.

A second experiment followed, ran in 2019 and described in Chapter 4 (page 49). With a narrower focus, we created an experimental setup that measured how participants associated different typographic modulations with speech utterances, classified according to both their labeled emotion and prosodic pattern. Objectively measured results were clearer, and we discuss how those, coupled with what we obtained from the open-ended replies some participants sent, led to some paradigm changes.

The lessons from experiments #1 and #2 led to a refined version of our speech-modulated typography model, whose development and structure we discuss in Chapter 5 (page 60). We go through how we measure and process prosody, annotate sound files, encode the typography modulations and output them, be it as static-images or animated closed-captions.

This refined model was then evaluated in a third experiment, ran at the end of 2020 and described in Chapter 6 (page 74). Here, we measured how well participants managed to infer which instance of speech-modulated typography corresponded with a specific audio utterance. Results were strong, which we discuss, along with some shortcomings of our approach.

Intersecting all discussions, Chapter 2 (page 21) presents some key concepts and literature substantiating our assumptions, including our understanding of how prosody models speech and is related to reading itself, how past researchers have dealt with the gaps between sound and letters and, finally, how the technological breakthrough of variable fonts allowed for our approach.

Finally, in Chapter 7 (page 83) we tie up how well our experiments and reflections answered our research questions, justifying directions for future research projects.

Chapter 2

Literature Review

2.1 *How prosody models speech*

To create *speech-modulated typography*, we need a model of how people expressively modulate their voices when speaking. They do this not only to articulate words but also to express themselves, their feelings, their intentions, etc. Such a model would describe a set of acoustic measurements able to capture the fundamental expressive qualities of the human voice. These measurements, in turn, could be applied graphically to a typeface, making it a visual echo of a modulated voice.

Imagine this: *a person reads out loud a passage of text.* A lot can be gathered from what we hear in their voice. At first, and more directly, there is the explicitly enunciated content: a stream of phonemes and pauses that we can more or less straightforwardly map into words of the text being read. But listen closely and new layers may start to unravel: Does the speaker's accent reveal their place of origin? Can we tell from the sound of their voice their gender, whether they are young or old, small or big, healthy or sick, drunk or sober, calm or agitated, etc? And — as will be the main focus in our work — do the modulations in their speech betray their feelings?

In this imagined scenario, two dimensions can be highlighted: one consisting of *what* is being said, as in the actual spoken words; the other describing *how* those words are being said. While it is possible to consider the two separately, we will show that meaning is inferred and decoded from their interaction (Wilson and Wharton, 2006). This interaction is a subject matter of *Prosody*, the branch of linguistics that investigates how meanings are shaped by the articulated human voice, and from where we hope to develop our human-voice model for *speech-modulated typography*.

2.1.1 *Acoustic and psychoacoustic features*

Prosody typically considers effects at the *suprasegmental* level. These effects are produced when a speaker modulates acoustic characteristics above the elementary speech segments of a language (phones). Suprasegmental modulation typically affects groups of phones and regulates the rhythm, the timing, the meter, and the stress of the words and sentences that we speak.

However, a key challenge in the visual modeling of prosody is that there is no 1-to-1 equivalence between the objective metrics extracted from the speech signal and how the same sound is perceived or experienced by the listener (Barbosa, 2019). While our model will have as its inputs measurable acoustic qualities of sound, if we are to create visual echoes of perceived sound, we have also to consider the psychoacoustic attributes of the human perceptual system.

Considering measurable acoustic parameters of the speech, prosody is related to three main features: *pitch*, *loudness* and *duration*.

2.1.1.1 *Pitch, or intonation*

This is a measure of frequency throughout an utterance. The acoustic-correlate generally used is f_o, or *fundamental frequency*, the perceptually loudest frequency. It is measured in Hertz but can be converted to a logarithmic scale, such as the semitonal, to more closely model how the human auditory system perceives differences in pitch[1] (Barbosa, 2019).

In tone languages — such as Mandarin Chinese, Yoruba, Navajo, Ticuna, etc — specific pitch patterns assume lexical or grammatical functions. In these, and unlike intonation languages such as English or Portuguese, "frequency patterns enter into arbitrary sound-meaning correspondences, and therefore restrict the availability of pitch patterns at the intonational level." (House, 2006)

Perceived *pitch* is influenced by the other two features, namely *loudness* and *duration*. Loud syllables are perceived as if having higher pitched and, inversely, weaker ones are perceived as lower. For the same f_o, a syllable with smaller duration will be perceived as if having a lower pitch than a longer one (Barbosa, 2019).

[1] Barbosa (2019) gives the minimum perceived value, technically referred to as *Just Noticeable Difference* (JND), as 6 Hz for frequencies of up to 1,000 Hz (in this range it is roughly equivalent to one semitone).

As one moves up the scale, however, our capacity to notice changes in frequency reduces, which is why a logarithmic scale, such as is the semitonal, is more appropriate to model the human auditory system

2.1.1.2 *Loudness, or magnitude, or amplitude*

This feature describes patterns of change in sound pressure that correlate to the perceived loudness of a given sound. It is expressed in *decibels*, a logarithmic scale that is conveniently similar to the human hearing response. Note, however, that there are differences in sensitivity to different frequencies, e.g. a signal at a certain dB level with an average 1,000 Hz frequency will be perceived as louder than a signal with the same dB signal with a 50 Hz frequency.

Loudness as an acoustic measure is very sensitive to recording conditions. The distance between speaker and microphone — which can vary naturally during a recording session — can significantly change measured values that are not related to similar changes in vocal effort. For this reason, a *relative loudness* value can be achieved by subtracting the intensity on low and high frequency ranges, giving a more robust measure of intensity.

2.1.1.3 *Duration*

Typically, *duration* will be measured in milliseconds when concerning phonemes or syllables, seconds when larger units. The conventions for determining the boundaries of each phoneme depend on whether it is a vowel — in which case one has to consider the rapid onset of energy and decay thereof in the second formant[2] —, or a consonant, for which there are distinct rules depending on its class (lateral, fricative, occlusive, etc).

Relating to duration, a specific syllable can be perceived as if longer or shorter than those around it, but this perception is dependent on its other acoustic qualities: it is easier to perceive duration in unaccented syllables than it is on tonic ones, and syllables where f_0 varies (a rising tone, for example) are perceived as longer than those where it is constant.

[2] Formants are frequency peaks related to resonances in the human vocal tract.

2.1.2 *Prosody and its functions*

The idea that there is a marked distinction between the *content* of an utterance and its *acoustic form* has been present since the first documented use of the word prosody. In *The Republic*, Plato recounts a discussion in which Socrates defends a "pure narrative" versus one based on "imitation," specially when the latter involved the mimicry of the ways in which certain Greek heroes spoke — whose reputations could thus be tarnished. It followed then that a contrast should be made between *phthongōs* (what is said) and *prosōidía*[3] (how it is said) (Barbosa, 2012).

[3] From πρός (*prós*), meaning "towards," or "close to," and ᾠδή (*ōidḗ*), meaning "song."

This idea of a clear-cut separation between content and form, as if content existed in a pure, abstract form, before its physical manifestation in one of many possible forms, is a notion that will resurface again in our work when we examine its touchpoints with graphic design and typography but, for our present discussion of prosody, a more useful understanding approaches the issue differently.

Relevance theory states that linguistic communication derives its meaning by the dual process of decoding and inference. First, signs and signals have conceptually encoded meanings, which the listener or reader will decode — e.g., the word *cat* brings to mind the concept of the corresponding feline animal. Second, there is a form of *procedural encoding*, where

> a word (or other linguistic expression) encodes information specifically geared to guiding the hearer during the inferential phase of comprehension. The function of such 'procedural' expressions would be to facilitate the identification of the speaker's meaning by narrowing the search space for inferential comprehension, increasing the salience of some hypotheses and eliminating others[.] (Wilson and Wharton, 2006)

Words, by themselves, allow for many different, possibly conflicting, hypotheses as to what meanings and intentions they encode. By incorporating the many different signals encoded in the acts of producing speech, of which prosody is but one, ambiguities can thus be reduced as the listener attempts to close in on one coherent hypothesis of the intended meaning.

2.1.2.1 *Linguistic and paralinguistic prosody*

Considering what information they may encode, these prosodic signals can be divided into *linguistic* and *paralinguistic* dimensions.[4]

The linguistic dimension is associated with changes in prosodic patterns related to the linguist structure of a sentence, including word/syllable boundaries and groupings, sentence modality (e.g. declarative, interrogative and imperative modes), and relative contrast between prosodic units, among other functions.

Paralinguistic prosody deals with expressive qualities of the speaker, be they circumstantial or more perennial. These include aspects such their attitudinal stance (how they relate to what and to who they are talking), indicial aspects (gender, social origin, age, etc), and affective qualities (emotions, moods, etc) (Barbosa, 2012).

Wilson and Wharton (2006) make a distinction between prosodic inputs serving as natural *signs* and natural *signals*: "Signs carry information by providing evidence for it; signals carry information by encoding it." A

[4] While in this document we will use this simpler taxonomy, different authors have grouped paralinguistic aspects under labels such as *extralinguistic*, *non-linguistic* or *(intra) linguistic* (Schötz, 2002).

speaker may thus produce a deliberately affectionate tone of voice (signal) while a simultaneous trembling high-pitch in their voice betrays an angrier disposition lurking beneath (sign). This places prosodic signs in a *showing-meaning continuum* in that it can contain communication artifacts that have both an involuntary and a deliberate dimension.

2.1.2.2 *Prosody and emotion*

There is much controversy in the study of how speech generally, and prosody specifically, encodes emotion. For one, there are no unanimous definitions for *emotion* itself. As a phenomenon, it can be viewed through its "evaluative, physiological, phenomenological, expressive, behavioral, and mental components," which are differently organized by several divergent frameworks and models (Stark and Hoey, 2020).

Hovering beyond such models, Boehner et al. (2005) defines emotion not as an informational layer embedded in speech[5] — something to be extracted from an acoustic signal, but contained within it —, to a culturally grounded, dynamically experienced phonemenon (a lens through which we analyzed some of our results of experiment #2. For more on this, see section 4.4.2, page 58).

[5] Among other signals. Much of this debate is similar to those had when discussing emotions in gestures, facial expressions, posture, etc.

Wilson and Wharton write how prosodic natural signs and signals can only be understood within a given context:

> [T]he effects of prosody are highly context-dependent: prosodic information interacts with information from many other sources during the comprehension process, and the same prosodic input may have different effects on different occasions.
>
> (Wilson and Wharton, 2006)

On the other hand, even if one accepts emotion from within what Boehner et al. call the information model, i.e., the encoding of context-independent emotions in a signal, it is still a matter of debate whether speech encodes specific emotions or only information about the speaker's level of emotional arousal (Bänziger and Scherer, 2005). Even among categorical emotion models, different approaches can include and exclude different emotions (see, for instance, Koolagudi and Rao (2012) for examples of how different speech corpus can have divergent models).

Many studies work from the standpoint of emotion recognition, considering how algorithms or humans can interpret paralinguistic dimensions in speech. For these, Rao et al. (2010) show that the prosodic features we are considering — magnitude, pitch, and duration — are good estimators of expressive speech dynamics, and can thus be used for emotion recognition.

Demonstrating this, Silva et al. (2016) describe an experiment where Swedish and Brazilian listeners were presented with speech samples in Brazilian Portuguese. After listening, they were asked to label each utterance using a set of categorical emotions. The results showed that, even if within the Brazilian group there was a stronger agreement, the Swedish participants — more closely focused on prosody itself since they were not Portuguese speakers — also had a significant agreement about how to label each utterance.

Using a categorical model for emotion, that in their case encompasses anger, joy, fear, or sadness, Moraes and Rilliard (2016) have shown how pitch and rhythm modulations in Brazilian-Portuguese can encode both linguistic and paralinguistic dimensions. Distinct patterns in f_0 and syllabic duration can be used to identify both the modality of an utterance (i.e., whether it is declarative, interrogative, or imperative) and if it represents one of four prototypical emotions.

2.2 *Reading and prosody*

Transcription of speech as written text is a lossy process. Writing systems are imprecise, and they overlook many phonetic and prosodic elements. Although we sometimes think of reading and speech as if "different ways of representing the same thing[,] because of the ways they are perceived and produced, spoken and written language are not simple variants of each other." (Seidenberg, 2017).

Generally speaking, writing systems "work" — people have after all been writing and reading for millennia. Yet, they rest in a delicate tension between complexity and communicative capacity. This balance varies between languages and writing systems, and is not static in time. Indeed, the history of writing tracks countless mutations in conventions — initially calligraphic, eventually typographic — often related to how certain phonetic and prosodic attributes are represented.

Scriptio continua, Latin for *continuous script*, presents an interesting example in western culture of how there is an interplay between writing conventions and how text is read. In it, words were not separated by spaces, which were only introduced around the 7th century. As such, written text was then a dense block of letters (See example in Figure 2.2). From this, some historians advance the theory that in classical antiquity reading was a predominantly *acoustic* phenomenon. Reading, then, was done out loud:

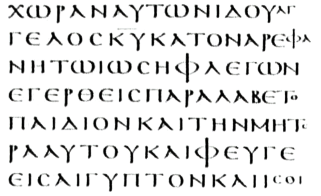


Fig. 2.1: Excerpt from the Codex Vaticanus, dating from the 4th century, an example of *scriptio continua* (Wikimedia Commons contributors, 2018).

> The common practice of listening to literature no doubt enabled ancient audiences to notice, appreciate and evaluate all those phenomena that are much better caught by the ear than by the eye[.] (Nünlist, 2016)

For Greeks then, learning to read a text with no marks for pauses, prosody, types of literary rhetoric, etc, involved not only learning how to decode sounds from the writing but also memorizing the rules for delivering it.

Regardless of silent-reading polemics[6] Küster (2016) writes that space as a word separator was invented between the 7th and 8th century in Ireland. Latin, in which religious texts were written, was there and then a foreign tongue. Parsing *scription continua* text depended on a previous command whichever language it codified, which was not the case for a typical reader of Latin in medieval Ireland.

Another important technical development was the standardization of modern punctuation conventions between the 15th and 16th centuries by the grandson of Aldus Manutius, Aldus Manutius the Younger. These, along with spaces, established a reading culture where *silent reading* was the norm — if not actually in practice, at least in how reading is idealized. This understanding of *progress* from oral reading to one that takes places exclusively in the mind carries a notion of progress with an implicit narrative of mind's conquest over matter (McCutcheon, 2015).

McCutcheon might have a point. A prejudice against supposedly-inferior voiced forms of readings can still be seen, for example, in speed-reading. Its proponents sell the idea that a reader that learns to suppress subvocalization will gain speed with no losses in comprehension as if an exclusively visual process would be more efficient. This, however, ignores the neurological basis for reading.

In a landmark study, Van Orden (1987) presents an experiment where skilled readers would be routinely fooled by homophonic words in a categorization exercise. When asked if a certain word belonged to the *flower* category, for instance, they would correctly discard ROBS, but not ROWS — a homophone of ROSE. This indicated that they were using a word's sound to access its meaning, and not (only) its visual shape.

Seidenberg (2017) writes that in fluent reading there are both ortographic and phonological neurological pathways involved. In fact, competent reading resolves information by considering in its decoding strategies many subtle and even contradictory signals — similary to how, as we saw in speech, there is the dual process of decoding and inference, as described by Relevance theory.

[6] McCutcheon (2015), strongly contests that *scriptio continua* necessarily equated to text exclusively read out loud: "There was actually no technological reason that prevented individuals in antiquity from silently perusing a text. (…) [Acknowledging that individuals in antiquity commonly read silently] would undermine the scholarly belief that the history of human cognition can be linked causally to the development of information technology."

2.2.1 *Phonetic and prosodic representation systems*

The phoneme is not a natural unit. Rather, it is best thought of as an abstraction that we learn to recognize and divide in speech. Importantly, phonetic awareness is only acquired when we learn to read (and, as such, is not available to the illiterate) (Morais et al., 1979).

Writing systems like the modern Latin alphabet codify phonemes through letters, and some elements of prosody through special marks. These are many and varied. Some are common across languages that use the Latin alphabet, e.g., question (?) and exclamation (!) marks, commas (,), periods (.), dashes (—), etc. Others are language specific, e.g., the inverted question (¿) and exclamation (¡) marks, used in Spanish. Lastly, there are some rare symbols, such as the interrobang (⁉), irony point (؟), etc — these are interesting, but their use is limited by lack of support in many typefaces and general recognition.

These symbols help a speaker meaningfully inflect his voice when reading aloud, but they work through convention rather than through specific directions. For instance, we know a comma is a shorter pause than a period but the exact ratio between the two is left to the speaker's discretion. Analogously, an exclamation mark gives a phrase an interrogative modality, and as we have seen these are indicated by stereotypical pitch patterns. Again Seidenberg:

> Spaces indicate the boundaries between words in alphabetic writing systems; there are no "spaces" between words in fluent speech. Typographical conventions such as using capitalization to indicate proper nouns do not have spoken equivalents in English. (Seidenberg, 2017)

Scripts are constantly changing. With time, their phonetic and prosodic gaps can increase or decrease. These vary depending on the writing system, and even when a writing system's "gaps" are not an issue when considering the script's typical day-to-day use, there are situations and groups of people for which this may not be true. This can be the case, for instance, when considering groups such as illiterate children, non-native speakers, or people with disabilities.

Written Hebrew, for instance, is typically only written with consonants, but children use a simplified — in the sense that there is a greater match between spoken phonemes and written graphemes — form which also includes vowels.

2.2.1.1 *Special alphabets and typefaces*

Sometimes these phoneme-grapheme gaps are embedded into the script. There have been many historic attempts to reform alphabets and their typographic representations to make reading more efficient, easier, or generally accessible by tackling these gaps and mismatches.

This is a difficult challenge, especially considering how it will be faced against naturally evolving, stable scripts that generally work in *most* cases. Indeed, an all-encompassing phonetic writing script already exists. While comprehensive, the International Phonetic Alphabet is overly complex — per its 2005 update, it has 107 letters, 52 diacritics and four prosodic marks —, restricting its use to specialized cases.

Considering the Dutch language, for example, which crams up to 48 phonemes into 26 letters, Verbaenen (2019) has explored how a typeface could represent differently the same letter to match the many different sounds that can be associated with it, as seen in the example in Figure 2.2. Knowing, for instance, when a vowel is pronounced in its open or closed, short or long variant depends on a previous knowledge of the Dutch language that a non-native speaker might lack. By explicitly illustrating these differences, Verbaenen argues that the Phonotype typeface could help non-native speakers parse written text.

Fig. 2.2: Modified version of the Times New Roman font, called Phonotype, as created by Verbaenen (2019). It visually differentiates how the same letters can be used to represent different phonemes of Dutch. (Reproduced with permission from the author.)

2.2.1.2 *Special typographic systems*

Some designers and researchers have tackled with expanded and prosodic representation through visual signs that are overlaid, or modify, an existing writing system.

Working within the confines of traditional orthography, dos Reis (2014) developed a notation system called *Speechant*, shown in Figure 2.3. It annotates vowels for their pitch and duration qualities, attributes that Portuguese speakers struggle with when learning English. Language students

who used this system produced better sounding readings of English utterances, when compared to a control group, according to the analysis of trained phoneticists.

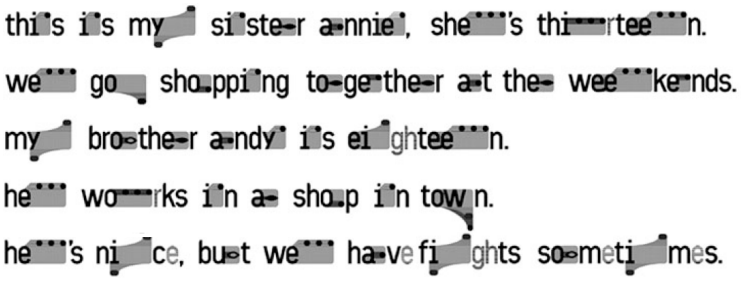


Fig. 2.3: Example of the *Speechant* prosodic annotation system, adapted from dos Reis and Hazan (2011). (Reproduced with permission from the Oxford University Press.)

2.3 *Modulated typography*

Some research projects attempt to create prosodic notation systems through what we are calling *modulated typography*. In them, a regular alphabet will have its glyphs[7] and typographic compositional parameters changed in tandem with prosodic modulations. These modulations can either serve to direct how a reader's tone-of-voice should be modulated as they read aloud a text, or respond to an existing audio recording — possibly through computational processes, as is the case of our research.

[7] A glyph is the specific shape of a character for a given typeface. In this thesis, for example, the same "Q" character is differently represented as three different glyphs from three typefaces: Q, Q, and Q.

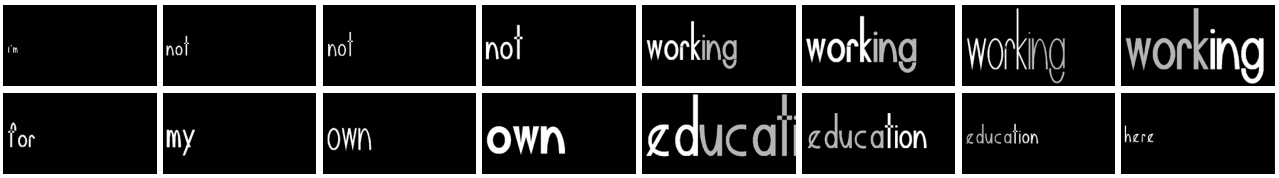


Fig. 2.4: Screenshots of animated prosodic font, adapted from Rosenberger and MacNeil (1999).

Rosenberger and MacNeil (1999) developed a system, shown in Figure 2.4 for an animated prosodic font where each glyph is composed as a set of primitive shapes that can be modulated according to measured prosodic features. Loudness was mapped to text size, and the pitch was mapped to an increase in height and decrease in width, for high pitched tones and an inverse decrease in height and increase in width for lower-pitched tones. The rhythm was represented by when the letters were shown on screen. So that each phoneme could be represented by a single glyph, there were some "phonetic ligatures" to represent sounds — such as /ð/ (*as in **th**y*), /ʃ/ (*as in me**sh**er*), and /ŋ/ (*as in thi**ng***) — as a single glyph. In an experiment, 12 out of 14 participants were able to use this prosodic font to identify correctly which audio the modulated font came from.

Rosenberger-Shankar's work carries similarities with Wölfel et al. (2015)'s Voice Driven Type Design (VDTD). In the latter as in the former, the prosodic features of each phoneme are mapped into a system that *rebuilds* a mathematically modeled typeface, allowing for its graphical attributes to echo the dynamics of sound. In VDTD, these specific visual parameters were *vertical stroke weight*, mapped to loudness, *horizontal stroke weight*, mapped to pitch, and *character width*, mapped to rhythm. A schematic of how these three dimensions change is shown in Figure 2.5. In an evaluation, participants had good recognition of changes in loudness and pitch, and not so much for rhythm.

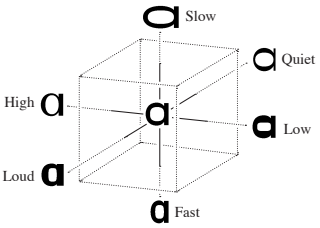


Fig. 2.5: How each of the three modulated visual attributes change in a VDTD typeface. (Reproduced with permission from the authors.)

Castro et al. (2019a)'s *Máquina de Ouver*, explored the relation of the human voice with the written text through typographic modulations that echoed prosodic inputs measured from recordings. Pauses between words and syllables were represented as smaller or longer white spaces, pitch as letter size, intensity as font-weight, as shown in Figure 2.6. These were created by scripts in Adobe InDesign, for static images, coupled with scripts in Adobe After Effects, for generating the animated versions.

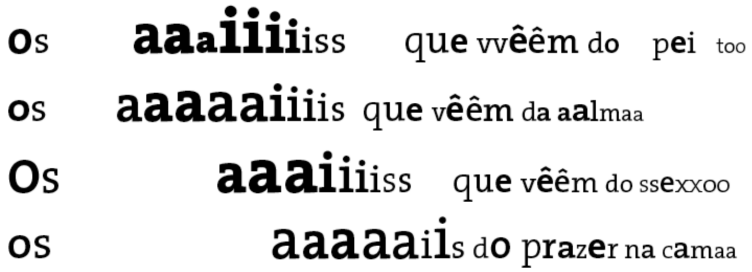


Fig. 2.6: Modulated typography in a poetry excerpt from Máquina de Ouver. Extracted from Castro et al. (2019a)

Bessemans et al. (2019)'s approach deals with the fact that a poor grasp of prosody when reading is associated with poor literacy outcomes in children.[8] Through typographic modulations, they attempt to stimulate a more richly expressive reading, hoping to uncover whether young readers will be able to follow the modulations with their voices. Loudness and duration were mapped to, respectively, font-weight and letter width. Pitch, "the most challenging aspect of prosody [to visually describe]," was represented either through shifts in the vertical position of syllables or through a vertical stretching of their shapes. Results indicated that children could, when instructed, follow all visual cues.

[8] Echoing Paige et al. (2017): "During silent reading readers generate mental prosodic-phonological written text representations[, auditioning] different prosodic interpretations until one is found that seems to be a best fit for meaning."

A summary of the similarities and differences between the previously presented research projects is shown in Table 2.1. It also anticipates our model, which will be presented in detail in Chapter 5, on page 60.

Research	Input	Modulation technique	Typographic mappings (sound → typography)	Rate of change	Proposed applications
Bessemans et al. (2019)	Writer's intents	Manually designed text	(1) Magnitude → Font-weight; (2) Pitch → Letter height and Baseline shift; (3) Duration → Letter width	Discrete	Literacy education
Verbaenen (2019)	Previous knowledge of language	Custom typeface	Additional set of glyphs to discriminate different phonemes represented by the same letter	Discrete	Learning of foreign-languages
dos Reis (2014)	Previous knowledge of a language	Manually designed text	(1) Duration → Syllable spacing; (2) Notation system overlay to represent pitch contour	Discrete	Learning of foreign-languages
Rosenberger and MacNeil (1999)	Recorded audio	Algorithmically-modelled typeface	(1) Magnitude → Font-size; (2) Pitch → Vertical and horizontal stretching	Continuous	Expressive text-messaging
Wölfel et al. (2015)	Recorded audio	Algorithmically-modelled typeface	(1) Magnitude → Vertical stroke thickness; (2) Pitch → Horizontal stroke thickness; (3) Duration → Letter width	Continuous	Language learning, subtitles, texting, etc
Castro et al. (2019a)	Recorded audio	Algorithmical modulation of typographic parameters	(1) Magnitude → Font-weight; (2) Pitch → Font-size; (3) Pauses between words → Word spacing; (4) Duration → Letter repetition	Discrete for (1) and (4), continuous for (2) and (3)	Design of poster and video artifacts
Our approach (final model)	Recorded audio	Algorithmical modulation of typographic parameters	(1) Magnitude → Font-weight; (2) Pitch → Baseline shift; (3) Duration → Tracking and changes in color	Continuous	Speech-modulated closed-captions

Table 2.1: Summary of how different authors approached closing the gaps between speech and text, for a myriad of different goals.

2.4 *Variable fonts*

An important technical breakthrough occurred in 2016, laying the groundwork for our approach to speech-modulated typography. Announced at the 2016 ATypI meeting at Warsaw by representatives from Adobe, Apple, Google, and Microsoft (Lemon et al., 2016), OpenType version 1.8 introduced the possibility of font variations.

In a (simplified view of a) typical font file, a type designer would create several glyphs, one for each letter, letter plus diacritic, punctuation, ligatures, etc. Importantly, one file would contain only one variation of this

typeface. So, for example, to represent regular and bold font-weights, the typographer would need to create two separate files. If they wanted to add italics, this typeface would now need four separate files: regular, regular-italic, bold, bold-italic.

This meant that complex type families would inevitably fall into combinatorial explosions — as an extreme example, the *TheSans* type family, which includes variations in font-weight and letter width, has 65 distinct files. This is complex from the point of view of the type designer, but also for the designer that will use these multiple font files, and also for the end-user: as custom fonts became common in web design, optimizing the bandwidth cost of a site using complex typographic combinations became an important issue.

Variable fonts have an internal structure that is fundamentally different from a traditional font file. Instead of considering each glyph as a static vector image of a letter, this same image is coupled with *variation axes* (Jacobs and Constable, 2018). Each variation axis defines how each vector point in a given glyph will shift its position as the value for that axis changes. In other words, instead of saving various different drawings for a glyph, you may have one base drawing and a series of *deltas* for the positions of each point.

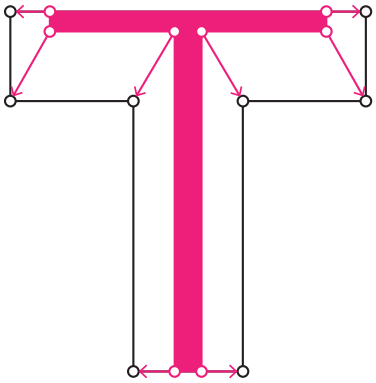


Fig. 2.7: Schematic example of how the vector of a *T* glyph will shift from a thin font-weight, in magenta, to a heavy font-weight, in black outlines. Note that the distance each point shifts is not constant, so this is not a simple mathematical transformation but, rather, a deliberate process that the type designer controls.

A common axis, for instance, is font-weight. It informs how each glyph changes as it morphs between a thin and a heavy variant (Figure 2.7).

A variable font can have any number of variation axes, and when a type shaping engine calculates the final output of a glyph it will consider the values for all axes, with which it will interpolate the values of all deltas. This way, instead of a discrete set of glyph shapes there is a continuous design space where the shapes move between as many dimensions as a type designer wants (Figure 2.8).

In terms of support, and maybe because of the coordinated involvement of major operating system and browser developers, Variable Fonts were

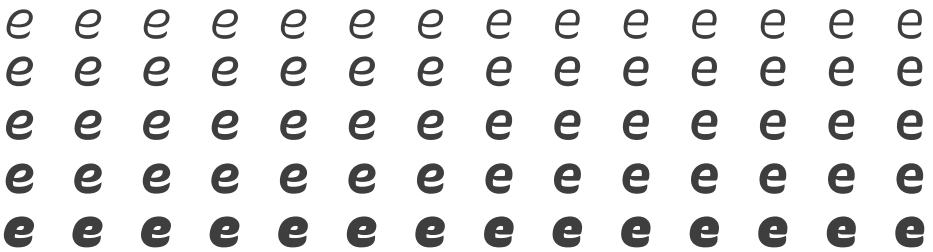


Fig. 2.8: Example of how a glyph can change arbitrarily in two dimensions: font-weight, on the y-axis, and slant, on the x-axis.

quickly adopted and are available in all major browsers (Deveria, 2021) and operating systems (Sherman, 2020).

Although a variable-font file will be typically larger than one single traditional font-file, it allows for the creation of innumerable variations that would necessitate many separate files, so the overall size tends to be smaller[9] even in relatively simple scenarios.

Bandwidth savings are a first and obvious advantage, but considering how text is an important building block of user interface design, the design-space approach to typefaces opens many other possibilities. These can involve allowing for greater visual finesse and expressiveness, animations, our approach for speech-modulated typography, among others. The designer Andrew Johnson, for example, has explored how typefaces in Augmented and Mixed Reality applications could respond to different inputs, such as distance, viewing angle, ambient lighting, etc (Johnson, 2019).

[9] In their initial presentation, Lemon et al. (2016) gave an example of how in a typical scenario with 5 font-files could have the sizes fall from 657 Kb to 199 Kb.

2.5 *Concluding remarks*

In this chapter, we presented key concepts related to prosody, over which much of our approach rests. We have also presented a panorama of research that tackles the gaps between how rich speech is and how flat typography typically portrays it.

These are varied both in terms of their approaches to how they modulate (or construct) glyph shapes and in their goals in doing so. We attempt to bridge the gap between prosody and typography, similarly to Bessemans et al. (2019) and not, unlike Verbaenen (2019), phonetic gaps. Also, this is done exclusively through typographic modulations, unlike the additional graphical vocabulary explored by dos Reis (2014).

Our work shares with Rosenberger-Shankar (1998), Wölfel et al. (2015), and Castro et al. (2019a) an algorithmic approach to generate typography. Unlike the first two, however, we do so using regular, off-the-shelf typefaces, leveraging the newfound possibilities granted by variation axes.

This is an important distinction because it potentially allows for easier integration of speech-modulated typography and commonly available systems — so much that we managed to run our third experiment's typographical framework in a common web-browser.

This emphasis on off-the-shelf typefaces is similar to Castro et al. (2019a)'s approach, as our work also built on the algorithmic processing of prosodic features and their use as input into typographic modulations. Some important differences lie in our use of variable fonts, where Castro et al. uses traditional, pre-OpenType 1.8 fonts, our use of WebVTT's class syntax to encode typographic modulations and animations, and in a potentially easier integration with existing typographic frameworks. More subjectively, Castro et al. (2019a)'s evaluation is similar to ours in that it asks participants to use speech-modulated typography to differentiate between utterances, but in their case these utterances have highly divergent prosodic patterns, making identification potentially easier.

Chapter 3

Can we infer emotions from reading speech-modulated typography?

In the first step of our methodology, we proposed a seminal translation model of expressive speech into modulated typography, based on assumptions derived from linguistics and design research literature. Following, we assessed if participants exposed to printed a number of phrases where the typography was modulated by expressive speech would be able to uncover the emotional labels in the original speech utterances.

The subjective evaluation was an attempt to answer whether readers of speech-modulated typography can infer from it emotions present in its originating audio but *not* its textual content. Indirectly, this could serve as evidence of our proposed speech-modulated typography model's effectiveness in capturing and expressing prosody, presumably related to the strength of the first question's answer, in turn helping answer our first research question, presented in Section 1.1.1.[1]

In this chapter, we will discuss the first version of our prosody-typography translation model. Also, we describe our evaluation, its results, and how we interpreted them. The results presented in this chapter are also discussed in Pataca and Costa (2019).

[1] Can readers of speech-modulated typography recognize prosodic features of its originating audio but *not* present in its textual content?

3.1 *A first version of our speech-modulated typography model*

In this section, we describe the inner workings of the first version of our speech-modulated typography model. As we developed it before we ran any of the experiments, we based its choices and assumptions on research literature from linguistics and design.

3.1.1 *Extracting prosody*

To generate speech-modulated typography, first, we needed to extract and process certain key features of audio. We used three key acoustic measurements to model prosody: *magnitude, pitch,* and *duration*. They were useful in our approach for two reasons. First, as previously discussed (Section 2.1), they communicate paralinguistic dimensions such as intentions, emotions, moods, etc. If successfully represented graphically, these dimensions could augment typography, infusing it with expressive qualities in the voice.

Secondly, in prosody these dimensions are meaningfully captured when considering an utterance at its local level (i.e., each syllable), as opposed to its global level (i.e., the whole utterance) (Rao et al., 2010). This is worth highlighting because we were interested in features not only because of their ability to capture subjective meanings enclosed in the voice, but also because they did so even when considering units in sound that had clear parallels to units in text. We posited that prosody that is meaningful when considering values for each syllable could be meaningfully represented in each syllable in printed text.

Thus, while there are some approaches in speech emotion recognition that use features considering time frames smaller or bigger than the syllable,[2] we opted for a prosodic-based algorithm because it allowed for a direct relationship between extracted acoustic measurement and its representation in text.

We used three prosodic features. First, *magnitude*, defined in decibels as the Root Mean Square[3] of a syllable's audio frame,[4] as calculated by the RMSE method of the Python *librosa* library.

A caveat with RMS is that it does not model how loudness is perceived psychoacoustically. In other words, while it is an accurate measure of the acoustic energy carried in a sound wave, there may be differences between its values and how loud a sound is *perceived*. This is because our ears are differently sensible to sound in different frequencies, so the spectral properties of sound matter. Yet, while there are more complete, albeit more complex, measures of loudness, in this instance we opted for simplicity rather than precision.

The second prosodic feature used was the *fundamental frequency*, or f_0, a correlate of perceived *pitch*. Here, it was measured in Hertz as the average frequency per syllable as extracted by the SWIPE method of the Python *pysptk* library, set up with frequency range between 75 and 600 Hz.[5]

[2] There are some, for instance, that correlate affective meaning to the results of the statistical analysis of data points taken from the *whole* utterance. (El Ayadi et al., 2011)

[3] The square root of the arithmetic mean of the squares of an audio time series, per the formula $\sqrt{\frac{1}{n}\left(x_1^2 + x_2^2 + \cdots + x_n^2\right)}$.

[4] An array of samples containing magnitude information at each instant in a given time range.

[5] This corresponds to a typical frequency range for higher pitched voices, as was the case of the female-voiced phrases of the speech corpus we used.

In Portuguese, as in other Latin-based languages, sounds are classified in many dimensions, one of which is divided into three classes: voiced, fricative or plosive (Hernández, 2016). Voiced sounds occur when there is a periodic vibration of the vocal folds, and only in them can pitch be reliably detected. Because we were interested in obtaining an average[6] value for each syllable's f_o value, we had to deal with audio frames where, because of a predominance of fricative or plosive sounds, there was no detectable pitch. For these, we ran a linear interpolation that considered the whole phrase, meaning that missing values in a given syllable could be derived from its neighboring syllables.

Lastly, the third prosodic feature measured was *duration*. This is the more straightforward span of time of each syllable, measured in milliseconds as the difference between their end and start timestamps.

[6] Again, a practicality-driven simplification, given that it is not uncommon to discuss a pitch contour in terms of its behavior throughout a syllable. Barbosa (2019), for instance, discusses the different perceptual effects caused by ascending or descending pitches.

3.1.1.1 *Syllable segmentation*

While there are approaches that promise to automatically define the boundaries of each phoneme and/or syllable in a given utterance,[7] because of the small number of short audio files used in the experiment, it was decided that a more pragmatic approach would be to define the boundaries of each syllable manually. This was done in the Adobe Audition software, from where one can analyze a sound file both from its magnitude contour and spectrogram. From these, it was possible to find the boundaries of each phoneme. It helped that the recordings were done in a noise-free setting by a clearly-articulating professional actress — not a given in any situation.

[7] For a recent example of one such approach with high accuracy see Ramteke and Koolagudi (2019) Alternatively, there are open source libraries that implement these functionalities, such as the *vosk* toolkit, and segmentation options are available in commercial cloud-based services such as Google's Speech-to-text API.

3.1.1.2 *Feature normalization*

Acoustic measurements of each prosodic feature were normalized. To achieve this, we considered the whole set of audio recordings in each of the four phrases used — the corpus is presented in more detail later, but for now it is sufficient to say that there were six versions for each of four phrases. The algorithm considered maximum and minimum values for each of the three prosodic features, grouped by phrase, which it then used to rescale each feature's individual value, as follows:

$$z_i = \frac{x_i - \min(x)}{\max(x) - \min(x)},$$

where x_i is the i^{th} value of the x feature, and z_i is its i^{th} normalized value.

This formula converted all measurements to values in the range between zero and one. These were easy-to-use inputs for the variable fonts, where

each axis defines its own scale, but, more importantly, the normalization made the phrases comparable: the loudest loud and quietest quiet would have, respectively, *1* and *0* values, which would be consistent throughout the phrases and independently of their local maximums and minimums.

3.1.2 *Representing prosody typographically*

Our first translation modeling approach consisted of visually representing prosodic features through the mapping of the normalized acoustic measurements as input to typographic axes in variable fonts. The question we posed ourselves was: *what axes should map each prosodic feature?*

Following Bessemans et al. (2019), we mapped the *magnitude* and *duration* features to the typographic axes of, respectively, font-weight and letter width (e.g., Figure 3.1).

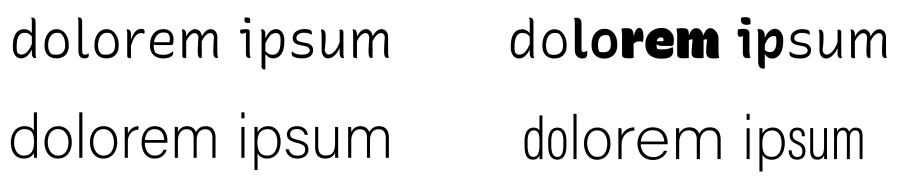


Fig. 3.1: Examples of changes in font-weight in the *Recursive* typeface and letter width in the *Acumin* typeface.

To represent *pitch*, however, we felt that the two alternative representation schemes proposed by Bessemans had unresolved issues. While the first solution of raising or lowering syllables relative to the baseline (e.g., Figure 3.2) could be made to work, it would impose challenges when speech-modulated typography was used in broader contexts, e.g., closed-captions, books, web-pages, etc.[8] In settings where the leading is tightly set, raising and lowering syllables could create juxtapositions between the lines.

[8] While this argument against a baseline shift still rings true, we eventually reconsidered it in the following models.

Fig. 3.2: Example of shifts in baseline in the *Public Sans* typeface. Note that this effect is not made with changes in a variable font axis, and could be applied thus to any font.

Their second approach consisted of creating a "font-height" axis to control the vertical extension of glyphs (e.g., Figure 3.3). While it remains to be seen if participants in our experiments would associate this attribute as a proxy for pitch, we could not find off-the-shelf fonts in which such an axis was defined, and this vertical shifting is too sophisticated to do programmatically, so this solution too was deemed inadequate.

The approach used by Wölfel et al. (2015) to represent pitch, which consisted of correlating a glyph's horizontal strokes' width to changes in pitch,

Fig. 3.3: Example of changes in the stretching of a glyphs' height in the *Dim* typeface. This example was done by manually editing the font's vectors.

could also not be adapted to our case — while this effect is rare, it is not unheard off,[9] but it too was not found as a variable-font axis in any off-the-shelf font. An example of reverse-contrast typography can be seen in Figure 3.4.

[9] Typefaces where horizontal lines are generally thicker than vertical are called "reverse-contrast."

Fig. 3.4: Reverse-contrast "Italian" type in an 1828 specimen book by the George Bruce company of New York. (Blythwood, 2015)

To represent pitch we ended up selecting the *slant* axis, which controls how inclined each glyph is — similar to traditional *italics*, although, unlike them, slanted fonts' glyph-shapes share the same structures as their unslanted versions. Although we found no mentions of previous, similar uses, we considered them a reasonable choice because of references about their use to suggest *brisk* and *energetic* movements, e.g., in 1919 the futurist Marinetti suggested italics denote "swift sensations." (Letters from The Temporary State, 2019).

An example of slant modulations can be seen in Figure 3.5.

Fig. 3.5: Example of changes in slant axis in the *Compressa* typeface.

3.2 *Method*

Experiment #1 attempted to measure how consistent participants' responses to our prosody-typography translation model would be. To achieve this, we designed a subjective evaluation where we could compare participants' responses to our speech-modulated typography to what we knew before hand would be their expected response if they were exposed to the typography's originating audio.

3.2.1 *The speech corpus*

At the time, we approached the issue of measurement from an Affective Computing standpoint. We were interested in using speech corpora in which utterances were labeled in terms of their perceived emotions. We posited that if participants' labeled the typographic instances in the same way as the audio, we would have evidence that, much like our model

translates prosody into typography, participants' were managing to decode from this typography its underlying prosody.

We used a speech corpus created and described by Costa (2015). In it, a set of Portuguese phrases is repeatedly read by actors and actresses, in each round instructed to attempt to depict in their voices one of the "Big Six" emotions,[10] as described by Ekman (1970), along a *neutral* variant. There was also an instruction to depict each emotion through different levels of extroversion, namely, shy, polite and extroverted.

[10] Anger, Disgust, Happiness, Fear, Sadness, and Surprise.

Considering our use, this corpus had a positive characteristic: the phrases, although in coherent, grammatically sound Portuguese, were constructed in such a way as to be somewhat, even if not overly so, nonsensical. This suited our experiment well, since we were interested in a decoding of emotion derived by speech and typographic expressiveness, and *not* by the particular semantics in any given phrase.

Based on their short length compared to the other phrases on the corpus, we selected the following four phrases: *Filha, rúcula para a pata* (in English, Daughter, arucula for the duck); *Passarinho, cuidado com a asa* (in English, Birdie, watch your wing); *Lilo, Kika, Luku, puxem o cavalo* (in English, Lilo, Kika, Luku, pull the horse); *Você tem certeza disso?* (in English, Are you sure of that?).

Fig. 3.6: An example of speech-modulated typography, as used in the test. This particular phrase was read with a *surprised* voice.

Although the phrases were read by many different actors and actress, to minimize variation between typographic instances we chose to generate the speech-modulated typography using only the versions read by one of the actresses, always using their versions read aloud representing the *polite* activation level. See Figure 3.6 for an example of the final typographic output, and appendix A, on page 94, for all typographic instances.

3.2.2 *Card-sorting*

Having selected the phrases, we processed each audio file and generated their speech-modulated typography versions. These were printed as A5-sized paper cards, 24 in total: 4 phrases, each representing one emotion.

These cards were to be used in card-sorting sessions. The method has some variations, but in the one we chose there are a number of cards denoting, for example, a set of objects, concepts, places, etc. The participants

are then asked to assign each card to a category, which can be previously defined or open to suggestions.

Card-sorting sessions are usually done to uncover a given audience' mental models about a certain topic,[11] but in our experiment it was deemed an useful way to uncover if the set of printed cards with speech-modulated typography would be classified similarly to their originating audios' represented emotions.

[11] This is useful, for example, in UX design, when one is creating the information architecture of a web site and wants it to echo how the site's users' expectations. It can uncover, for instance, that a given number of participants believe a *contact us* page belongs in a *help* section.

These categories, in our case, were the Big Six emotions, which were expressed in the ways the actress read aloud each of the utterances. Since we knew beforehand this originating emotion, the test would measure how frequently participants correctly matched each card to this original emotion, an indication that they were able to have a sense of the audio behind the typography even without having listened to it.

To analyze the results, we planned on using a metric called *edit distance*. As presented in Nawaz (2012), it measures the distance between two given card sorts. In our case, we had an implicit *ideal* card sort where each card would match the emotion of its originating audio. Therefore, calculating the edit distance between each card sort and this ideal configuration would give us a measure that was inversely correlated to the efficiency of our prosodic-typographic model.

Additionally, through Monte Carlo simulations, it is possible to estimate what the average edit distances would be if our model was *totally* inefficient, i.e., in case participants' card sorts were indistinguishable from a random shuffle of the cards.

To calculate a given card sort's edit distance, we have to measure how many operations are necessary to reorganize it in such a way as to make it equal to our reference organization. Consider the following examples, the first a reference card sort:

$$A = \{1, 2, 3, 4\} \text{ and } B = \{2, 3, 1, 4\}$$

To change B into A we need 2 operations, which is their edit distance (swapped cards in red, unchanged cards in gray):

$$B_0 = \{2, 3, 1, 4\}$$
$$B_1 = \{1, 3, 2, 4\}$$
$$B_2 = \{1, 2, 3, 4\}$$

3.2.3 *How the evaluation sessions were organized*

While it can be done in software, one of the reasons we opted for the card-sorting method was its feasibility in low-tech settings (Goodman and Santos, 2006): setting up the test consisted of basically laying out paper cards on a table. This tied it well with the three rounds of tests we ran: a first, with students from a specialization course in graphic design; a second, with affective computing students from an engineering graduate course; a third, with undergraduate design students. In all of these, students were in the middle of, or had just finished, class, so our "testing-facilities" were the necessarily unsophisticated settings around a typical college classroom.

Participants did the sessions individually: we found a corner as quiet as we could[12] and set up the cards on a table along the six emotion-categories. Participants could take their time to organize the cards, and while we explained that the visual modifications correlated with speech quirks, we gave no explanation of their underlying logic.

The low-tech setting also allowed for an important addition to the card-sorting method: after each session, we conducted a short, open-ended interview, where we asked participants about their strategies to organize the cards, mental models about what they thought were the inner workings of the algorithm that modulated the typography, and whatever else comments they wanted to share. If necessary, we asked follow up questions to clarify or expand on the issues that came up.

[12] We were successful for our first two rounds, done in quiet, isolated rooms. The third session was done in a very loud and busy classroom.

3.3 *Results*

The three sessions had, respectively, 13, 5, and 16 participants. The first and third sessions involved design students. The second, computer engineering ones. A photo from the first session can be seen in Figure 3.7.

3.3.1 *Card-sorting results*

Each participant organized 24 cards, divided into six emotions and four phrases. With 34 participants, this meant we had a total of 816 sorted pairs,[13] each comprised of one card being matched to one emotion.

One way to view these pairs is in terms of the how frequently each card which represented a given emotion was placed in its correct category, as shown in Table 3.1.

The results can also be viewed in Figure 3.8 in terms of how their edit

[13] 34 participants × 4 phrases × 6 emotions.

Fig. 3.7: A participant sorts cards into the six available emotions. Photo taken from the first evaluation session we ran, done in a empty classroom adjacent to a design workshop they were taking.

	Anger	Disgust	Happiness	Fear	Sadness	Surprise
Filha, rúcula...	29%	32%	6%	6%	12%	21%
Passarinho, cuidado...	20%	24%	24%	18%	41%	15%
Lilo, Kika, Luku...	6%	12%	6%	15%	15%	18%
Você tem...	24%	21%	18%	21%	15%	18%

Table 3.1: Average rate of correctly placed cards per emotion (columns) per phrase (rows). A confusion matrix of these results is presented in Pataca and Costa (2019).

distances were distributed. Since there were six categories, these can range between 0 — which, had it happened, would equate to a card sort where all emotions were correctly assigned — and 5. Note that mistakes here come in pairs, e.g., if a mistake is made where a card of category A is misplaced in category B, a second mistake will necessarily occur with whatever card is placed in category A.

3.3.2 *Interviews*

Interviews were always short: we had a limited time to run all rounds and had previously decided to prioritize collecting as many card sorts as possible. In this section we will present a summary of some of the main ideas that emerged from our conversations with the participants, complemented with some sparse notes we took while watching them sorting the cards.

3.3.2.1 *What visual interventions did participants consciously perceive?*

Many participants made the association of font-weight modulations as if representing loud volumes. However, this relation was mostly perceived in

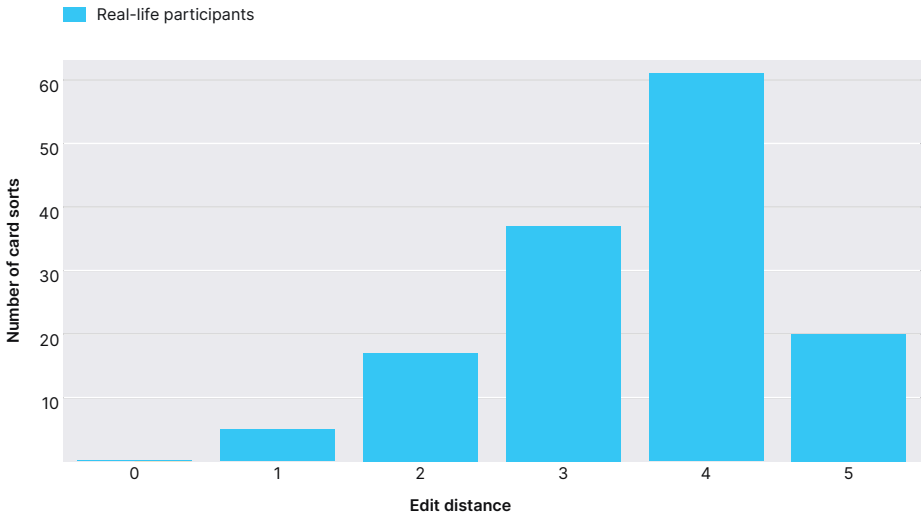


Fig. 3.8: Distribution of the sum of edit distances for real-life participants.

one direction only: many understood heavy syllables as being loud, but few cited the inverse relation of light syllables being quiet.

The two other typographic modulations had more diffuse interpretations. Slant was related to speed by one participant and as a weakness in the voice by another. A third participant thought that it could be associated not with prosody, but with specific emotions — a highly slanted phrase could be perceived as sad, for instance. Some participants mentioned having noticed that slant was a modulated dimension in typography, but could not decipher what it meant.

The third visual parameter — letter width — was mentioned only once by a participant, who correctly deciphered it as being associated with the duration of syllables.

3.3.2.2 *What sorting strategies emerged?*

Participants' responses here came in two groups. The more common strategy involved looking at each card and attempting to imagine how it could sound[14] according to their interpretation of the visual modulations. It was this voiced reading, rather than the images themselves, that would be classified in one of the six emotions.

A more rare approach went the other way: they would look at each emotional category and, with the phrase in mind, try to vocally interpret it in such a way as to represent the wanted emotion. With this sound in mind, they would seek a card that looked like the phrase had sounded.

For both strategies, there were accounts of how some emotions seemed easier or harder to find than others, even if the specific groupings of emo-

[14] Generally in silence, although we heard a few participants lightly murmuring the phrases to themselves.

tions varied a lot between participants. For some, this gave rise to the strategy of starting with easier emotions and, when these were done, work by exclusion with the ones left.

3.3.2.3 *Other comments*

Many participants told us they found the experiment too hard. In fact, while they were sorting the cards many looked lost, some even asking for help. We also noted that the first set of phrases usually seemed to take much longer than the rest.

Lastly, one participant mentioned that the *Você tem certeza disso?* phrase, unlike the other three, was hard to imagine being said under all six emotions because it seemed to allude to an unknown something — a possible indication that it was not as senseless as we had hoped.

3.4 *Discussion*

Perhaps due to a faulty experimental design, results from Experiment #1 were inconclusive. With them, it was not possible to reject the null hypothesis, i.e. the results were statistically indistinguishable[15] from what would be expected had the participants randomly assigned each card. This indicates that they were unable to recover from the speech-modulated typographic whatever sounds were behind each card, or that even if this effect was present it was too small to be detectable by our experimental set up.

[15] $\chi^2(5, N = 816) = 6.97$, $p > 0.05$

Another way to sense the negligible magnitude of the effect can be gathered by considering how the observed edit distances, also shown in Figure 3.8, barely differ from a random distribution generated through Monte Carlo simulations in Figure 3.9.

While all of this does not seem to bode well for our prosodic-typographic model, there are some design flaws in the experiment that can also be at work. First, we had not previously validated the specific pairings between prosodic features and typographic modulations. Their varying degrees of appropriateness seem to be indicated by the interviews. These showed that while magnitude $\rightarrow$ font-weight was generally a well-understood pairing, both pitch $\rightarrow$ slant and duration $\rightarrow$ letter width pairings were harder to interpret. In fact, not a single participant mentioned that pitch could be somehow represented.

Secondly, even if participants were being exposed to the audio files themselves we should not expect perfect scores for every emotion. In

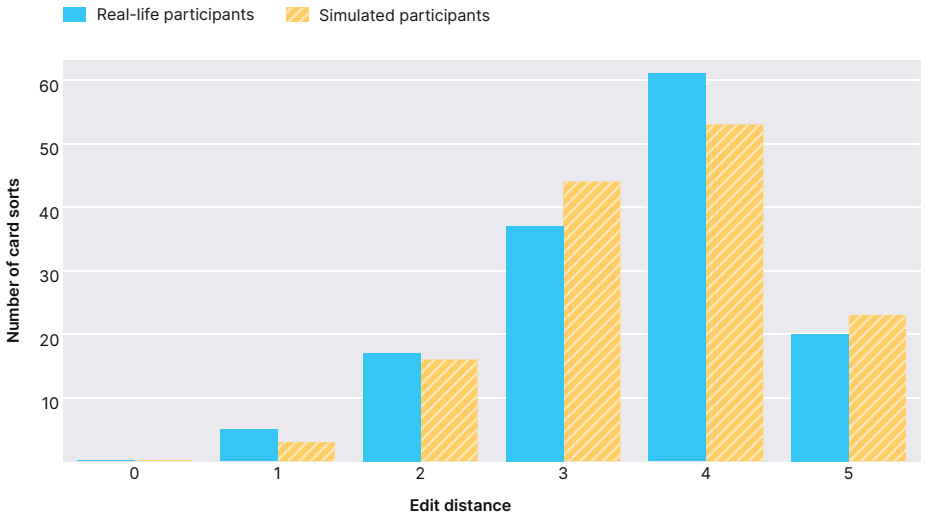


Fig. 3.9: Distribution of the sum of edit distances from all four phrases, in blue, versus the simulated distribution of randomly sorted cards, in yellow with white stripes.

fact, although her specific tests had different goals and methods, Costa (2015)'s recognition experiments show that, while some emotions tend to have high recognition rates (e.g., anger), others are somewhat more diffuse (e.g., fear and sadness). As a perfect mapping between sound and typography could not be reasonably expected, our results would presumably only show effects of a smaller magnitude than those of a test with the sound files themselves.

Lastly, while the card-sorting method had the advantage of allowing for observing participants doing the test and interviewing them later, it also showed major drawbacks. For one, it is hard to prepare, tabulate, and harder still to scale up the number of participants.

A significant issue was the difficulty to control for some noise-inducing characteristics of our setup. For instance: as we noted, participants would spend much more time organizing the first phrase as compared to the other three. While this could be explained by them taking their time to formulate mental models, an alternative explanation seemed plausible when we watched them take the test: the second, third, and fourth rounds were easier because, instead of attempting to "sound" mentally each of these cards, they were simply comparing them with their choices for the first round (which they could see). In a digital test, it is easier to shuffle phrases, hide past choices, measure how long each round takes, etc. On a paper-based test, not so much.

3.5 *Conclusion*

Taken at face value, Experiment #1 seemed a failure: while it did point towards some conclusions, these were vague. *Maybe* font-weight was a good proxy for loudness, and *maybe* more work should be done exploring alternative pairings for the other prosodic features, but either way the experiment was inconclusive. Questions were also raised about how adequate a categorical emotion model is when used as a proxy to understand if we are indeed representing speech typographically.

The vagueness of these answers could have been caused by our experimental methodology, or our model, or a bit of both. As an attempt to untangle these issues, Experiment #2 takes a step-back and attempts to uncover a clearer picture of how participants relate sound with typographic modulations.

Chapter 4

How did participants pair prosodic features with typographic modulations?

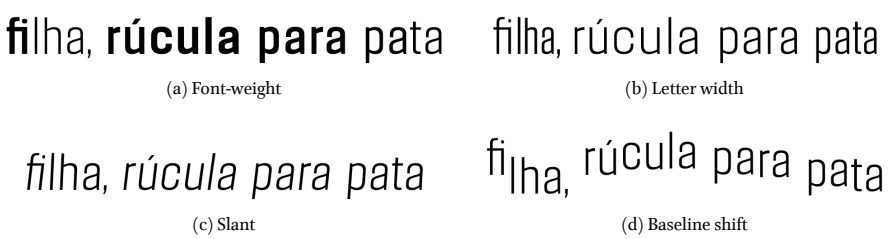
While in the first perceptual evaluation participants had been nearly unanimous in relating font-weight as a proxy for loudness, the perception of other prosodic features and typographic modulations were erratic.

In order to investigate the human perceptual mechanisms that enable the recognition of prosody features in typography, we designed an experiment[1] to measure how strongly (or weakly) participants related each typographic modulation to each prosodic feature. The results presented in this chapter are also discussed in de Lacerda Pataca and Costa (2020).

[1] Related to our second research question: What typographic modulations can be generally recognized as visual proxies for the prosodic features of magnitude, pitch, and duration?

4.1 *Generative algorithm*

We mostly used the same algorithm as we had for the previous experiment, repeating its extracting, feature-normalizing, and text-shaping methods. But, because of our now more specific goals, there was a key difference: for each typographic instance, we only mapped one prosodic feature to one typographic modulation.[2] Figure 4.1 shows an example of pitch being used as the input for four different typographic modulations, applied separately in each image. For the complete set of images we used, see Appendix B.



(a) Font-weight (b) Letter width (c) Slant (d) Baseline shift

[2] On the previous experiment all cards had the three typographic modulations, each tied to one prosodic feature, applied simultaneously.

Fig. 4.1: Four examples of speech-modulated typography where but one prosodic feature was mapped to each of four typographic modulations.

Since both letter width and slant were so rarely mentioned by participants in experiment #1's interviews, we supplemented the original set of three typographic modulations with a fourth one: *baseline shift*.

As stated, baseline shift has already been mentioned in previous works (Bessemans et al., 2019), but its use presents a challenge in situations where text is to be laid out in more than one line at a time, e.g., paragraphs. This is surely a limit of the approach, imposing that either loose leading[3] values be used or that low limits are set for how far syllables can shift vertically up or down — or, alternatively, accepting that there will be the occasional clash of syllables from one line on the other.

[3] The typographic term for the spacing between lines.

Yet, the apparent failure of our previous approach to represent pitch called for the exploration of a new typographic modulation. In Tatit (2007), a vertical shift in syllables is used as an informal melodic notation in studies of popular Brazilian songs (see, for example, Figure 4.2). This was an indication that, as an intuitive representation of pitch, baseline shift could be a promising modulation to include in the test.

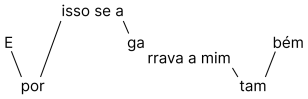


Fig. 4.2: Melody notation adapted from Tatit (2007).

4.2 *Method*

4.2.1 *Paired comparison*

To compare how adequate participants thought each typographic modulation was as a representation of each prosodic feature, we created an experiment where they would be able to rank their preferences. The idea was not to create an absolute scale, with preferences not determined by specific criteria but, rather, the intangible quality of something being more or less "adequate." Thus, we set to measure the relative proportions of each typographic modulation's strength against the other three, e.g., how likely it would be for it to be chosen when paired against the other three.

We used the speech corpus by Costa (2015). Having each utterance be associated with a different emotion made them also have very different prosodic patterns, allowing the created rankings to be analyzed through two different lenses: not only would we be able to measure the performance of each typographic modulation against each emotion, it would also allow us to correlate these performances with different prosodic patterns.

In other words, if for instance a given utterance was generally louder than others, it would be possible to measure if participants preferred font-weight to represent its louder prosodic features. In fact, we hypothesized that this would be precisely the case, as our previous experiment had pointed to. Unlike experiment #1, however, here we would have independent measures for the performance of each of the four typographic modulations against each utterance, giving us a more precise understanding of how

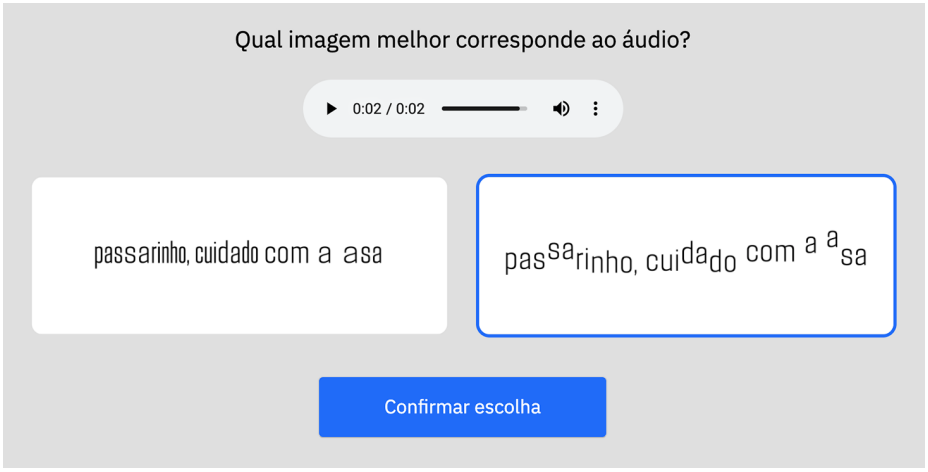


Fig. 4.3: Screenshot of how experiment #2's online platform's interface presented audio utterances and two corresponding modulated-typography choices. A choice could only be selected after the audio was played at least once. The translated title over the audio player reads "Which image better corresponds to the audio?"

these patterns were perceived.

We opted for a *paired comparison* experimental design, wherein each round participants would be exposed to one audio utterance and two instances of speech-modulated typography, both derived from the extraction of the same prosodic feature but each having used a different typographic modulation. After having listened to the audio, participants would choose which of the two options was its most adequate representation (see Figure 4.3 for a screenshot of the test).

As enough participants were exposed to all possible typographic matches for a given utterance, from the distribution of their preferences would either emerge statistically significant "winners" or indifference towards the four given options. This could either allow us to refine our model, ensuring its typographic choices were intuitive representations of prosody, or be an indication that the approach as a whole should be reassessed.

4.2.2 *Combinatorial explosion*

Combining the four typographic modulations gave us six pairings, which had to multiplied by the six emotions (plus one, representing a *neutral* stance[4]), three prosodic features and the four phrases to give us the total number of pairings. So,

$$\frac{4!}{2! \times (4-2)!} \times 7 \times 3 \times 4 = 504 \text{ pairings},$$

a rather impractical test setting — if each participant took 30 seconds to judge each of the 504 pairings, the test would take more than four hours! So, simplifications had to be found. The issue was what cuts would have the least impact on how diverse the pairings would be.

[4] A *neutrally* emoted phrase, where prosodic parameters can be assumed to have little variation, could act in this experiment as a sort of control-utterance, where we expected participants' preferences to be close to a random distribution.

4.2.2.1 *Reducing the number of phrases*

Instead of the four original phrases, we narrowed the set to two: *filha, rúcula para pata* (in English, *daughter, arugula for the duck*), and *passarinho, cuidado com a asa* (in English, *birdie, watch your wing*). With this, we were left with 252 pairings.

4.2.2.2 *Reducing the number of prosodic features*

For the specific subset of phrases we chose, we found that there was a moderate to strong linear correlation of -0.6 between measures of magnitude and duration for each syllable, while values of pitch and magnitude where more divergent (0.2 linear correlation). We thus removed from the test the duration prosodic feature, since its typographic modulations would have a somewhat similar (even if inverted) dynamic with magnitude's. With this, we were left with 168 pairings.

4.2.2.3 *Reducing the number of emotions*

As we had introduced a *neutral* utterance, we considered that emotions with little variation in their prosodic patterns should be excluded, since they would generate similar looking speech-modulated typography instances with neutral and among themselves. Considering average, normalized values of RMS, we removed both fear and disgust utterances in both phrases. With this, we were left with 120 pairings.

4.2.3 *The experiment's online platform*

A decision was made that, instead of exposing all participants to all possible combinations, we would define an adequate test duration and find the number of pairings that a participant would, on average, be able to rank in an equivalent time. This had a clear disadvantage in that, as the percentage of pairings per participant decreased, we would need an inversely larger number of participants to cover all possible pairings. It was necessary, however, as we found that after arriving at 120 pairings any further cuts would start imposing too great a loss in the granularity of the experiment's results.

This was planned from the start as an online test. Considering participants were at all moments but a click away from abandoning the test, we defined that a good test duration would be in the range between 10 and 20 minutes. This came from a prototype version of the test, where we found 30 pairing reasonably comfortable to rank — if we discount an

(excessively) conservative 5 minutes for test instructions, we would have $^{15\ \text{min}}/_{30\ \text{rounds}} = 30$ seconds per comparison, which seemed adequate.

We divided the 120 pairings in two batches. For the first, participants as a whole would judge the 60 pairings generated from the extracted RMS; for the second, pitch was used. Since each test had only 30 rounds, each participant would only be exposed to half of all possible pairings.

To counter for participants' initial ineptitude and eventual tiredness, not only the choice of pairings but also their order were randomized every time a new test started.

The test itself consisted on a website,[5] where there was (1) a very general introduction to speech-modulated typography[6] and its possible applications, (2) a guide on how the test would work, (3) an informed consent form, (4) a simple demographic form asking academic formation, age, and sex of the participant (none of which were required), and, lastly, (5) the test itself, as shown in Figure 4.3. After finishing 30 rounds of ranking typographic pairings, participants were thanked for their participation and could optionally send us comments on whichever topics they wanted — which, like the rest of the test, were anonymous.

[5] Developed in the VueJS framework, its source code is available in de Lacerda Pataca (2019).

[6] We did not explain how certain prosodic features were translated as typographic modulations but, rather, gave an account of an attempt to *imitate* the human voice through typography, which, we wrote, had echoes in a common expressive resource used in comic-book lettering.

4.3 *Results*

Participants were recruited through email invitations sent to graduate students in our department and invitations posted on Twitter. Of the 78 participants that completed the test, 46% were female, 53% male. Around 90% had bachelor degrees or higher. In terms of age, 13% were between 18 and 24 years, 59% between 25 and 39 years, 19% between 40 and 59 years, and 5% had 60 or more years. Thirty-four participants ranked the typographic modulations generated with magnitude, and 44 participants ranked those generated with pitch.

The 2,340 comparisons we measured gave us the ranking shown in Table 4.1. We have divided it in two groups, one for each of the two prosodic features. Each row gives us the ranking for each emotion, with percentages of preference for each typographic modulation shown in each column. Note that each cell shows the (rounded) percentage of times that, when that specific typographic modulation was shown against any of the other three modulations, it was chosen by participants.

Since each emotion appears twice, once for each extracted prosodic-feature, we highlighted in blue those cells that point to the highest overall

Rankings for magnitude as represented prosodic feature				
emotion	font-weight	letter width	slant	baseline shift
anger	83%	34%	43%	39%
happiness	34%	54%	47%	64%
neutral	47%	52%	76%	25%
sadness	45%	52%	46%	58%
surprise	72%	38%	49%	43%
Rankings for pitch as represented prosodic feature				
emotion	font-weight	letter width	slant	baseline shift
anger	87%	46%	45%	26%
happiness	28%	66%	40%	69%
neutral	47%	55%	63%	35%
sadness	21%	57%	39%	79%
surprise	71%	43%	41%	44%

Table 4.1: Participants' preference rankings for each prosodic feature (each of the sub-tables), per emotion (each row) and per typographic modulation (each column). Gray-tinted cells indicate the highest performing typographic modulation for that emotion and prosodic feature. If the cell is in blue, that typographic modulation has the highest performance for that emotion considering both prosodic features.

performance considering all utterances representing a given emotion. For instance, Anger's Pitch's 87% performance is blue because it is higher than Anger's Magnitude's 83%. No preference was highlighted for the sad utterance generated with magnitude. Running its distribution through a Chi-test,[7] gave a p-value of more than 0.05, indicating that its answers were indistinguishable from random choices.

From Table 4.1, font-weight and baseline shift were the typographic modulations with the best overall performances. Font-weight captured participants' preferences for the surprise and anger emotions when we used, respectively, the magnitude and pitch prosodic features. Baseline shift came out on top for happiness and sadness, in both cases representing pitch. Slant won general preference once when it was used to represent magnitude in the neutral emotion.

[7] $\chi^2(3, N = 190) = 2.6, p > 0.05$

4.3.1 *Consistency between emotions and participants' preferences*

With the rankings, it was possible to measure if participants' preferences were consistent when controlled for *emotion*, i.e., would typographic modulations respond to a semantic interpretation of prosody?

Let us consider, for instance, how font-weight ranked as a representation of magnitude taking into account only the *Filha* sentence. Analyzing separately its angry, happy, neutral, sad, and surprised emotions, performances were, respectively, of 91%, 33%, 44%, 51%, and 87%. For the same sentence and same emotions, but using font-weight to represent pitch, we had performances of, respectively, 84%, 24%, 50%, 21%, and 86%.

With the *Passarinho* sentence, anger's pattern of high preference for font-weight is repeated, but virtually disappears for the surprised emotion: from a performance of 87% (magnitude) and 86% (pitch) in *Filha*, in *Passarinho* we get 58% (magnitude) and 57% (pitch).

This inconsistency repeats itself in other sentences. In the two happy sentences, for instance, when using baseline shift as proxies for pitch preferences varied from 60% to 80%.

This calls for relating these performances not necessarily to their corresponding emotions but, instead, to each recording's specific prosodic patterns. Here, we use f_o to measure pitch, but instead of the direct values of RMS to measure magnitude we preferred the metric given by $C - f_o$, where C stands for a spectral centroid.[8] It goes that a loud utterance will have a wider spectral range than a quieter one and, therefore, higher centroid values, which come without equivalent increases in f_o. The relationship between both values is largely unchanged by, for instance, changes in distance between the speaker and the microphone, making it a more resilient metric than RMS when considering changes in the recording environment.

[8] Calculated as the weighed mean of frequencies in a signal, giving its spectrum's center of mass.

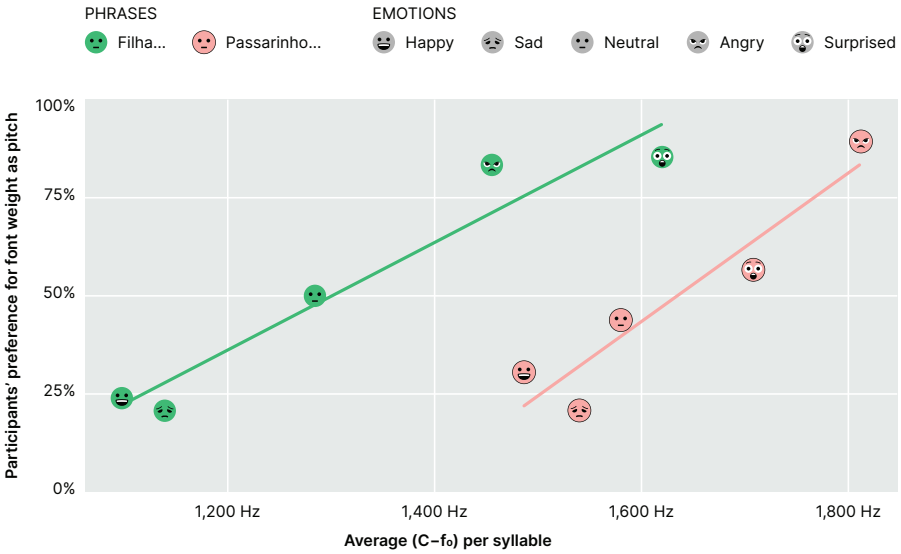


Fig. 4.4: How each sentence's average $C - f_o$ value mapped against font-weight's performance when used to represent pitch.

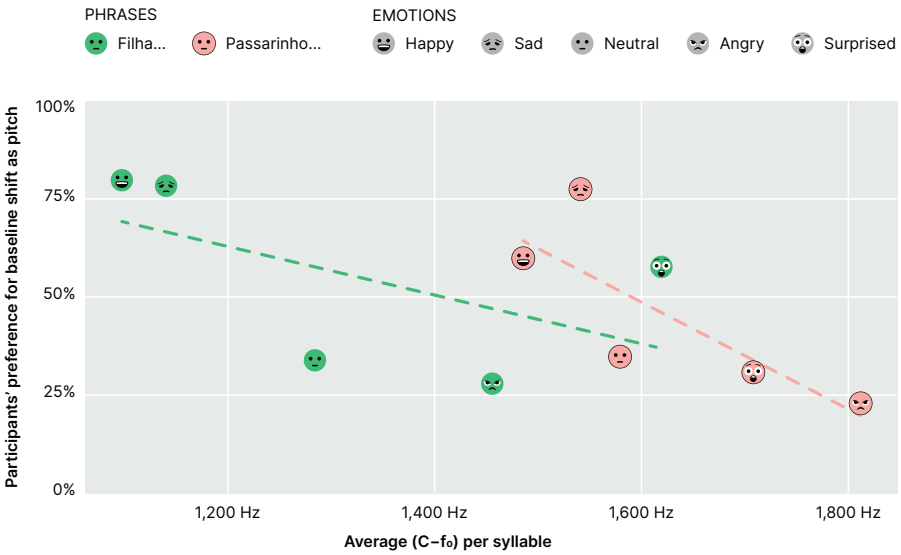
In Figure 4.4 we relate each utterance's average $C - f_o$ value against participants' preferences for font-weight used to represent pitch. For the *Filha*[9] and *Passarinho*[10] phrases, the regression equations were statistically significant and show that, as $C - f_o$ increases in value (i.e., the actress' voice becomes louder), so does the performance of font-weight.

Figure 4.5 is similar to Figure 4.4 in that it relates average $C - f_o$ values against participants' preferences, specifically considering how they rated

[9] ($F(1,3) = 39.27$, $p < 0.05$), RSE = 0.09. R^2 of .93 and expected performance of $(C - f_o) \times 10^{-3} - 1.29$.

[10] ($F(1,3) = 24.07$, $p < 0.05$), RSE = 0.10. R^2 of .89 and expected performance of $(C - f_o) \times 10^{-3} - 2.61$.

the use of baseline shift as a representation of pitch. There is an inverse correlation between $C - f_o$ and this preference, albeit not significant (shown therefore as dashed lines).[11]



[11] For the *Filha* phrase, ($F(1,3) = 1.36$, $p > 0.05$), RSE = 0.23. For the *Passarinho* phrase, ($F(1,3) = 5.26$, $p > 0.05$), RSE = 0.16.

Fig. 4.5: How each sentence's average $C - f_o$ value mapped against baseline shift's performance when used to represent pitch.

4.3.2 *Open-ended comments*

There was no specific prompt asking for participants' comments, meaning they could tell us whatever they wanted. Twenty-one of the 78 did. Many described their mental models to interpret the typographic modulations, while others sent us suggestions for possible additions or changes to the four we used. There were also some more general comments about the experiment and our model of speech-modulated typography as a whole.

4.3.2.1 *Typographic modulations*

Font-weight was often associated with an aggressive, firm tone of voice.

Baseline shift, on the other hand, carried more diverse readings: some related it with a singing, melodic tone of voice, while others interpreted in through more colorful lenses, such as one participant who saw it representing a "playful, happy, kind of drunk" tone of voice.

Slant modulations were said to be difficult to discern, with one participant writing that they could only discern changes in slant in the faster phrases. One participant commented that it seemed to portray a formal, monotonous tone of voice.

Letter width was associated with a weak, timid tone of voice. One participant related it with insecurity, and another with a slow enunciation.

There were also some comments about possible expansions or modifications to the model and set of typographic modulations. One suggested that a font size modulation could have been used to suggest emphases in the voice. Another mentioned that trembling letter shapes could be used to suggest a fearful, submissive tone of voice.

4.3.2.2 *Other comments*

Some participants complained about how the test was set up. One complained about a lack of academic formation and gender identity options.[12] Another expressed suspicions that our setup might not be managing to isolate the effects we were measuring — as they said, "in most of my choices, two or more aspects (often conflicting) were taking place." Some participants wrote that the test was difficult and/or too long. Also, some thought that it was hard to know what to do when both options seemed equally appropriate to represent the sounded utterance.

Lastly, there were comments about how reading long texts in our proposed speech-modulated typography could become uncomfortable, with one participant specifically warning us about how baseline shift would create juxtaposition issues in multi-line text settings, and another writing that the more extreme values of the typographic modulations felt strange.

[12] For this experiment we presented participants with only two gender options and the possibility of not answering. In experiment #3, we added *non-binary* and *other* options.

4.4 *Discussion*

4.4.1 *What about the typographic modulations?*

The consistency of the results indicated that it is indeed plausible to construct a speech-modulated typography model that intuitively translates spoken prosody into typographic modulations. There was strong evidence that *font-weight* is a good visual representation of changes in magnitude,[13] and, while not as robust, *baseline shift* seems adequate to represent changes in pitch.

The experiment only evaluated one modulation per typographic instance. When the actress' voice was loud, font-weight won preferences and, when quieter, baseline shift was preferred. This does not imply, however, that a combination of the two would or would not work.

Slant's role is more ambiguous. It was the preferred modulation for the neutrally voiced utterances, where we had hypothesized that no clear preference would emerge. But maybe it was best not to take this at face value: as one participant said, slant[14] is generally hard to discern. In utterances

[13] Echoing Van Leeuwen (2006), who says that *[b]old can be made to mean 'daring,' 'assertive,' (...) and its opposite can be made to mean 'timid,' or insubstantial.*

[14] At least for the particular typeface used: Compressa.

where the effects are themselves subtle, it could just be that slant stood-in for a missing "no modulation" option. While the test does not specifically answer this, Figure 4.6, which contains the set of *Passarinho* neutral utterances in which we represented the magnitude prosodic feature, shows that slant is plausibly the least expressive option of the four, making it a possibly adequate choice for a monotonous tone of voice.

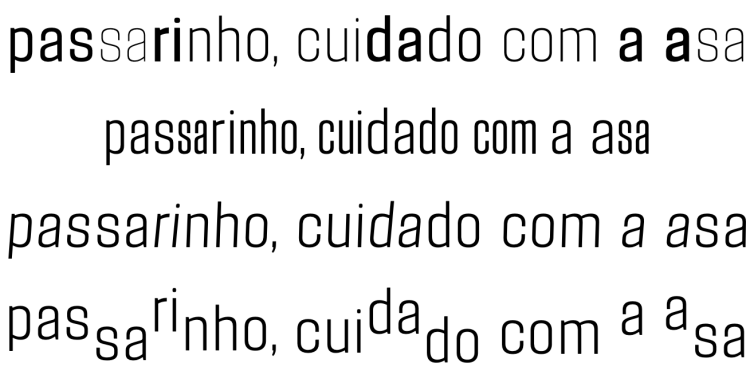


Fig. 4.6: The four typographic modulations being used to represent magnitude in the *Passarinho* phrase. From top to bottom: font-weight, letter width, slant and baseline shift.

While *letter width* was not universally despised, it did not achieve eminence in any emotions either. Unlike slant, this does not seem caused by it being visually imperceptible and is, therefore, a sign that it would not be an intuitive representation of either magnitude or pitch.

4.4.2 *What underlying mechanisms guided participants' preferences?*

While generally anecdotal, participants' comments do point to some possibilities about how our model was being interpreted. First, it seems that only very rarely did participants see the modulations as if representing the specific, categorical emotions the actress was portraying in the audio files. There were also very few descriptions of the modulations as if representing the specific prosodic features we used in the experiment. Instead, the interpretations suggest participants were seeing prosodic features that were *changed* by emotions, e.g., an aggressively strong voice.

While this is a tenuous, speculative interpretation, it would be in line with what Boehner et al. (2005) writes about "communication of affect in an interactional model (...) [being] more than transmission, [and requiring] active interpretation." Considering, as they do, a model where emotion is a social, contextually-embedded construct, we can speculate that a categorical model attempting capture and represent universal signs of emotion would not work or, if so, only in the most constrained settings.[15]

While it is convenient to implement and evaluate, using in our research a categorical model such as Ekman's Big Six could just be a case of how "[we]

[15] Besides issues with the model or experimental setup, this could be a additional factor behind the diffuse results in experiment #1: different people may idealize each of the six emotions in a different way.

designers choose to reduce the polysemy of emotion for the convenience of technical constraint." (Stark and Hoey, 2020)

It is therefore a promising, and not unfortunate, sign that participants' subjective interpretations were all over the place even if their objectively measured preferences converged. Instead of an attempt to resolve what is inherently ambiguous in speech, our model could enhance and enrich how one experiences sounds and words by amplifying prosody's expressiveness.

4.4.3 *Shortcomings*

Since the typographic modulations were shown as a complement to audible speech, our results may not necessarily replicate in situations where text is shown by itself. The experiment also did not explore how typographic modulations could interact if applied in tandem, which could affect the measured effects.

It is also worth considering that most of the participants were college-educated, from which it is reasonable to presume a greater-than-average familiarity with the written word (Instituto Pró Livro, 2020).

4.5 *Conclusion*

The experiment showed strong preferences associated with specific pairings of prosodic features and typographic modulations, specifically magnitude $\rightarrow$ font weight and pitch $\rightarrow$ baseline shift. While this deals with but a dimension of many in a complete speech-modulated typography model, it was a positive indication of how such a model could effectively visually translate aspects of prosodic expression.

It also strengthened the case for abandoning categorical emotion recognition as a proxy for the model's effectiveness. The key to understanding participants' preferences was found in prosody rather than the categorical emotions portrayed in each utterance.

Chapter 5

The Model (and a pilot implementation thereof)

In this chapter, we will discuss our proposed model for generating speech-modulated typography, as summarized by Figure 5.1, after which we will discuss aspects of a pilot implementation of the model.

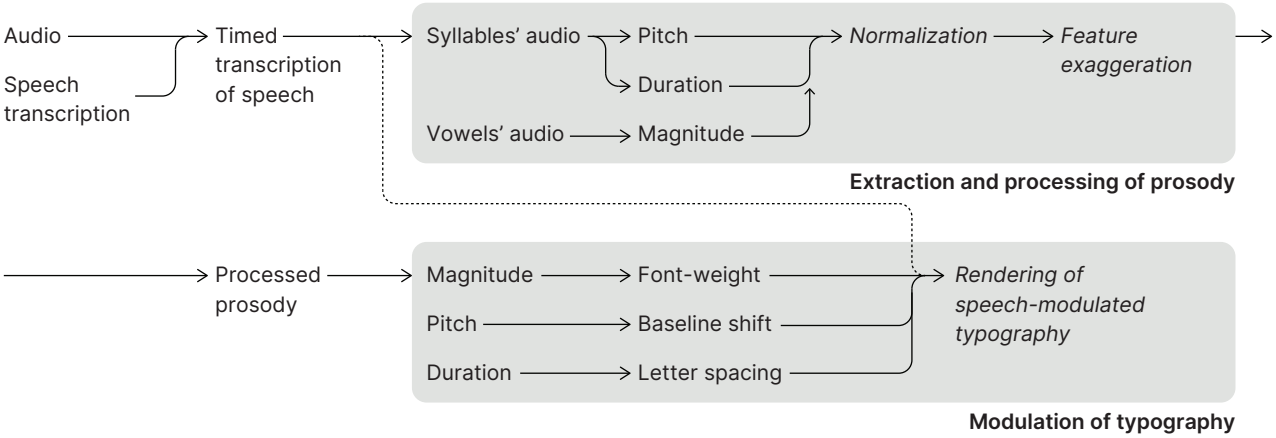


Fig. 5.1: Diagram of the speech-modulated typography model.

An important factor distinguishing this version from those discussed in the previous chapters is that, here, we were interested in a model that allowed for speech-modulated closed-captions. While we still constrained ourselves with audio from only one speaker, having it animated and synchronized to a sound recording imposed new challenges. Now, the extraction/normalization algorithm would have to deal with the chaining of multiple utterances, and the typographic shaping engine would need to deal with multi-line text, as needed for a moving-image context.

The model presented here was tested in a final perceptual evaluation, which we present in the next chapter.

5.1 *The model*

Broadly speaking, the model consists of a step where a speech's prosodic features are extracted and processed, followed by a mapping of these features into typographic modulations, which can be used to render speech-modulated typography.

5.1.1 *Extraction and processing of prosody*

5.1.1.1 *Timed transcription of speech*

Prosodic features are extracted considering either the whole syllable or only its vowels' sounds. For this, a timed transcription of speech needs to discern not only the timestamps of each phrase and word in the audio file, but also each word's syllable and each syllable's vowels.

In previous versions of our algorithm we had not made this distinction, and each of the three prosodic features were extracted directly from syllables. Here, however, and especially after some rounds of internal testing, we noted a discrepancy between the measured values for magnitude and how we were perceiving loudness, notably in syllables that had long segments of consonantal sounds. We eventually realized that, by only extracting magnitude from vowel-regions, the obtained values better echoed our own perceptions of loudness of each syllable.

Thus, pitch and duration should be extracted considering the whole syllable, while magnitude should only consider vowels.

5.1.1.2 *Feature normalization*

Having extracted the raw acoustic measurements for the three prosodic features, these need to be prepared for their use as input in a typographic shaping engine. Unlike previous versions of the algorithm, we considered that a simple normalization between maximums and minimums taken from the whole of the audio file — which, as we would be working with closed-captions, would be in a scale or minutes or even hours instead of a few seconds — would make each individual syllable too unaffected by the local prosodic dynamics of its immediately-surrounding syllables.

It would also ignore how, in speech, each prosodic feature is not perceived independently from one another but, rather, in their relation to each other in a given utterance. Prosody marks different levels of salience, which it does through the differences between its values in each syllable more so

than through their absolute values. For example, the same syllable that we might register as *loud* in a quiet environment could be perceived as *normal* in a noisy environment, where all its surrounding syllables are equally as loud. Prosodic effects are thus perceived as the contrast and interaction of measures against the general backdrop of the utterance.

In other words, we had to find a way for values to reflect not only their intensity relative to the whole audio, as they did previously, but also how they paired against their immediate neighbors. This echoes the process Barbosa (2019) describes for doing rhythm analysis in speech utterances, in which one considers how each measurement relates to the average value of the whole utterance, after which the value is smoothed by a weighted moving average that considers only its surrounding 5 data-points.

In searching for this mix between each syllables relation to a global and local contexts, we used a mix between our previous global normalization and a new, local globalization, comprised of a moving window that considered maximum and minimum values of the last 10 syllables and next 5 syllables at each given moment. The formula is as follows:

$$z_i = \frac{1}{2} \times \frac{x_i - \text{Local min}(x)}{\text{Local max}(x) - \text{Local min}(x)} + \frac{1}{2} \times \frac{x_i - \text{Global min}(x)}{\text{Global max}(x) - \text{Global min}(x)}$$

- where x_i is the i^{th} value of the x feature, and
- z_i its i^{th} normalized value considering local and global normalizations.

An example of how this dynamic between two normalizations changes the final value is demonstrated in Figure 5.2.

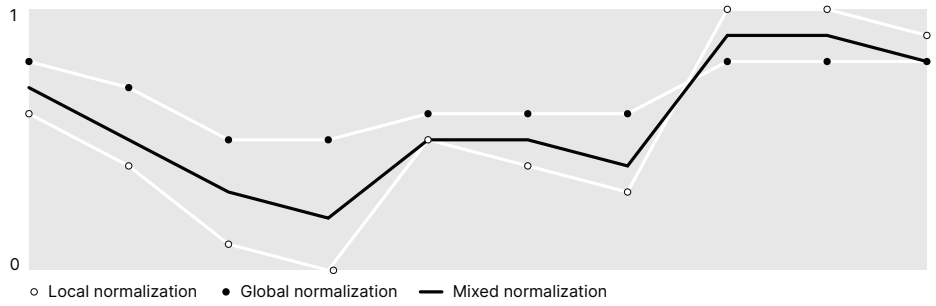


Fig. 5.2: Example of how a mix between local and global normalizations, as used in the model, changes through time.

Note how the local normalization is finicky and fast changing, while the global normalization is steady and slow moving — but, as we found, too indifferent to the local dynamics at each instant. The mixed normalization is, thus, our attempt of balancing both approaches. An example of how this dynamic is expressed in typography is shown in Figure 5.3. To heighten the

effect, in this particular image we applied a set of values (generated from a gradient noise function) to each letter, rather than the typical per-syllable division of our model.

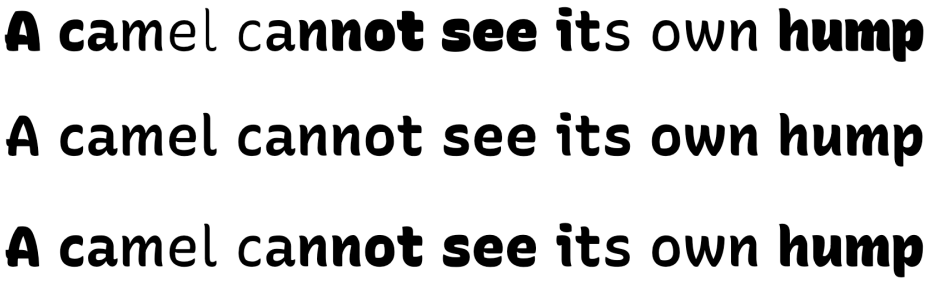


Fig. 5.3: Example of the effects of the three normalizations. From top to bottom, respectively, local, global, and mixed normalizations.

Two of the parameters here — the windows of 10 syllables back, 5 syllables forward, and the $1/2$ weight of each normalization — were arrived at through trial and error, searching for values where the typographic modulations were responding to the local dynamics at each instant, but not overly so. This same normalization logic was applied to the three prosodic features used: magnitude, pitch and duration.

5.1.1.3 *Feature exaggeration*

Unlike with the psychoacoustic-perception of prosodic features, while testing with speech-modulated typography we felt as if the dynamics of sound, when translated visually, had to be exaggerated to be meaningfully perceived. This could of course be the topic of a separate, systematic study, but here we settled for a provisional solution of squaring or cubing the normalized values. This softened each value, with already-small quantities losing more than closer-to-one values.

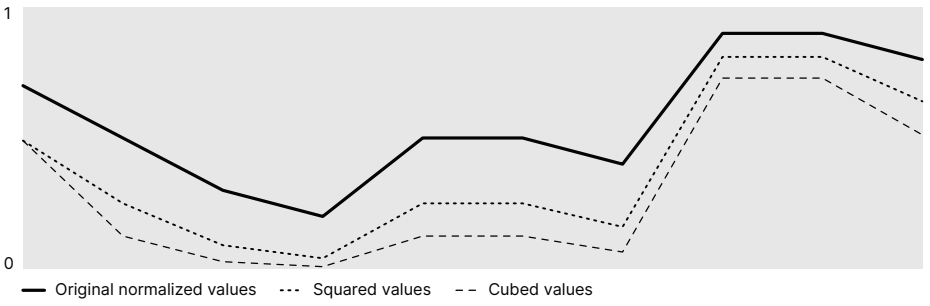


Fig. 5.4: Example of how the normalized values coming from the prosodic features change when squared or cubed to serve as input to the typographic modulations.

Choosing the operation depended on the specific prosodic feature (and later typographic mapping) in question: magnitude seemed to benefit from its values being cubed, while pitch and duration worked better squared.

5.1.2 *Modulation of typography*

5.1.2.1 *Prosody-typography mappings*

Following our previous experimental findings, we mapped the processed prosodic features of pitch and magnitude to the typographic modulations of, respectively, baseline shift and font-weight. The mapping formulas between these features and their corresponding typographic modulations depends on the attributes of the rendering context, e.g., what font is being used, what is its font-weight range, at what size type is being displayed, how loose or tight the leading is set, etc, so it is expected that a designer will fine-tune these ranges when implementing the model.

We also introduced in the model the use of duration for a new typographic modulation: letter-spacing. This consists of the straight-forward changing of spaces between letters in tandem with how long a syllable is relative to its context. Since negative letter-spacing[1] values hamper legibility (Bigelow, 2019), we used only positive values for letter-spacing, i.e., a faster-than-average syllable is represented not by its letters being brought closer together but, rather, an average or slower-than-average syllable will always have its letters displayed with positive letter-spacing values.

[1] Also known as tracking.

5.1.2.2 *Rendering of speech-modulated typography*

There are two rendering modes for speech-modulated typography. The first, for use in static contexts — e.g., printed matter, images, digital text, etc —, consists of laying out the text with each of the typographic modulations applied to each syllable, as shown in Figure 5.5.

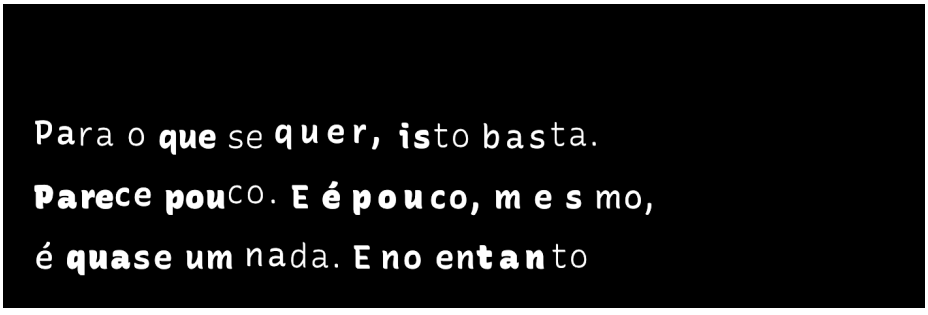


Fig. 5.5: Speech-modulated closed caption example, rendered as a static image, i.e., all typographic modulations applied at the same time.

In the context of closed-captions, however, where text will be displayed while the speech is being heard, animating the transitions allows the typography to better represent how speech changes in time, an important dimension in prosody.

To animate the typographic modulations, we defined that when the text surfaced on the screen all of its syllables should be in their default setting, almost a *resting* state before the changes began in sync with the audio being played. As soon as the speech starts, the three typographic modulations are applied synchronously with the audio. However, for the animations to be perceived as if echoing the sound, they have to start before the syllable's zeroth second in the audio.

This accommodates two issues: first, we wanted the change between default and fully modulated typographic states to follow a smooth transition curve, avoiding a jarring visual impression. Secondly, we have to consider that each syllable's prosodic values are equal to an average of the whole syllable, which ignores how the feature changes through the duration of the syllable. If we synced start and end times, we would have the typographic modulation peak at the end of the syllable, which in our testing gave the impression of the typography always being a little late.

To achieve the perception of typographic modulations being changed in sync with speech, we set the duration of each syllable's animation to twice its actual duration, with the animation's half-point aligned with the syllable's start-point. In other words, the animation starts before the syllable is heard, but they both end together.

Lastly, the model heightens the perception of typographic modulations being changed synchronously with speech by also changing the color of each syllable. In their initial, resting state, all syllables will have a subtle gray hue.[2] When they are active, i.e., when the syllable is being heard and its corresponding typographic modulations are being applied, this color will change to a pure white. After this moment passes, the syllable transitions back to a slightly darker gray hue than before. This creates a highlighting effect that seems to accentuate the typographic changes while guiding the eyes through the text as it changes states.

[2] This assumes light text against a dark background.

5.2 *Pilot implementation of the model*

To fine-tune the model and evaluate it in the experiment described in the next chapter, we developed a pilot implementation of its "Extraction and processing of prosody" module in Python, and of its "Modulation of typography" module in HTML / CSS / Javascript. In this section, we also discuss aspects of the segmentation/syllable annotation procedures of the speech transcription, which was manually done in Praat.

This pilot version of the software was not planned as a stand-alone, consumer-facing product. Nevertheless, since we foresee this being a future possibility — or, at least, of one day when speech-modulated typography will be implemented as a set of tools usable by tech-savvy interested-parties —, there was an attempt to avoid arcane, custom processes and data-structures whenever a suitable, well-established alternative could be found. This was the case both for our syllable-annotation procedures and, later on, for the closed-captions' file format.

5.2.1 *Measuring prosody*

5.2.1.1 *Defining syllables and vowels*

Where we had previously used a manually edited JSON file to define the boundaries of each syllable in a given utterance, here we decided to use the TextGrid file format. This is a native format of the Praat software — an open-source tool developed by Boersma and Weenink (2020) that has become the de facto standard for speech analysis in the linguistics community (Barbosa, 2019), where it has many users and extension developers.

A TextGrid organizes annotations of a digital sound file. Through Praat's graphical user interface, one can subdivide an utterance into labeled *regions*,[3] themselves organized in *tiers*. For our purposes — sectioning speech utterances that would be then translated as modulated typography —, we were interested in defining two tiers: *syllables* and *vowels*.

As discussed in Section 5.1.1.1, we needed *syllables* to be defined in order to extract from them values of duration and pitch, and we needed *vowels* to extract magnitude. An example of this process is shown in the Praat screenshot, in Figure 5.6, where the two tiers are visible. In the top one (*Syllables*), we can see how the "ríc-" syllable is delimited. Note how, on the magnitude visualization on the first row and the spectrogram on the second, there are three distinct visual patterns contained in the same syllable: a first, /ɾ/ sound, has a weak, irregular magnitude and a diffuse spectral signature. Next, the /ɪ/ sound, which has both a robust magnitude shape and more clearly defined spectral patterns. Lastly, an almost silent passage marked by the short, explosive /k/ sound.

While the "ríc-" section on the *syllables* tier comprises the three excerpts, on the *vowels* tier we have sectioned only the /ɪ/ sound, with a hash symbol labeling both /ɾ/ and /k/ sounds that surround it. We will delve into this syntax in greater detail when we discuss the subtitle files but, for now, the #

[3] While the process of delimiting and labeling sound files takes some time getting used to, Praat's interface is optimized for it and there are many readily-available educational resources on the topic, making it easier than our previous process.

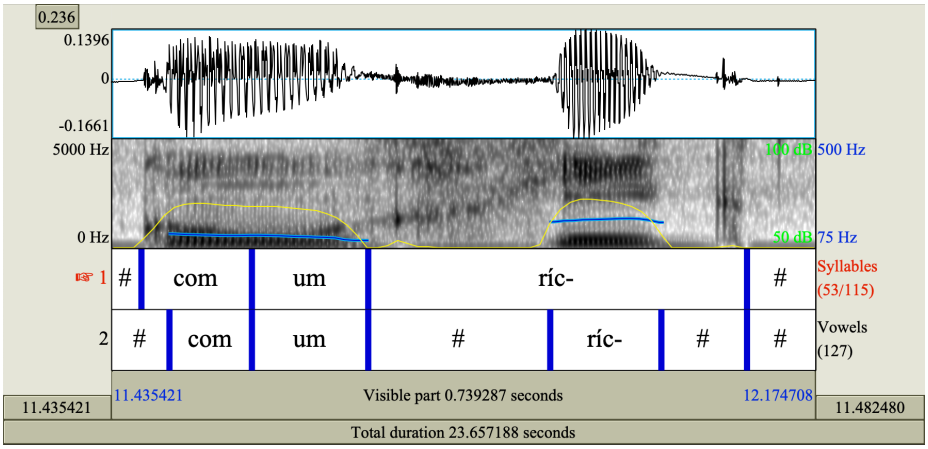


Fig. 5.6: Screenshot of a small segment of an audio file open in Praat. There are four rows, that are, from top to bottom: (1) the amplitude contour, (2) the spectrogram, (3) the *syllables* tier (tier #1, indicated by the red manicule, on the left, and by its name in red, on the right), and (4) the *vowels* tier.

stands in for a region of the audio file that will be ignored in the extraction phase. For syllables, these contain all silent regions. For vowels, all regions outside of vowels themselves.

5.2.1.2 *Praat as a sound-processing engine*

In the tentative versions of this model used in the previous evaluations, we had been using the *librosa* and *pysptk* Python libraries to extract, respectively, RMS and f_0. While they were adequate, we decided to migrate our magnitude and pitch extraction functions to Praat, while still working within Python.

The change was relatively simple: TextGrids are a native Praat format, making the integration[4] between segmentation/labeling and extraction straightforward. As a baseline, Praat's native extraction algorithms are excellent — it is an industry-standard speech-analysis tool —, and having a Praat integrated workflow opens future possibilities, e.g., use of extraction, analysis, and even automatic segmentation and annotation scripts made by its community of extension developers.

[4] We used the praatlo library, by Mahrt (2020), which wraps Praat's command-line functionality in Python functions.

5.2.2 *Encoding speech-modulated closed-captions*

Once prosodic features were extracted and processed, and considering the fact that we would be working with animated closed-captions, we had to encode how these values would be temporally mapped into typographic modulations.

To define these speech-modulated closed-captions, we decided to use the WebVTT (Web Video Text Tracks) file format, a text-based standard created by the World Wide Web Consortium (W3C). It allows for the dis-

playing of timed-text tracks, such as closed-captions and subtitles, through the <track> HTML element. It can be associated with a <video> element, but also allows integration with other formats.

This decision was driven by three main factors. First, WebVTT files are human-readable and relatively easy to edit manually, if needed. Secondly, as of 2020[5] they are compatible with all major web browsers. Lastly, and most importantly, the standard allows for the definition of arbitrary style classes associated with each text unit, which can be associated with a specific, otherwise effectless <c> tag.

There are no practical limits to how many classes are used, and, like with regular HTML/CSS classes, if they are not defined they will simply be ignored when the text element is rendered. This was important, as per our goal of using off-the-shelf formats, we could use these classes as a way of embedding encoded typographic-modulation information inside a typical WebVTT file without breaking its compatibility with typical video software. Our custom classes could be programmatically interpreted in our speech-modulated typography engine but, in any other setting, would be ignored and the file would render as any normal timed-text file would.

The TextGrid file which was used to subdivide the recorded speech utterance into its constituent syllables (and vowels) also contains the transcription that will allow the timed-text file to be constructed. As stated, we used the hash symbol (#) to define tier segments that should be ignored (e.g. silences, in the syllables' tier, and all non-vowel segments, in the vowels' tier). A double hash (##) stands in for line-breaks, and the pipe symbol (|) indicates the end of a cue.

A WebVTT file starts with a WEBVTT identifier, a blank line and, then, all its text cues. Each cue begins with two timestamps, marking its starting and ending time. They are separated by a --> arrow, and correspond to the period when the text will be showing on screen relative to the start of the video/sound/etc file.[6] In the following lines there is the text itself. An empty line separates each cue. For example:

```
WEBVTT

00:00:00.000 --> 00:00:10.599
Para o que se quer, isto basta.
Parece pouco. E é pouco, mesmo,
```

[5] Per *Can I Use* (https://caniuse.com/?search=webvtt, accessed on December 22, 2020), WebVTT is currently compatible with browsers used by 96% of measured users.

[6] While typically used in tandem with films, it is possible to use closed-captions for a sound file, for instance, and is in fact recommended "for people who are hard of hearing to get the richness of listening to the audio." (Henry, 2020)

```
é quase um nada. E no entanto

00:00:10.599 --> 00:00:18.422
cabe um bocado, cabe tanto
que é até preciso dar um basta.
Quanto ao assunto – o si-mesmo –
```

5.2.2.1 *Class naming scheme*

We defined a class syntax that followed the pattern:

v_cccc-nnnn

It contains three parts: the `v_` identifier — a namespacing prefix used in all class names for our model —, the `cccc` code, which specifies which typographic modulation to use or animation setting to define, and the `nnnn` code, which sets its numeric value.

The four-letters at `cccc` serve to encode instructions about how to modulate each syllable enclosed in the class. In some cases, this will represent the specific variable font axis to me manipulated. `wght`, for example, stands-in for changes in a variable font's weight axis.[7]

In other cases, the code signals instructions for other typographic attributes or animation instructions. Table 5.1 references six of the codes we used in our prototype.

[7] When plausible, this class name echoed the so-called "design-variation axis tag" set in the variable font itself, also a four-letter code. (Constable and Jacobs, 2018)

Table 5.1: Four-letter codes used to communicate typographic modulation information to the text shaping engine that processes the WebVTT file.

Code	Description
`wght`	Changes to the *font-weight* axis of a variable font.
`slnt`	Changes to the *slant* axis of a variable font.
`ltsp`	*Letter-spacing* changes, which will impact the `css letter-spacing` property of the that encloses the affected syllable.
`topp`	*Baseline shift* changes, which will impact the `css top` property of the that encloses the affected syllable.
`time`	Used for animating the typographic modulation. Defines the start time, relative to the start of the specific cue where it is found, when the animation of that syllable should start.
`dura`	Used for animating the typographic modulation. Defines the duration of the animation of the affected syllable.

5.2.2.2 *Encoding values*

As we are constrained by CSS' identifier naming syntax (Bos et al., 2011), some workarounds had to be found for defining how the third code block of the class name (nnnn) would represent the actual numbers used to control the typographic modulations, which could include positive, negative, and floating-point numbers.

A positive number is easy enough to represent. A font-weight value of 800, for instance, can become: `v_wght-800`. Since two hyphens are allowed, negative numbers are also straightforward: `v_slnt--15` gives `slnt` a value of –15.

Lastly, the decimal point in a floating-point number can not be represented as a period (.), since this would clash with how one can attach multiple classes to a tag. For example, `<c.A1.B2>`[8] identifies the <c> tag with `A1` and `B2` classes. `<c.A1.B2.3>`, on the other hand, has three: `A1`, `B2` and `3` — a syntactically invalid third class name. Our workaround was to replace the period in the number with a letter p. A `v_dura-0p84` code, for example, would set 0.84 (seconds) as the duration of the syllable's animation.

With many typographic modulations combined, the resulting WebVTT file loses its readability somewhat. Applying changes to `wght`, `ltsp`, and `topp` to the line "Para o que se quer, isto basta," from the previous WebVTT example, will give:

```
<c.v_wght-505.v_ltsp-2.v_topp-3>Pa</c><c.v_wght-316.v_ltsp-2.v_topp-2>ra
</c> <c.v_wght-336.v_ltsp-2.v_topp-2>o</c><c.v_wght-695.v_ltsp-0.v_topp-
0>que</c> <c.v_wght-300.v_ltsp-0.v_topp-0>se</c><c.v_wght-580.v_ltsp-5.v
_topp-4>quer,</c> <c.v_wght-673.v_ltsp-2.v_topp-0>is</c><c.v_wght-307.v_
ltsp-1.v_topp-1>to</c> <c.v_wght-428.v_ltsp-3.v_topp-0>bas</c><c.v_wght-
-2133.v_ltsp-1.v_topp-2>ta.</c>
```

[8] Note that in WebVTT, the <c> tag has a leaner class-setting syntax than an `html` equivalent, where for something like `<c.classname>` we would need something along the lines of `<c class="classname">`.

5.2.3 *Rendering the speech-modulated closed-captions*

Our prototype to render the speech-modulated closed-captions was created as a JavaScript that interpreted the WebVTT file, using it to coordinate a web page where an audio file[9] was played in sync with closed-captions.

The basis of our solution came from how Nieskens (2019) proposed to deal with HTML / CSS' variable font inheritance problem. The issue is that the `font-variation-settings` property, which allows one to set specific values for variable font axes, resets to their default values all axes *not* ex-

[9] Originally, we planned on having the actor appear along the closed-captions, but we decided against this since it would introduce too many variables to the test.

plicitly specified. This breaks CSS' predictable cascading behavior, i.e., that styles applied to an element will inherit those from more generic selectors *unless* overriding styles are explicitly set in a more specific selector.

Nieskens proposes using CSS variables. In our case, this implies you would have a `:root` declaration such as:

```
:root {
  --wght: 400;
  --slnt: 0;
  --topp: 0;
  --ltsp: 0;
}
```

setting `--wght`, `--slnt`, `--topp`, `--ltsp` to their default values, which can be then applied to all `<span>` tags (that enclose each syllable) as such:

```
p span {
  font-variation-settings: "wght" var(--wght),
    "slnt" var(--slnt);
  top: calc(var(--topp) * -1px);
  letter-spacing: calc(var(--ltsp) * 1px);
}
```

Note that only two of these variables were actual variable font axes (`--wght` and `--slnt`), with `--ltsp` controlling `letter-spacing` changes and `--topp` changes in the `<span>`'s `top` position. Also note that, for these two CSS properties, a unit had be added to the variable's value — in our current approach, the WebVTT file encodes numbers but not units.

5.2.3.1 *A note about the* `<track>` *element*

We soon found that the `<track>` element, which reads the WebVTT file and shows its cues in sync with its parent `<video>` element, does not allow for our proposed CSS variable manipulation. As a workaround for this limitation, we had to hide the `<track>` element and only use its timing events, which control when text goes on and off-screen, to manually sync a custom text element where the closed-captions are set.

5.2.3.2 *Syncing changes*

With this setup, what the JavaScript will do is that, for each cue, a timed function will be set using the `time` class' value for each syllable. When it is called, it will set — and only for the scope of the `<span>` that encloses that particular syllable — the values of these four `CSS` variables to their corresponding values read from the WebVTT file, with a `transion-duration` property set to the value set in the `dura` class. An `ease-out` transition timing function was used, which makes properties change fast at the beginning of the transition and slow at the end.

When each `timeout` function is called, it sets another timed function. When this is called, it returns the syllable's to its default values by resetting the changed `CSS` variables in the syllable's scope. In our testing, we found that this was necessary especially for the `--topp` `CSS` variable, which controls the baseline shift property. As a participant of a previous evaluation had noted, this particular typographic modulation weakens legibility. Having it return to a resting state after the syllable was sounded seemed to weaken this negative effect.

Figure 5.7 demonstrates how these changes occur in time.

5.3 *Concluding remarks*

The model we propose here allows for the representation of three important prosodic features through typographic modulations. Not only that, it allows for these modulations to be displayed synchronously with speech, paving the way for the model's use in speech-modulated closed-captions.

While the model draws from what we have learned in our previous experiments, its implementation leaves space for many creative decisions, left to designers' and developers' subjective evaluations. This is not to be seen as a negative quality, but rather an opening for the creative exploration of how the model will unfold itself in different contexts. Indeed, many of our own implementation decisions came from our subjective evaluations of how to create readable but graphically potent typographic images considering both a static and a closed-captioning context.

There is, of course, the question of whether these modulated typographic images will indeed serve to communicate the prosodic features they seek to represent, which we evaluated empirically in the experiment discussed in the next chapter.

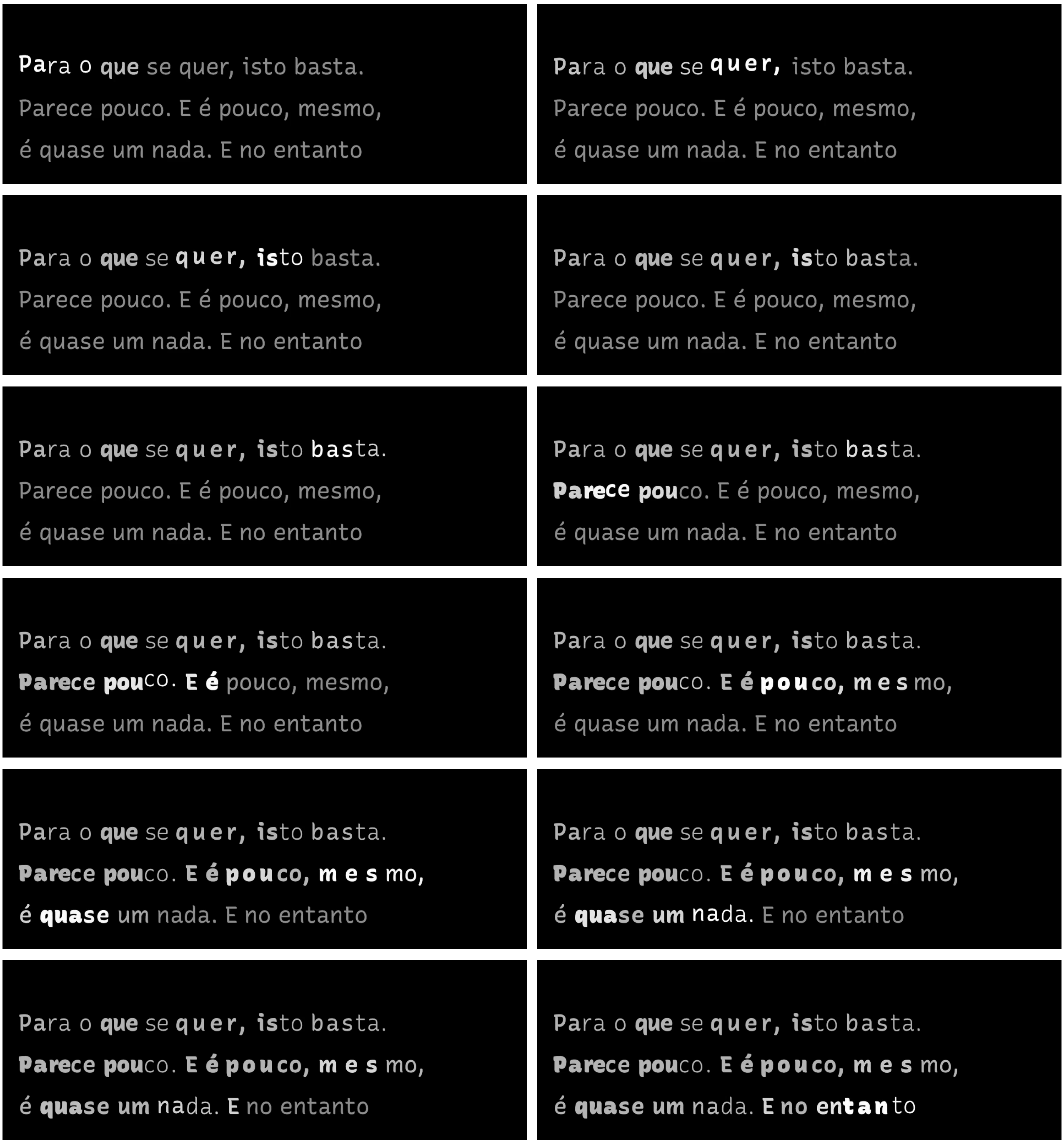


Fig. 5.7: Screenshots of speech-modulated closed-captions changing in time, each image captured roughly one second apart.

Chapter 6

Can prosody be inferred from speech-modulated typography?

Having distilled the results of the previous experiments into the speech-modulated typography model and implementation discussed in Chapter 5, we ran a third experiment that sought to validate the strength of the choices we took when creating said model. Measuring whether the model allows participants to distinguish which of two similar typographic instances corresponded to a specific audio utterance would help respond our first research question,[1] while also giving evidence for how well our model answered research question number two.[2]

In this chapter, we will discuss the objectives and preparation steps for experiment #3 — which involved creating a speech corpus of poetry readings and overhauling the online platform we had used on the previous experiment —, its results, and how we interpreted them.

[1] Can readers of speech-modulated typography recognize prosodic features of its originating audio but *not* present in its textual content?

[2] What typographic modulations can be generally recognized as visual proxies for the prosodic features of magnitude, pitch, and duration?

6.1 *Method*

If in experiment #1 participants attached emotional labels to different instances of speech-modulated typography, and in experiment #2 we measured how frequently they matched certain typographic modulations with different prosodic patterns, we designed experiment #3 as a way to measure *how well* participants would discern which of two similarly looking options corresponded to specific audio utterances.

Again an online experiment, the idea was to have participants be first exposed to an audio file with an expressive reading of a text. After it finished playing, they would be shown two instances of speech-modulated typography. Both represented the same text, but only one had its modulations created from the audio they had just heard.

The rate at which participants chose the correct image would give us hints of how easily understandable[3] our model was in practice. Also, the experiment would be able to measure the model's effectiveness both considering static images and animated closed-captions — while a significant difference between the two options was not expected, if it did emerge it could point to future inquiries.

[3] As with previous experiments, before the test participants would be given a vague overview of our algorithm, hopefully giving enough instructions and sparking interest without overly influencing their interpretations of the model.

6.1.1 *Creating a speech corpus*

After repeatedly using the same speech corpus for both past experiments, we now saw the need for some changes. For one, we would be evaluating the model's behavior not only when applied to static images, but also in the form of animated, speech-modulated closed-captions. This called for more natural sounding, longer phrases.

Despite their expressiveness, the previous phrases' short-bursts-of-nonsense structure gave them an unnatural quality which now felt as a hindrance: while the experiment would not yet simulate any real-life scenarios where speech-modulated typography could be used, it was a step in that direction. Since what we were measuring was related more to sound/visual perception than to sound/text semantics, the text being understandable would not muddy our results, and, inversely, nonsensical texts could be harder to follow.

We were also interested in working with a speech corpus in which the creative-direction had been of creating prosodically-divergent variations of the same texts, rather than a focus on interpreting labeled emotions.

So, we set out to create a small speech corpus for the experiment.

6.1.1.1 *Selecting four poems*

We opted to record a set of poetry readings. Poetry is a form of oral speech that accommodates well an exaggerated prosodic interpretation, as shown by the promising visual results of Castro et al. (2019b). It would have to be recorded because, while there are many examples of readily available, professionally recorded poetry readings, in Portuguese or otherwise, none matched our specific need for a set of versions of the same poem read differently, but not *too* differently (as we will see).

We defined two main criteria for our choice of poems. First, the stanzas had to be short, ideally with between three and four lines. We had planned an experiment with one stanza per round. In it, participants would play

the audio and watch the two animated closed-captions' videos, totaling no more than 30 seconds per round.[4] A 4-line stanza fit that bill comfortably.

The second criterion was that we needed a total of around 15 stanzas, again because of the total time the experiment should take. They need not be necessarily from the same poem, because all stanzas would be shuffled anyway (a control for participant's initial ineptitude with the test and eventual tiredness at the end).

We selected four poems by Brazilian poet Paulo Henriques Britto (read them in Appendix D).

[4] As in experiment #2, we had in mind that a comfortable test would last at most 20 minutes, and preferably less. 30 seconds would be a minimum duration per round if each file played only once, but we accounted for the fact that in the previous experiment participants played each audio multiple times.

6.1.1.2 *Recording the poems while guiding their prosodic patterns*

To record the poems we hired a vocal actor. They read each of the four poems 11 times. The first two times were free-form readings where the actor could freely inflect their voice according to what felt natural within each poems' structure. The following six readings emphasized the three prosodic features we use in our model. Readings 3 and 4 had thus a focus on exaggerating the magnitude, 5 and 6 on the pitch, and 7 and 8 on the rhythm.

The following readings 9 and 10 were again free-form, as we found that, after the more specific instructions for readings 3 to 8, some of the exaggerated ideas that emerged would find more natural ways of mixing themselves in the utterance. The last reading was an attempt at a neutral, monotonous tone of voice.

6.1.1.3 *Excerpt selection and manipulation*

Having 11 recorded takes for each poem, we set out to create the digital audio files that would be used in the experiment, two for each stanza. Participants would only be exposed to one instance of speech-modulated typography per stanza. It would either be either generated from audio A or audio B, but, regardless of this, both versions of the stanza's recording would be available in each round.

A prototype version of the experiment was run with around 10 colleagues from our lab, and in it we found that they tried to manually synchronize the closed-caption video file (which contained no sound) with the audio files.[5] In doing so, they were trying to find the pairing between the two files not through a match between prosody and typography, but by searching for files of the same duration. To counter this effect, we realized that the different versions for each stanza had to have roughly the same duration.

To build these artificial readings, we opened all recorded takes and, in

[5] In so doing, they were echoing a behavior Rosenberger and MacNeil (1999) had seen in their similar experiment: "We observed that people paid especial attention to the similarity of rhythm and pacing across the mediums. As a result, any inconsistencies in computer animation speed will negatively affect people's ability to understand the prosody correctly."

the multi-track interface of the Adobe Audition software, divided the files into blocks, each corresponding to one of the poem's lines (see Figure 6.1).

Fig. 6.1: Screenshot of Adobe Audition's multi-track interface. Each row contains one take in our recording session, with each colored block corresponding to one verse of the poem. While they are not aligned here, these blocks could be freely manipulated to construct a file of arbitrary durations.

This allowed us to align each of the poem's lines between different takes. We selected and shifted blocks to create a poetic reading where we maximized prosodic differences while controlling for the length of their silences, making resulting files of roughly the same duration.

We ended with 30 audio files. We divided their syllables and vowels in a TextGrid file, then processed to create the WebVTT files used to generate the static and animated speech-modulated typography for the test. These images and the links to the used videos can be seen in Appendix C.

6.1.1.4 *How we created the video files used in the test*

A small note about how the videos were generated. As stated, our Javascript interprets each cue in the closed-caption file, generating a series of timed functions that coordinate when each typographic modulations starts, with animation tweening controlled by CSS transitions, and another set of functions controlling their return to resting states. Especially with complex cues — which the poems, with up to 4 lines per cue, inevitably generated —, the sequences became rather processor-hungry. This could be felt particularly when we tested the system in older smartphones, where animations would stutter and fall out of sync with the audio.

Since the test would inevitably be run on many kinds of devices, including slower ones, we decided that the videos should be pre-rendered and made available not as a dynamic process running on participants'

browser but as a simple YouTube video. This would also bring the benefit of sidestepping the issue of testing and debugging the browser compatibility of our Javascript and CSS features[6] — for this first version of our model, they would only have to run once, in our own controlled environment.

To ensure that animations ran smoothly, we created special versions of the WebVTT files where the duration of typographic changes was slowed by a factor of 4. We screen-captured these slow-motion animations, where eventual performance hiccups were made imperceptible, and the results were then sped up by a factor of $1/4$.

[6] Although per *Can I Use* (`https://caniuse.com/variable-fonts`, accessed on December 28, 2020), the most cutting-edge feature we have used, which are variable fonts, is as of this writing already supported by 93% of browsers.

6.1.2 *The experiment's online platform*

The online platform for the experiment was similar to experiment 2's, with some key differences. First, we would have two simultaneous versions of the same experiment running: one with static images, the other with animated closed-captions. Participants would be randomly assigned to one version when entering the site, and all subsequent instructions and rounds reflected this path.

A second difference was in the structure of the test itself. As shown in Figure 6.2, where we once had one audio file and two images, we now had one video (or image), two audio files and a Likert scale measuring how strong participants felt the relation between the shown typographic instance and the selected audio was.

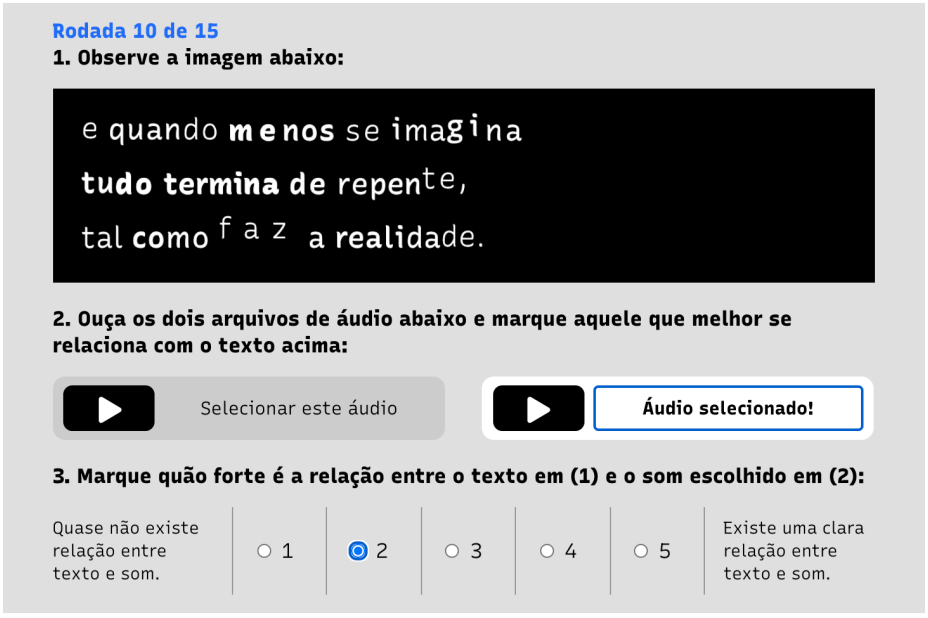


Fig. 6.2: Screenshot of the interface of one round in the static-typography version of the experiment. On top, the speech-modulated typography itself (step 1), followed by the two audio players (step 2) and the Likert scale (step 3). Translated labels are: *Round 10 of 15; 1. Examine the image below; 2. Listen to the two audio files below and indicate that which best matches the text above; 3. Indicate how strong is the relation between the text in (1) and the chosen audio in (2).*

There were also a long list of small usability and cosmetic refinements

made considering what we had learned in experiment #2 and reflecting the new speech-modulated typography model. The platform was designed in the same typeface[7] as the typographic choices in the test itself; we made it clearer at each step how many steps remained until the end of the experiment;[8] the steps to peek into the experiments' partial results were made simpler; among other changes.

[7] The five-axis *Recursive* typeface, designed by Arrow Type's Stephen Nixon.

[8] Still, 19% of participants that actually started the test abandoned it before finishing, similar to experiment #2's rate of 16%.

6.2 *Results*

The experiment was conducted from November to December 2020. Email invitations were sent to students in our department, and open invitations were posted on Twitter and LinkedIn.

Altogether, 117 participants concluded the test. Their demographics were similar to those in the second experiment: 47% were female, 51% male and 2% non-binary. 90% had undergraduate degrees or higher. 25% had between 18 and 24 years, 49% between 25 and 39, 23% between 40 and 59, and 3% had 60 or more years. 52% of participants were assigned the static-image based test, and 48% the animated closed-captions variant.

6.2.1 *How frequently did participants chose the right choice?*

There were slight differences in how well participants did between the two versions of the test. For the static-image one, participants chose the correct typographic choice in 67% of their answers. Figure 6.3 shows how well each of these 61 participants did.

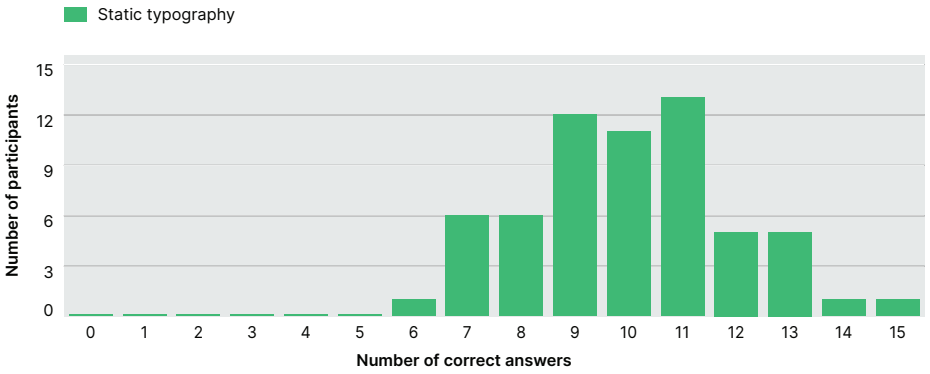


Fig. 6.3: Distribution of performances of the 61 participants who did the static-image based version of the experiment. 0 on the x-axis would represent participants who got all answers wrong, while 15 meant all answers were correct.

For the animated closed-captions based version of the experiment, participants chose the correct typographic choice for 63% of their answers. Figure 6.4 shows how well each of these 56 participants did.

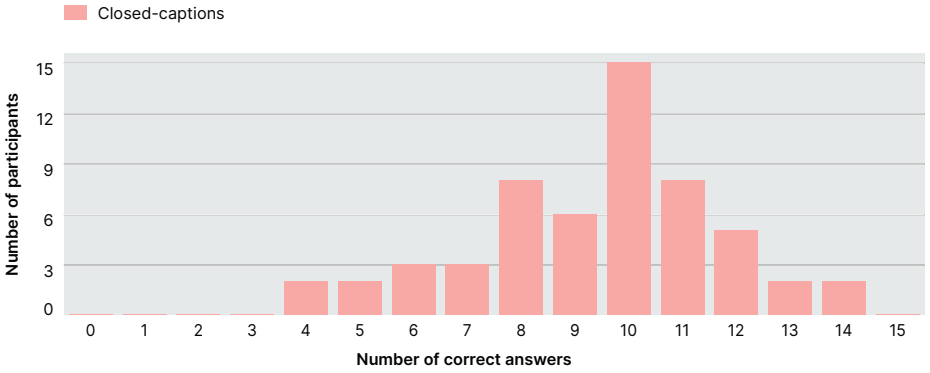


Fig. 6.4: Distribution of performances of the 56 participants who did the animated closed-captions based version of the experiment. 0 on the x-axis would represent participants who got all answers wrong, while 15 would mean all answers were correct.

6.2.2 *Likert scale*

Along each choice, participants also assigned a value between 1 and 5 to how strong they found the relation of typography with sound. The difference between average Likert value for correct answers ($M = 3.89$, $SD = 0.9$) versus incorrect answers ($M = 3.76$, $SD = 0.95$) was statistically significant.[9]

Comparing average Likert values only for the static-image version of the test gives us that the difference between correct ($M = 3.84$, $SD = 0.91$) and incorrect ($M = 3.67$, $SD = 0.93$) answers was statistically significant.[10]

For the animated closed-caption version, average Likert values for correct ($M = 3.95$, $SD = 0.88$) and incorrect ($M = 3.84$, $SD = 0.96$) answers were not statistically different.[11]

[9] $t(1,753) = 2.95$, $p < 0.05$.

[10] $t(913) = 2.650$, $p < 0.05$.

[11] $t(838) = 1.687$, $p > 0.05$.

6.2.3 *Open-ended comments*

Roughly one-third of participants sent comments at the end of the test. Some were complimentary remarks about the research as a whole and its potential applications in different realms. Others shared their interpretations of the modulations, suggestions for the model, etc.

6.2.3.1 *How did participants interpret the model's modulations?*

Many participants understood how the three typographic modulations related to the three prosodic features. There was also complaints against baseline shift's behavior — which some participants thought was at times erratic and not always clearly related to audio, e.g., one participant thought it seemed to represent only high pitches and not lower ones.

There was mention of an apparent discrepancy between how strong the typographic changes *looked* versus how strong the prosody *sounded* like. One person wrote that reading the typography seemed to have a much

smaller emotion than the audios. Another argued that large changes in pitch and loudness that sounded natural at the end of phrases looked strange in our model's typographic interpretation.

6.2.3.2 *Suggestion for new typographic modulations*

One participant suggested that many additional dimensions in speech should be represented in typography, especially regarding vocal timbre. Missing in our model, a rough voice could have pointy borders, while a calm one should be softer. They wondered: how would we represent voices that were feminine, child-like, stuttering, nasal, foreigner, illiterate, etc?

Some participants thought that the typography should be more expressive, like the lettering in comic-books.[12] One person suggested using changes in lower and upper case.

[12] It is worth mentioning that at the experiment's presentation text we showed some images of expressive lettering in comic-books as an analog (in spirit) to how our research involved the expressive manipulation of typography.

6.2.3.3 *Legibility and readability concerns*

Some participants felt that letter-spacing changes made the space between words hard to tell apart. One particular participant, who mentioned being dyslexic, said that the animated typographic-modulations were difficult to read and the video had to be played many times before its text could be understood.

6.3 *Discussion*

The measured performance was strong enough to safely state that participants were able to deduce how sound was being translated into typographic modulations. While there is a four percentage points difference in favor of performances for the static-image based test ($M = .668$, $SD = .129$) versus the animated closed-captions one ($M = .627$, $SD = .154$), seen in Figure 6.5, it is not significant[13] and could be due to chance.

[13] $t(115) = 1.54$, $p > 0.05$.

The results, particularly participants' comments, point to some limits of our use of perceptual evaluations. We created in this test a scenario that suggested (but was not equal to) a real-world use case where we imagine speech-modulated typography could be used. While this gave our participants hints about how the model could be used, the experimental setup nevertheless stimulated in them unrealistic attention towards minute typographic and sonic details that we would not expect in real-world scenarios.

In other words, while we successfully measured the recognition of prosody in typography, the experiment's artificiality limits its use towards

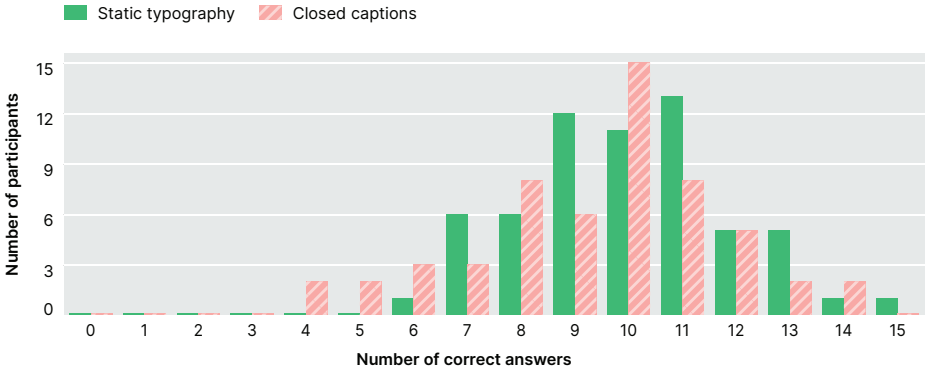


Fig. 6.5: Chart comparing the performance distributions between both versions of the experiment. Static typography, in solid-green, is slightly better.

understanding how speech-modulated typography could function in concrete scenarios. This is not to say that the experiment was inadequate — which it was not —, but rather to point some limits into what our measured average of 65% correct answers tells us.

This value is a good indicator that our model's proposed typographic mappings are efficient in encoding prosody,[14] and that as such it allowed readers to recognize prosodic patterns in typography.[15] What it does not tell us is *how* this speech-modulated typography changes readers' perceptions about what a text means. The model's effectiveness, however, gives support for future studies that investigate these further dimensions not captured by our current methods.

[14] Related to our second research question.

[15] Related to our first research question.

6.4 *Conclusion*

We have successfully created a performance baseline for the interpretation of prosodic features shown as typographic modulations, both in static-image and animated closed-captions contexts. Future versions of the model could be compared to its current implementation in terms of how they impact this performance.

While adjustments to the model can still be made to boost its prosody-recognition performance, the current experiment shows it to be in a robust enough state to be used in investigations into other aspects of speech-modulated typography, particularly related to how it can change the experiencing of media in which it is used.

Chapter 7

Conclusion

In this work, we contributed to the field of *visual design* with an assemblage of methodological approaches to express and measure the perceptual relationships between sound and the typographic form; to the field of *affective computing*, by illustrating some short-comings of the use of a priori categorical models of emotion to classify subjective phenomena; and lastly, by developing and evaluating a model to process and represent a dynamic input such as is speech, we bring contributions to the field of *human-computer interaction*, which is also enriched by how speech-modulated typography can become a building block for affect-sensitive text interfaces.

The primary goal of our research was to design, implement, and evaluate a speech-modulated typography model. And so we did, bringing new insights and knowledge in how we (1) selected, extracted, and processed a set of prosodic features; (2) used the modulation of variable font axes and other typographic attributes to represent sound; and (3) developed an experimental approach to inform the development and allow the evaluation of our model.

Chapter 3 presented the first version of our speech-modulated typography model, inspired by our literature review. Participants' responses to the model were largely inconsistent and, if there was a prosody-recognition effect, our experimental setup could not measure it. This showed us that, regardless of the possible inadequacy of the card-sorting method for our particular purpose, the model would need to be investigated with greater attention to its constituent parts.

In Chapter 4, we describe how we did so. An online test platform allowed us to investigate how each typographic modulation could be associated with a different prosodic feature. We managed to obtain strong indications of certain prosodic-typographic associations, which would later inspire a revamped speech-modulated typography model. This experiment also led

us to reevaluate our reliance on categorical emotion models as a way to measure our model's successes and failures in capturing prosody.

These lessons fed into our updated model, which we describe in Chapter 5. In it, we managed to combine the lessons from our previous experiments with what other researchers have been doing. In this chapter, we describe how our model extracts and processes prosody, which it uses to modulate a set of typographic parameters to create speech-modulated typography. We also present how we implemented a pilot version of the model, discussing our approach's advantages — backward compatibility, relative independence between prosody extraction and typographic modulation, the flexibility of typographic modulations, etc —, and drawbacks — low performance, some steps of the process are still heavily reliant on manual inputs, etc.

Lastly, this model was put to a test, which we describe in Chapter 6. This finally gave robust answers to our first[1] and second[2] research questions. Our results validate how our model encodes prosody through typography, and serve as a stepping stone towards future research that can begin to investigate the effects speech-modulated typography can have on readers.

[1] Can readers of speech-modulated typography recognize prosodic features of its originating audio but *not* present in its textual content?

[2] What typographic modulations can be generally recognized as visual proxies for the prosodic features of magnitude, pitch, and duration?

7.1 *Main contributions*

In summary, our work's main contributions are:

- Evidence that a categorical-emotion model is not an adequate measurement proxy to investigate the mapping of prosodic features as typographic modulations.
- Robust evidence supporting the use of font-weight as a representation of magnitude, and baseline shift as a representation of pitch.
- A speech-modulated typography model which was empirically validated both in its use to generate static images and speech-modulated closed-captions.

Additionally, some technical developments are worth noting:

- A method to annotate, extract and process prosodic features at a vowel level, receiving as input a sound-file with a speech utterance and outputting a JSON file with an array of normalized features per syllable.
- A method to use said JSON file to create a WebVTT closed-captions file that specifies how to modulate typography in tandem with speech.

- A method to use said WebVTT file to output animated speech-modulated closed-captions, which can be played in any modern browser synchronously with a film or audio file.
- Two web-frameworks that can be repurposed for perceptual tests similar to the ones we ran in experiments #2 and #3.

7.2 *Future work*

From where we stand, some enhancements to the model can be envisioned. It is also important to investigate the effects of speech-modulated typography on different audiences. Both will be discussed in this session.

7.2.1 *Automation*

In terms of enhancements, the automation of the processes involved in the syllabic-segmentation of audio utterances should be explored, echoing the suggestions of Castro et al. (2019b) and Rosenberger-Shankar (1998). Even if a general-purpose solution is improbable in the short-run, approaches that tackle specific, simpler scenarios — e.g., poetry readings, monophonic vocal music, stand-up comedy, etc — would be important if not only because they would help shorten the duration of development cycles.

7.2.2 *Design explorations*

This is important when we consider how a second front for further research is found in the refinement of our audio-visual translation model. This could be subdivided into, first, an exploratory study of alternative combinations of typographic modulations and prosodic features, i.e., different axes and typographic parameters could be experienced. It is worth noting that the axes we used to represent prosodic features were not originally created by typographers for such uses. Italic, bold, thin, condensed, etc, are typographic aspects generally used for the highlighting of passages, and *not* for echoing the expressive quirks of the human voice. While our experimental results are discussed here with no regard to this fact, one wonders if a typeface created with prosodic representation as an explicit consideration could potentially produce stronger effects than those we measured.

Secondly, the logic of how prosody affects typography should be better understood. As we tried to make clear in chapter 5, our model contains a set of parameters whose inter-dependencies are not easily optimized. We

opted for a simple, linear translation approach, where changes in prosody are reflected in typography, but this could be better explored, e.g., maybe certain typographic changes should only kick-in after a certain threshold of significance is reached in the prosodic patterns, thus creating a less "noisy" visual image in the outputted typography?

7.2.3 *Effects of Speech-Modulated Typography*

7.2.3.1 *Immersion in film-media*

Lastly, future research should start exploring the effects that speech-modulated typography can have in terms of changes in how a medium is experienced. We foresee two immediate possibilities, but further reflection could point to others. First, using immersion-measuring questionnaires, such as those specifically geared towards closed-captions presented in Kruger et al. (2016), could be used to compare if viewers exposed to a film with speech-modulated closed-captions would present higher levels of immersion than those watching the same film, but traditionally-captioned.

7.2.3.2 *Accessibility*

Secondly, accessibility-focused uses of speech-modulated closed-captions should be explored. This is of course a broad category that accommodates many issues and approaches. Speech-modulated closed-captions could be used, for instance, as a way of highlighting the relative importance of words (in a way, echoing one of prosody's functions in the spoken word). This venue has been explored, for instance, by Kafle et al. (2019). They have shown that the highlighting of words increases readability and understandability for the d/Deaf and hard-of-hearing (DHH) test-subjects when used in online-lecture videos, and have presented a methodology for investigating the efficacy of word-highlighting which could be adapted for speech-modulated closed-captions.

Speech-modulated closed-captions could also be integrated into automatic captioning systems, especially those in the up-and-coming field of head-mounted displays, such as those presented by the design probes of Jain et al. (2018) and Olwal et al. (2020). These devices mediate communication between persons by transcribing speech and displaying its transcription on digital overlays. Considering how important prosody is in conveying meaning, integrating our approach into these devices could increase their emotional resonance.

Bibliography

Bänziger, T. and Scherer, K. R. (2005). The role of intonation in emotional expressions. *Speech Communication*, 46(3-4):252–267.

Barbosa, P. A. (2012). Conhecendo melhor a prosódia: aspectos teóricos e metodológicos daquilo que molda nossa enunciação. *Revista de Estudos da Linguagem*, 20(1):11–27.

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

Bessemans, A., Renckens, M., Bormans, K., Nuyts, E., and Larson, K. (2019). Visual prosody supports reading aloud expressively. *Visible Language*, 53:28–49.

Bigelow, C. (2019). Typeface features and legibility research. *Vision Research*, 165:162–172.

Blythwood (2015). Reverse-contrast "italian" type in an 1828 specimen book by the george bruce company of new york. `https://commons.wikimedia.org/wiki/File:Reverse_contrast.png`. Accessed on December 4, 2020.

Boehner, K., DePaula, R., Dourish, P., and Sengers, P. (2005). Affect. In *Proceedings of the 4th decennial conference on Critical computing between sense and sensibility - CC '05*. ACM Press.

Boersma, P. and Weenink, D. (2020). Praat: doing phonetics by computer [computer program]. version 6.1.10. `https://www.fon.hum.uva.nl/praat/`.

Bos, B., Çelik, T., Hickson, I., and Lie, H. W. (2011). Cascading style sheets level 2 revision 1 (CSS 2.1) specification. `https://www.w3.org/TR/CSS21/syndata.html#characters`. Accessed on December 23, 2020.

Britto, P. (2003). *Macau*. Companhia das Letras, São Paulo, Brazil.

Britto, P. (2007). *Tarde : poemas*. Companhia das Letras, São Paulo.

Castro, J. C. et al. (2019a). «máquina de ouver»—representação do discurso oral pela tipografia. Master's thesis, Universidade de Coimbra. License: Creative Commons BY-NC-ND 4.0.

Castro, J. C. e., Martins, P., Boavida, A., and Machado, P. (2019b). «máquina de ouver»-from sound to type: Finding the visual representation of speech by mapping sound features to typographic variables. In *Proceedings of the 9th International Conference on Digital and Interactive Arts*, pages 1–8.

Chiang, T. (2019). *Exhalation: Stories*. Knopf.

Constable, P. and Jacobs, M. (2018). Opentype design-variation axis tag registry. `https://docs.microsoft.com/en-us/typography/opentype/spec/dvaraxisreg#registered-axis-tags`. Accessed on December 23, 2020.

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

de Lacerda Pataca, C. (2019). Msc 2nd evaluation. `https://github.com/caluap/msc-2nd-evaluation`.

de Lacerda Pataca, C. and Costa, P. D. P. (2019). Tipografia modulada pela fala. In *Proceedings of the 12th edition of the EADCA, Encontro de Alunos e Docentes do DCA*.

de Lacerda Pataca, C. and Costa, P. D. P. (2020). Speech modulated typography: towards an affective representation model. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, pages 139–143.

Deveria, A. (2021). Can i use: Variable-fonts. `https://caniuse.com/variable-fonts`. Accessed on January 6, 2021.

dos Reis, J. (2014). Speechant: Chanting & speeching: Sistema de notação tipográfica para a educação de adultos. *Matéria Prima*, 2(3).

dos Reis, J. and Hazan, V. (2011). Speechant: a vowel notation system to teach english pronunciation. *ELT Journal*, 66(2):156–165.

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

El Ayadi, M., Kamel, M. S., and Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44(3):572–587.

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*.

Henry, S. L. (2020). Does my media need captions? `https://www.w3.org/WAI/media/av/captions/#checklist`. Accessed on January 12, 2021.

Hernández, M. (2016). A tutorial to extract the pitch in speech signals using autocorrelation. *Open Journal of Technology & Engineering Disciplines (OJTED)*, 2:1–11.

House, J. (2006). Constructing a context with intonation. *Journal of pragmatics*, 38(10):1542–1558.

Instituto Pró Livro (2020). Retratos da leitura no brasil. `https://prolivro.org.br/wp-content/uploads/2020/09/5a_edicao_Retratos_da_Leitura_no_Brasil_IPL-compactado.pdf`. Accessed on December 20, 2020.

Jacobs, M. and Constable, P. (2018). Opentype specification version 1.8. `https://docs.microsoft.com/en-us/typography/opentype/otspec180/`. Accessed on January 5, 2021.

Jain, D., Chinh, B., Findlater, L., Kushalnagar, R., and Froehlich, J. (2018). Exploring augmented reality approaches to real-time captioning: A preliminary autoethnographic study. In *Proceedings of the 2018 ACM Conference Companion Publication on Designing Interactive Systems*, pages 7–11.

Johnson, A. (2019). AR optical typography. `https://www.aetherpoint.com/casestudy/AR-optical-typography/`. Accessed on December 6, 2021.

Kafle, S., Yeung, P., and Huenerfauth, M. (2019). Evaluating the benefit of highlighting key words in captions for people who are deaf or hard of hearing. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*, pages 43–55.

Koolagudi, S. G. and Rao, K. S. (2012). Emotion recognition from speech: a review. *International journal of speech technology*, 15(2):99–117.

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).

Lemon, D., Constable, P., Esfahbod, B., Holbrook, N., and Daniels, S. (2016). Atypi 2016 warsaw, special opentype session. `https://youtu.be/6kizDePhcFU`. Accessed on November 1, 2017.

Letters from The Temporary State (2019). Italics. `http://letters.temporarystate.net/entry/4/`. Accessed on September 23, 2019.

Mahrt, T. (2020). Praatio. `https://github.com/timmahrt/praatIO`. Accessed on December 20, 2020.

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

Moraes, J. A. and Rilliard, A. (2016). Prosody and emotion in brazilian portuguese.

Morais, J., Cary, L., Alegria, J., and Bertelson, P. (1979). Does awareness of speech as a sequence of phones arise spontaneously? *Cognition*, 7(4):323–331.

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 *10th Asia Pacific Conference on Computer Human Interaction (Apchi 2012). Matsue-city, Shimane, Japan*, pages 28–31.

Nieskens, R. (2019). Boiling eggs and fixing the variable font inheritance problem. `https://pixelambacht.nl/2019/fixing-variable-font-inheritance/`. Accessed on December 23, 2020.

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.

Olwal, A., Balke, K., Votintcev, D., Starner, T., Conn, P., Chinh, B., and Corda, B. (2020). Wearable subtitles: Augmenting spoken communication with

lightweight eyewear for all-day captioning. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*, pages 1108–1120.

Paige, D. D., Rupley, W. H., Smith, G. S., Rasinski, T. V., Nichols, W., and Magpuri-Lavell, T. (2017). Is prosodic reading a strategy for comprehension? *Journal for educational research online*, 9(2):245–275.

Pataca, C. de L.. and Costa, P. D. P. (2019). Tipografia modulada pela fala: avaliação de um algoritmo de geração de prosódia visual em textos. In *Anais do 9º CIDI | Congresso Internacional de Design da Informação, edição 2019 e do 9º CONGIC | Congresso Nacional de Iniciação Científica em Design da Informação*. Editora Blucher.

Ramteke, P. B. and Koolagudi, S. G. (2019). Phoneme boundary detection from speech: A rule based approach. *Speech Communication*, 107:1–17.

Rao, K. S., Reddy, R., Maity, S., and Koolagudi, S. G. (2010). Characterization of emotions using the dynamics of prosodic features. *Speech Prosody*, page 4.

Rosenberger, T. and MacNeil, R. L. (1999). Prosodic font: translating speech into graphics. In *CHI'99 Extended Abstracts on Human Factors in Computing Systems*, pages 252–253.

Rosenberger-Shankar, T. (1998). Prosodic font: The space between the spoken and the written. Master's thesis, Massachusetts Institute of Technology.

Schötz, S. (2002). Linguistic & paralinguistic phonetic variation in speaker recognition & text-to-speech synthesis. In *GSLT Papers: Speech Technology 1.*

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.

Sherman, N. (2020). Variable fonts support. `https://v-fonts.com/support/`. Accessed on January 6, 2021.

Silva, W. d., Barbosa, 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.

Stark, L. and Hoey, J. (2020). The ethics of emotion in ai systems. Accessed on December 19, 2020.

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.

Van Orden, G. C. (1987). A ROWS is a ROSE: Spelling, sound, and reading. *Memory & cognition*, 15(3):181–198.

Verbaenen, W. (2019). Phonotype. the visual identity of a language according to its phonology. Master's thesis, PXL-MAD.

Wikimedia Commons contributors (2018). Codex vaticanus, matthew 1:22-2:18. `https://commons.wikimedia.org/w/index.php?title=File:Codex_Vaticanus_Matthew_1,22-2,18.jpg`. Accessed on July 28, 2018.

Wilson, D. and Wharton, T. (2006). Relevance and prosody. *Journal of pragmatics*, 38(10):1559–1579.

Wölfel, M., Schlippe, T., and Stitz, A. (2015). Voice driven type design. In *2015 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, pages 1–9.

Appendices

Appendix A

Cards used in experiment #1

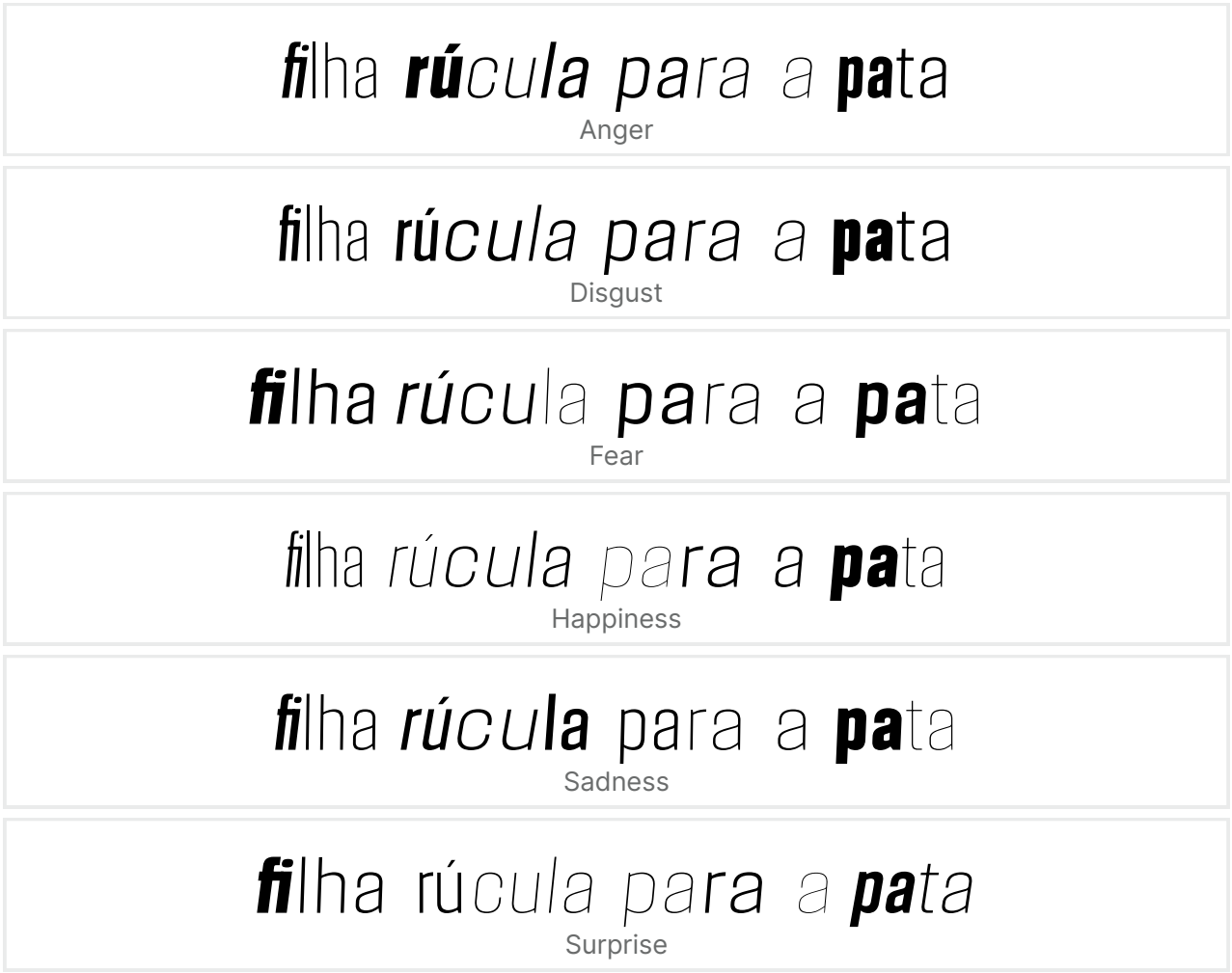


Fig. A.1: Six instances of the *Filha…* phrase, as used in experiment #1.

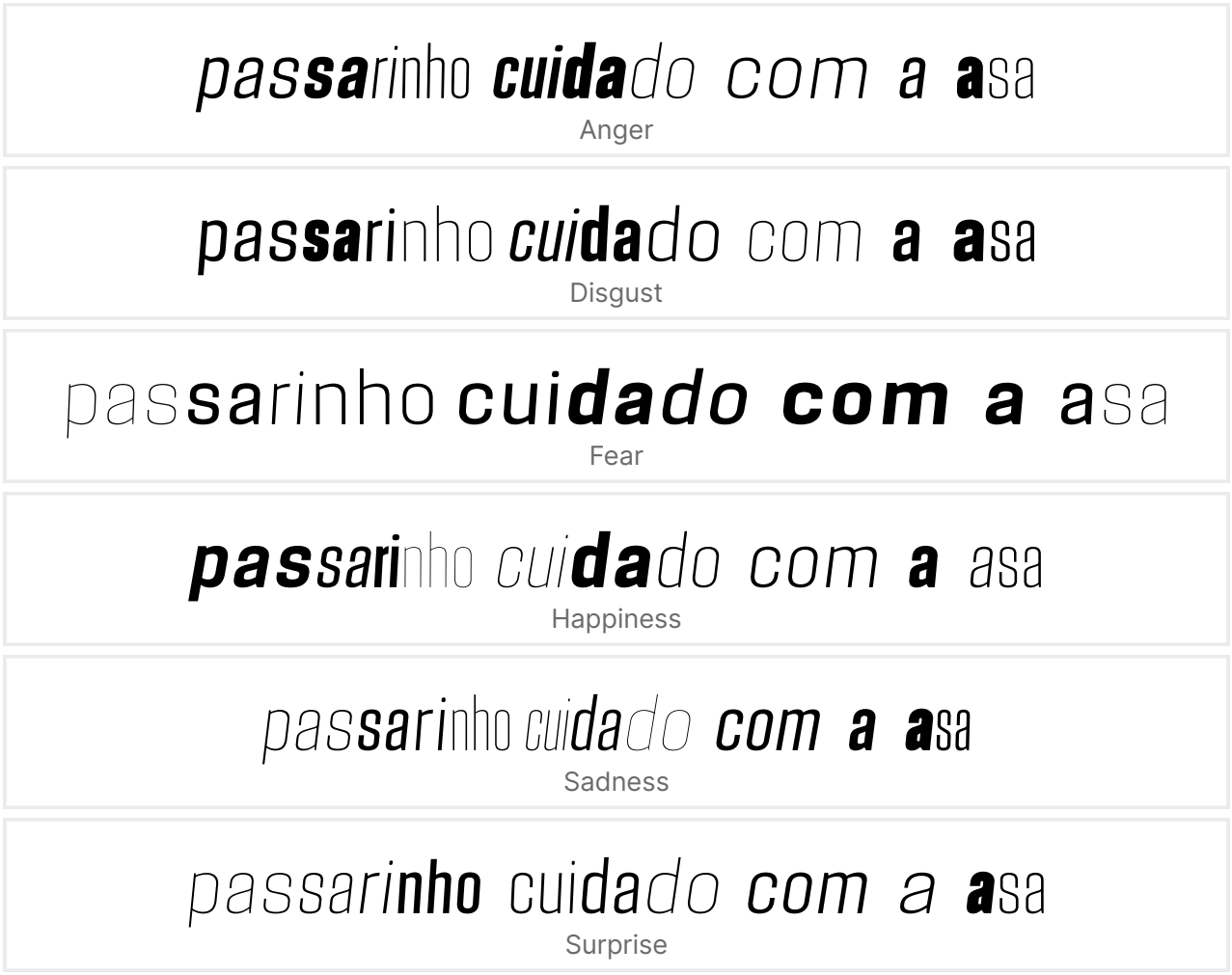


Fig. A.2: Six instances of the *Passarinho…* phrase, as used in experiment #1.

Fig. A.3: Six instances of the *Lilo…* phrase, as used in experiment #1.

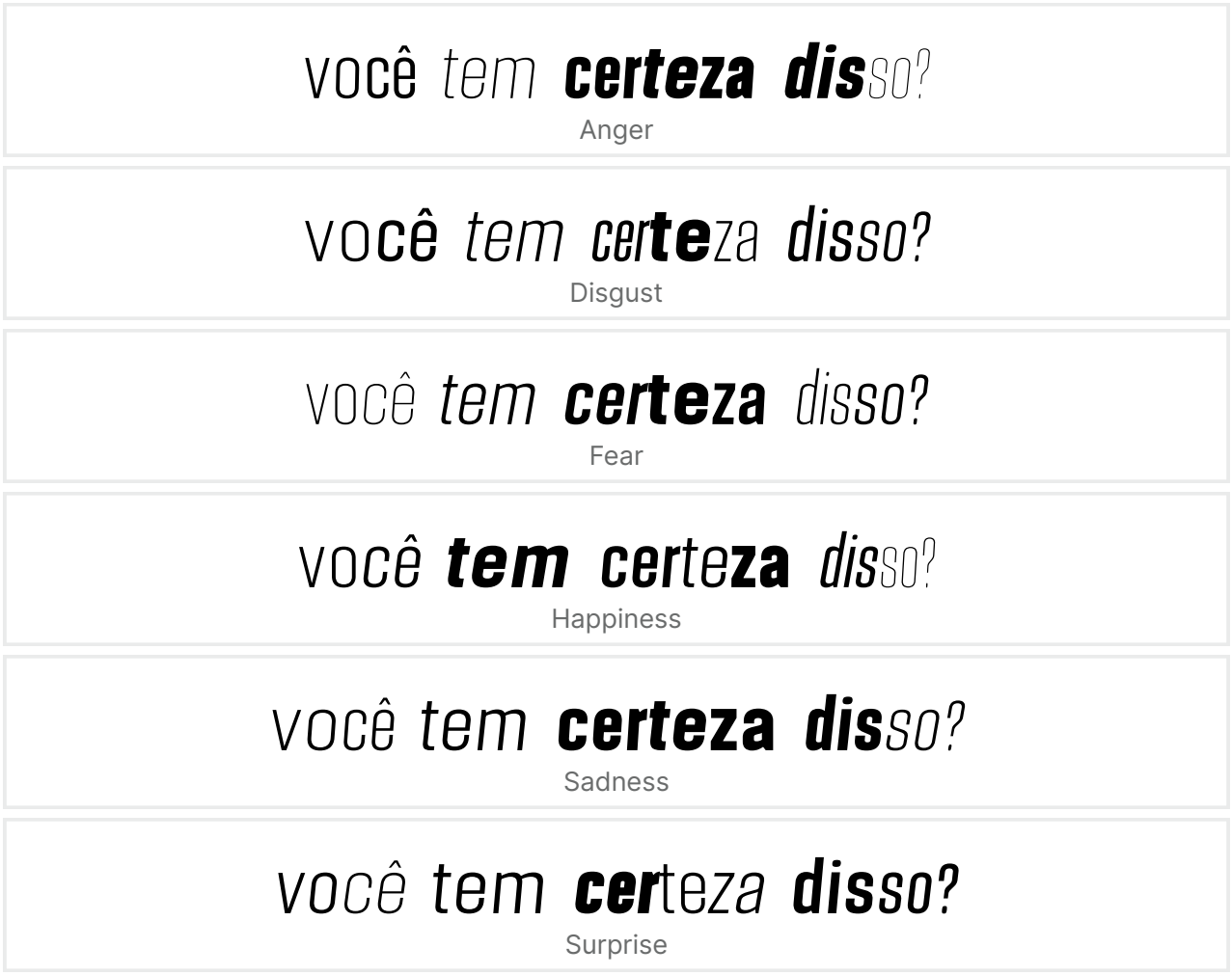


Fig. A.4: Six instances of the *Você...* phrase, as used in experiment #1.

Appendix B

Images used in experiment #2

Listed in the following pages are the sets of images participants were shown in experiment #2. We have grouped the images in blocks of four. In each block, any one image was always shown paired with one of the other three.

Figures B.1 to B.10 were created using the extracted *amplitude* from the recorded utterances, while Figures B.11 to B.20 were created using the extracted *pitch* from the same recorded utterances.

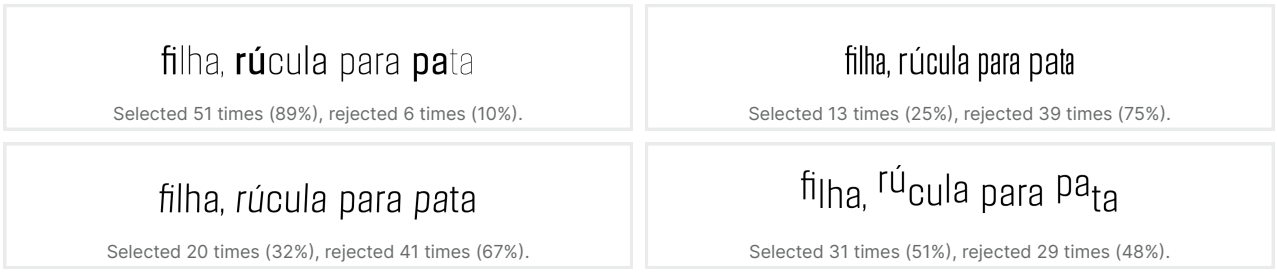


Fig. B.1: Four ways to represent an angry utterance's amplitude.

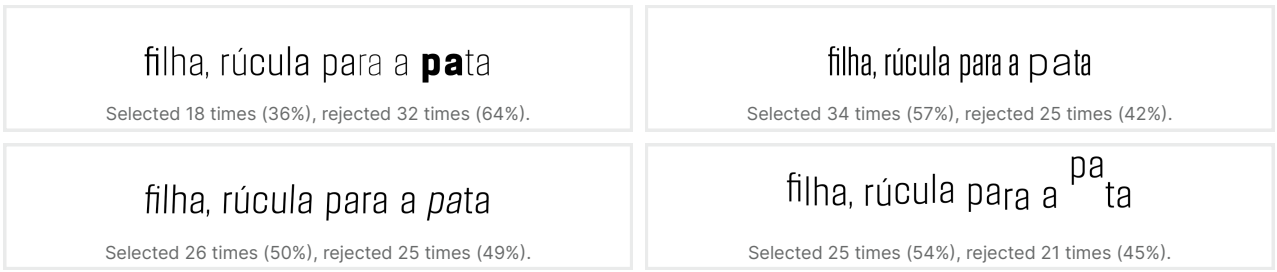


Fig. B.2: Four ways to represent a happy utterance's amplitude.

filha, rúcula para a **pa**ta

Selected 28 times (43%), rejected 37 times (56%).

filha, rúcula para a pata

Selected 23 times (42%), rejected 31 times (57%).

filha, rúcula para a *pata*

Selected 57 times (83%), rejected 11 times (16%).

filha, rúcula para a pata

Selected 16 times (26%), rejected 45 times (73%).

Fig. B.3: Four ways to represent a neutral utterance's amplitude.

filha, rúcula para a **pata**

Selected 27 times (56%), rejected 21 times (43%).

filha, rúcula para a pata

Selected 23 times (45%), rejected 28 times (54%).

filha, rúcula para a *pata*

Selected 21 times (45%), rejected 25 times (54%).

filha, rúcula para a pata

Selected 26 times (53%), rejected 23 times (46%).

Fig. B.4: Four ways to represent a sad utterance's amplitude.

filha, rúcula para **pa**ta

Selected 43 times (84%), rejected 8 times (15%).

filha, rúcula para pata

Selected 19 times (30%), rejected 43 times (69%).

*fi*lha, rúcula para *pa*ta

Selected 23 times (37%), rejected 39 times (62%).

filha, rúcula para pata

Selected 25 times (55%), rejected 20 times (44%).

Fig. B.5: Four ways to represent a surprised utterance's amplitude.

pas**sa**rinho, **cuida**do com **a** asa

Selected 50 times (76%), rejected 15 times (23%).

passarinho, cuidado com a asa

Selected 28 times (45%), rejected 34 times (54%).

*passa*rinho, *cuida*do com *a* asa

Selected 31 times (51%), rejected 29 times (48%).

passarinho, cuidado com a asa

Selected 15 times (24%), rejected 46 times (75%).

Fig. B.6: Four ways to represent an angry utterance's amplitude.

passari**nho**, cui**da**do com **a** asa

Selected 26 times (38%), rejected 41 times (61%).

passarinho, cuidado com a asa

Selected 31 times (51%), rejected 29 times (48%).

*pas*sari*nho*, cui*da*do com *a* asa

Selected 24 times (42%), rejected 33 times (57%).

pas$_{\text{sari}}$nho, cui$^{\text{da}}$do com $^{\text{a}}$ asa

Selected 46 times (65%), rejected 24 times (34%).

Fig. B.7: Four ways to represent a happy utterance's amplitude.

pas**sa**ri**nho**, cui**da**do com **a a**sa

Selected 26 times (54%), rejected 22 times (45%).

passarinho, cuidado com a asa

Selected 30 times (61%), rejected 19 times (38%).

*pas*sa*rinho*, cui*da*do com *a a*sa

Selected 30 times (61%), rejected 19 times (38%).

pas$_{\text{sa}}$$^{\text{ri}}nho, cui^{\text{da}}$do com $^{\text{a}}$ $^{\text{a}}$sa

Selected 15 times (26%), rejected 41 times (73%).

Fig. B.8: Four ways to represent a neutral utterance's amplitude.

pas**sa**rinho, cui**da**do **com a a**sa

Selected 21 times (38%), rejected 33 times (61%).

passarinho, cuidado com a asa

Selected 37 times (63%), rejected 21 times (36%).

passarinho, cuidado *com a asa*

Selected 24 times (39%), rejected 37 times (60%).

pas$^{\text{sa}}$rinho, cui$^{\text{da}}$do com $^{\text{a}}$ $^{\text{a}}$sa

Selected 33 times (57%), rejected 24 times (42%).

Fig. B.9: Four ways to represent a sad utterance's amplitude.

Fig. B.10: Four ways to represent a surprised utterance's amplitude.

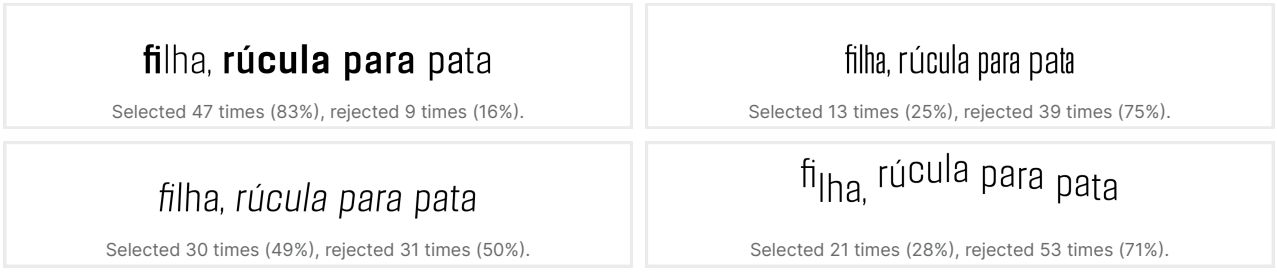


Fig. B.11: Four ways to represent an angry utterance's pitch.

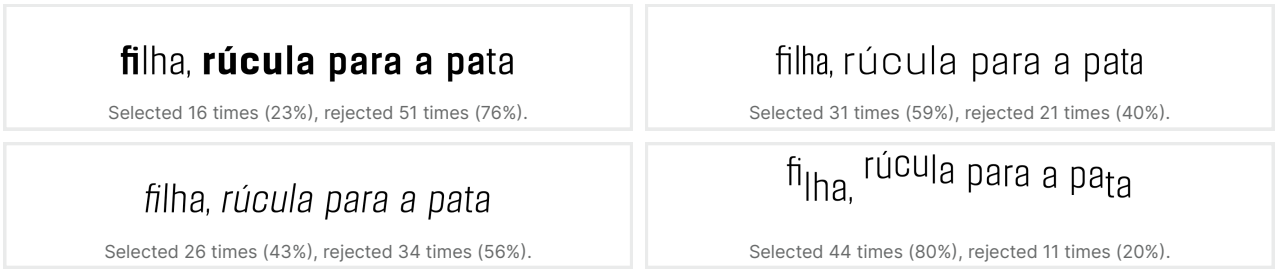


Fig. B.12: Four ways to represent a happy utterance's pitch.

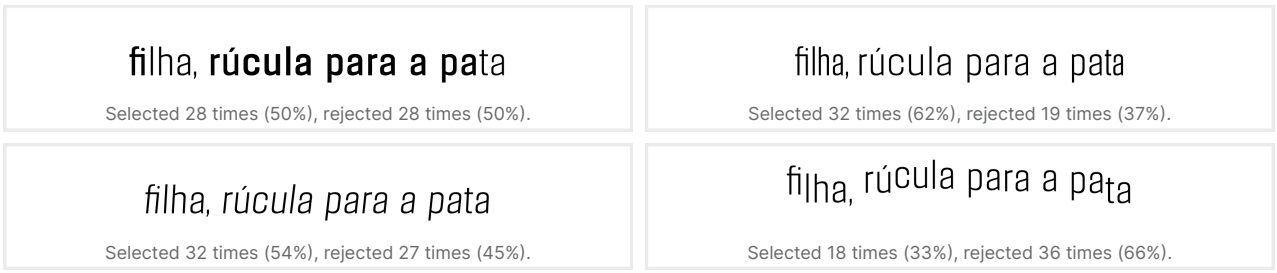


Fig. B.13: Four ways to represent a neutral utterance's pitch.

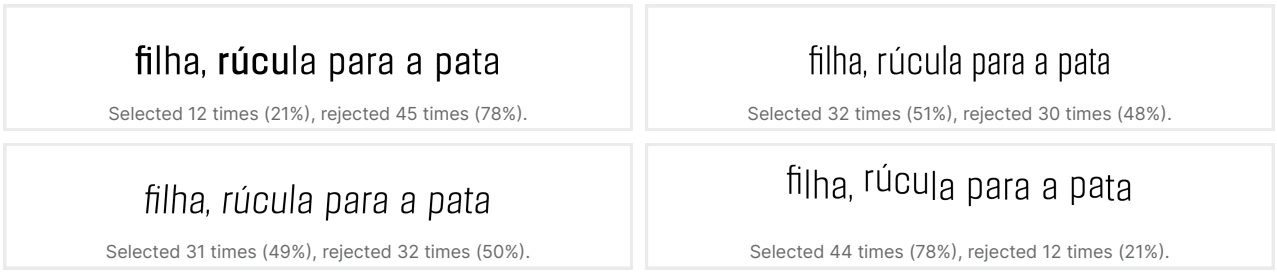


Fig. B.14: Four ways to represent a sad utterance's pitch.

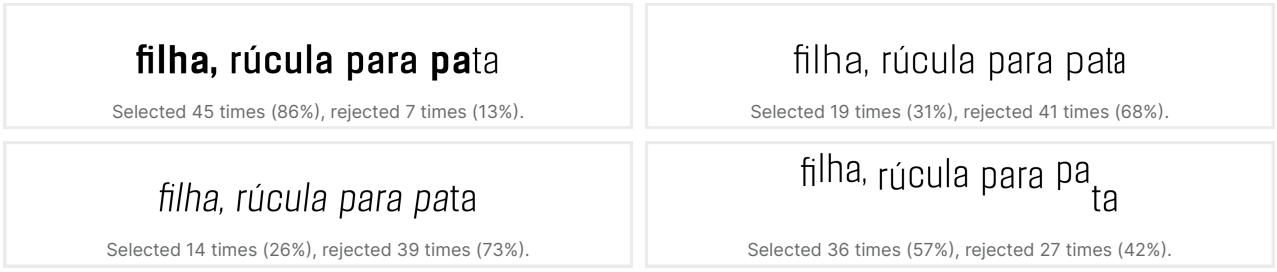


Fig. B.15: Four ways to represent a surprised utterance's pitch.

Fig. B.16: Four ways to represent an angry utterance's pitch.

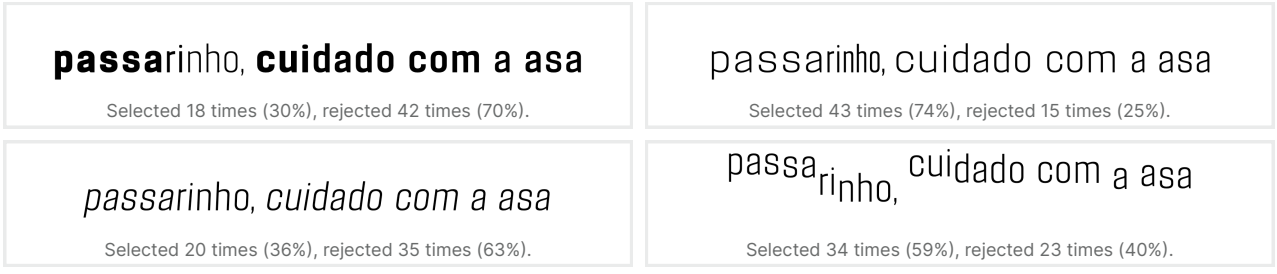


Fig. B.17: Four ways to represent a happy utterance's pitch.

passarinho, cuidado com a asa

Selected 28 times (43%), rejected 36 times (56%).

passarinho, cuidado com a asa

Selected 25 times (49%), rejected 26 times (50%).

passarinho, cuidado com a asa

Selected 43 times (72%), rejected 16 times (27%).

passarinho, cuidado com a asa

Selected 22 times (35%), rejected 40 times (64%).

Fig. B.18: Four ways to represent a neutral utterance's pitch.

passarinho, cuidado com a asa

Selected 11 times (20%), rejected 42 times (79%).

passarinho, cuidado com a asa

Selected 40 times (61%), rejected 25 times (38%).

passarinho, cuidado com a asa

Selected 18 times (31%), rejected 39 times (68%).

passarinho, cuidado com a asa

Selected 51 times (78%), rejected 14 times (21%).

Fig. B.19: Four ways to represent a sad utterance's pitch.

passarinho, cuidado com a asa

Selected 30 times (55%), rejected 24 times (44%).

passarinho, cuidado com a asa

Selected 32 times (56%), rejected 25 times (43%).

passarinho, cuidado com a asa

Selected 33 times (55%), rejected 26 times (44%).

passarinho, cuidado com a asa

Selected 18 times (32%), rejected 38 times (67%).

Fig. B.20: Four ways to represent a surprised utterance's pitch.

Appendix C

Images & videos used in experiment #3

Presented here are the videos and images used in the third evaluation.[1] We prepared two typographic instances from two recorded readings, processing them both as animated closed-captions and static images. We then randomly selected which of the videos/images would be used in the test (which, as presented in section 6.1.2, 78, used only one video/image per round and two sound files).

[1] For a demonstration of the test platform itself, see youtu.be/KYGTtV6RqzU

In Table C.1, we list links for the animated closed-caption versions of each poetic reading. Note that these videos show the whole poems, i.e., they are not divided by their stanzas, as they were in the experiment. Also, they have the audio track, which was also not the case in the experiment, where the audio was available in a separate player.

Poem	**Video A link**	**Video B link**
Três Prenúncios, III	youtu.be/CdgyhU9r54o	youtu.be/b0uffq9pjSo
Súcubo	youtu.be/JOOw5qvXAvc	youtu.be/ywLkw1ox15k
Dez sonetóides mancos, VI	youtu.be/YlZG4mTEHiY	youtu.be/Rcz556nJoAo
Três tercinas, I	youtu.be/BoEg8zjjm0w	youtu.be/TuAqDnryjLM

Table C.1: Links for all speech-modulated closed-captions used in experiment #3.

In Table C.2, we list which version of each video and image was shown to participants. Both audio files were always presented, but only one image/video, which was defined which audio was the *correct* alternative.

Poem	**Stanzas used in the image evaluation**	**Stanzas used in the video evaluation**
Três Prenúncios, III	A, B, B	A, B, A
Súcubo	A, B, A, B	B, A, A, A
Dez sonetóides mancos, VI	B, A, B, B	A, B, A, A
Três tercinas, I	B, A, A, A	A, B, A, B

Table C.2: Which image or video was shown in each round of experiment #3?

Fig. C.1: Two typographic instances for poetic readings of *Três Prenúncios, III*.

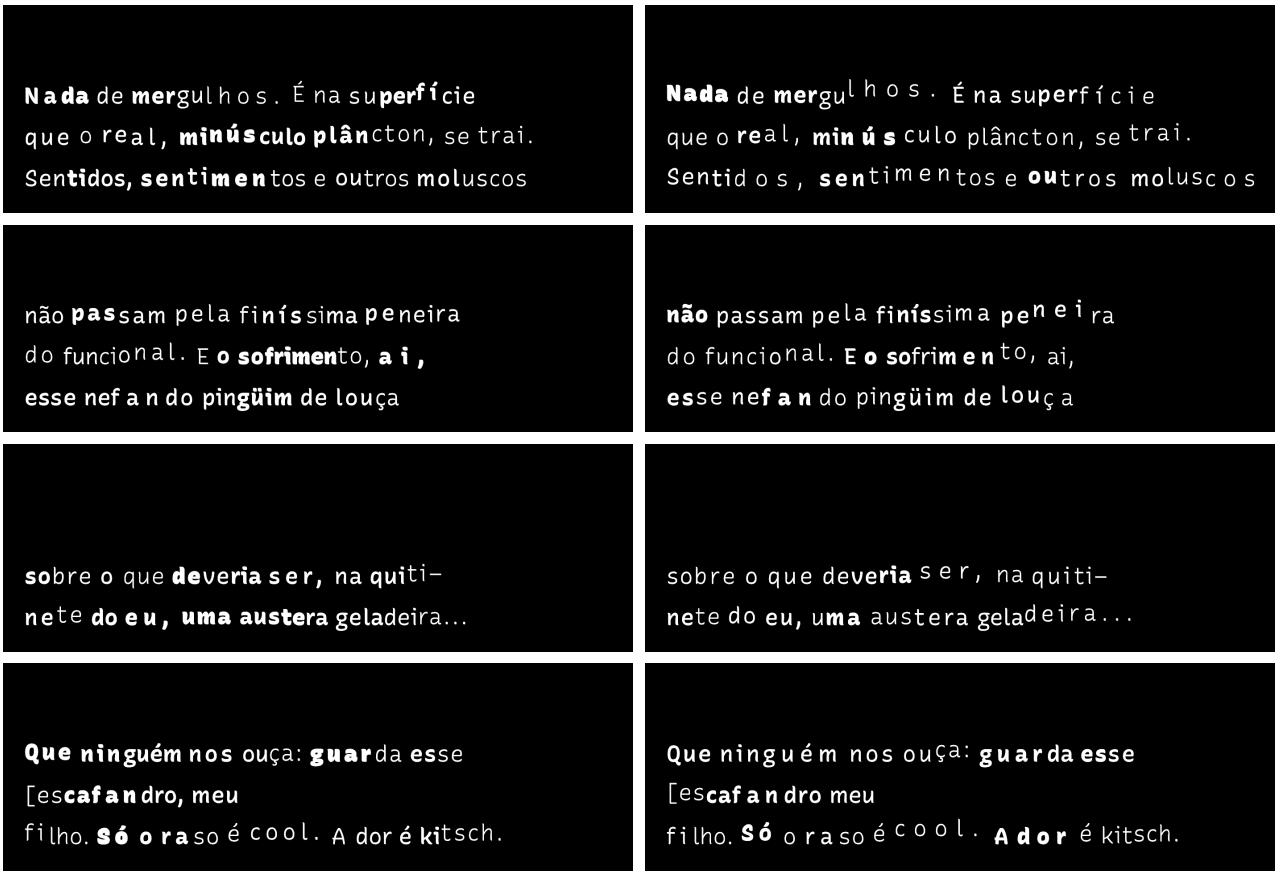


Fig. C.2: Two typographic instances for poetic readings of *Súcubo*.

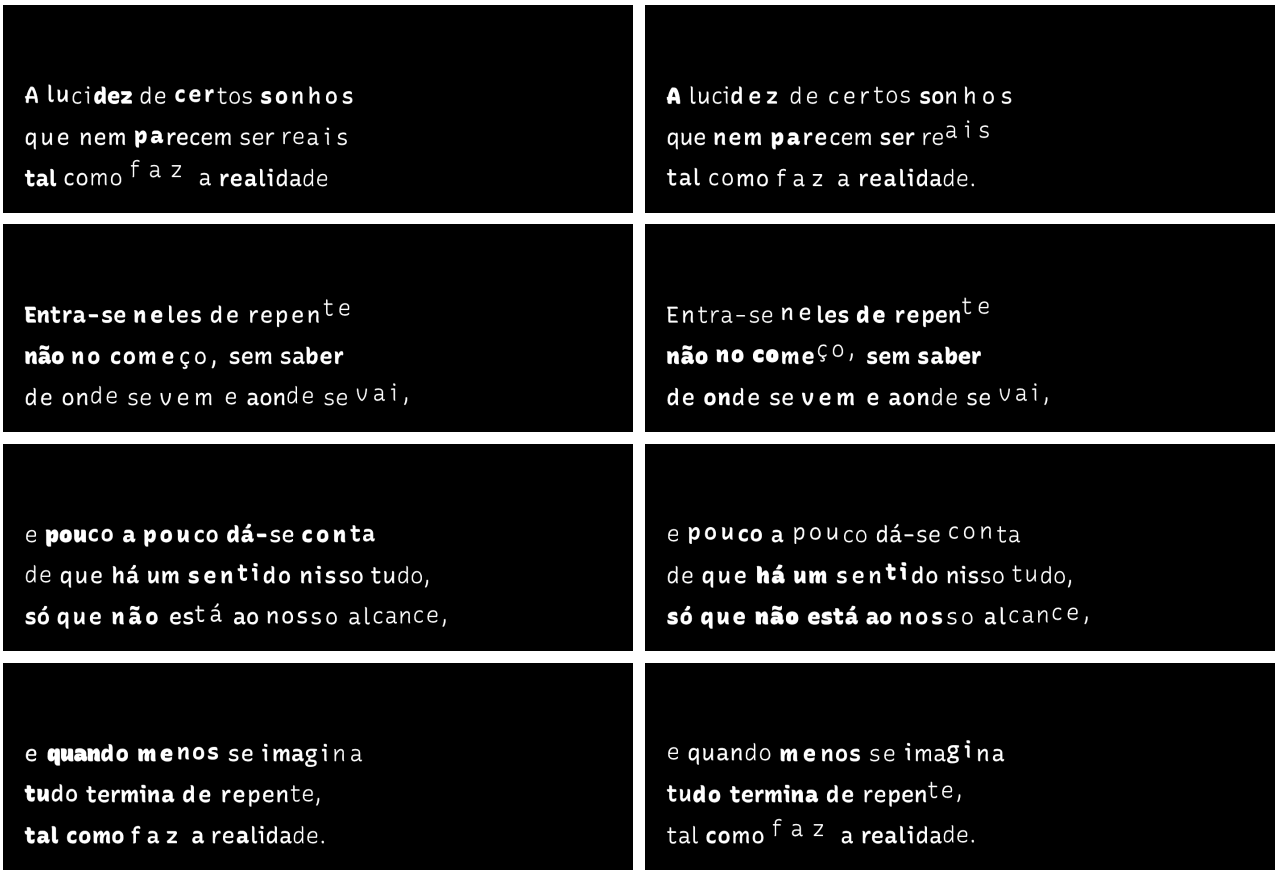


Fig. C.3: Two typographic instances for poetic readings of *Dez sonetóides mancos, VI.*

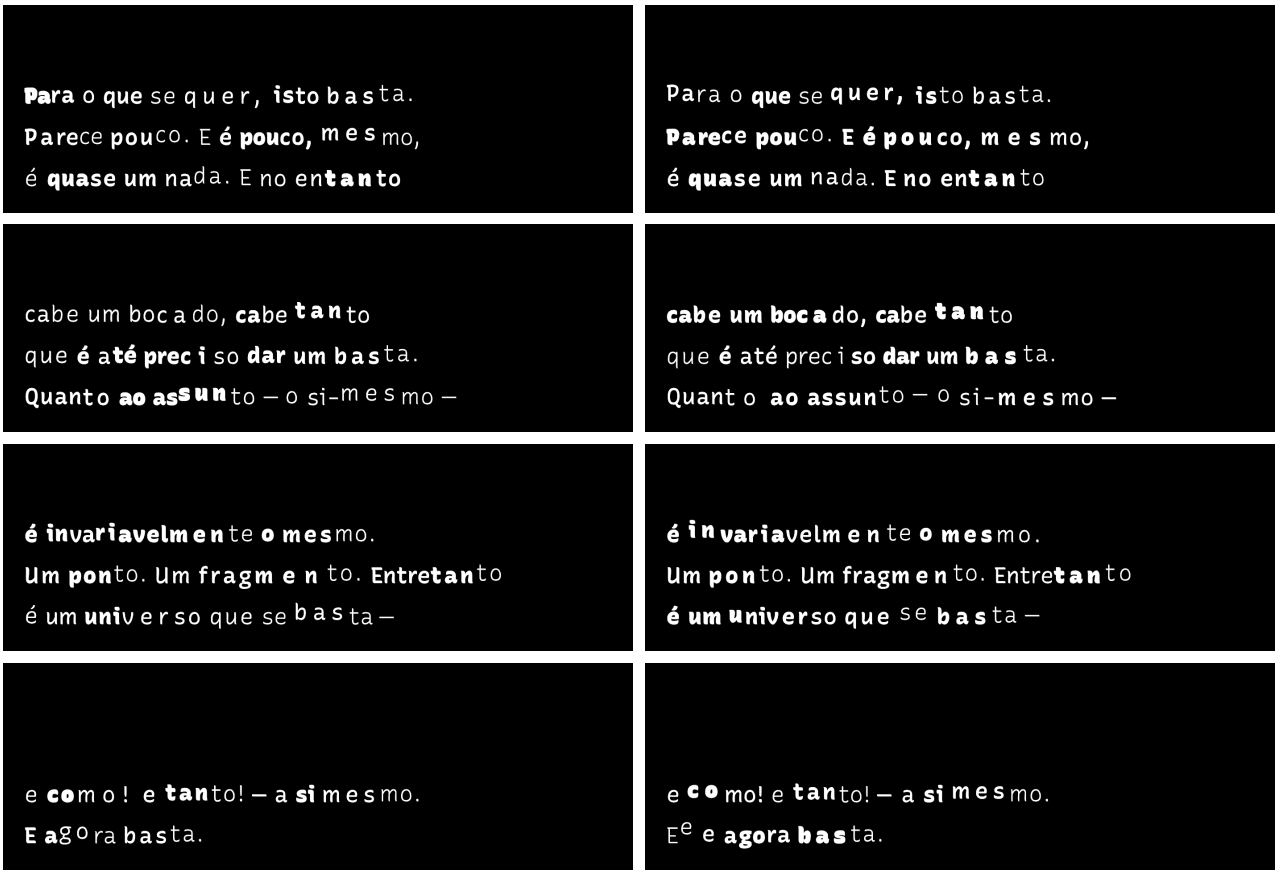


Fig. C.4: Two typographic instances for poetic readings of *Três tercinas, I.*

Appendix D

Poems used in experiment #3

1.

O fim nos acena
com um gesto discreto:
um pouco de pena
e escárnio secreto.

Mas não, ainda é cedo —
dizemos, com um ríctus
de explícito medo
(embora convictos

de que não seria
a nós destinado
bilhete premiado
de tal loteria).

Três Prenúncios, III, Britto (2007, p. 62)

2.

Nada de mergulhos. É na superfície
que o real, minúsculo plâncton, se trai.
Sentidos, sentimentos e outros moluscos

não passam pela finíssima peneira
do funcional. E o sofrimento, ai,
esse nefando pingüim de louça

sobre o que deveria ser, na quiti-
nete do eu, uma austera geladeira...

Que ninguém nos ouça: guarda esse escafandro, meu
filho. Só o raso é cool. A dor é kitsch.

Dez sonetóides mancos, VI, Britto (2003, p. 62)

3·

A lucidez de certos sonhos
que nem parecem ser reais
tal como faz a realidade,

Entra-se neles de repente
não no começo, sem saber
de onde se vem e aonde se vai,

e pouco a pouco dá-se conta
de que há um sentido nisso tudo,
só que não está ao nosso alcance,

e quando menos se imagina
tudo termina de repente,
tal como faz a realidade.

Súcubo, Britto (2003, p. 75)

4.

Para o que se quer, isto basta.
Parece pouco. E é pouco, mesmo,
é quase um nada. E no entanto

cabe um bocado, cabe tanto
que é até preciso dar um basta.
Quanto ao assunto — o si-mesmo —

é invariavelmente o mesmo.
Um ponto. Um fragmento. Entretanto
é um universo que se basta —

e como! e tanto! — a si mesmo.
E agora basta.

Três tercinas, 1, Britto (2003, p. 23)

Appendix E

Chronology of this research

This research project started in brainstorms done in meetings when I was attending Paula's *Affective Computing* classes in the Computer Engineering Graduate School of the University of Campinas. Since then, I have been recording my hours every time I worked on tasks related to the research project. In figures E.1, E.2, E.3, and E.4 we have plotted these hours, roughly divided into meaningful categories. Some important landmarks are jotted down, telling the story of how the project advanced through the years.

Not included here is the time I spent in classes related to the Master's program, which officially started in the second semester of 2018. Also, we arbitrarily defined a cut-off point as of December 2020, even though work did not end on New Year's Eve.

We share this because, although the story it tells is of one and only one student, it can nevertheless give other students a sense of the relative effort that went into each of the different steps we took, which is something we feel could be shared and discussed more often.

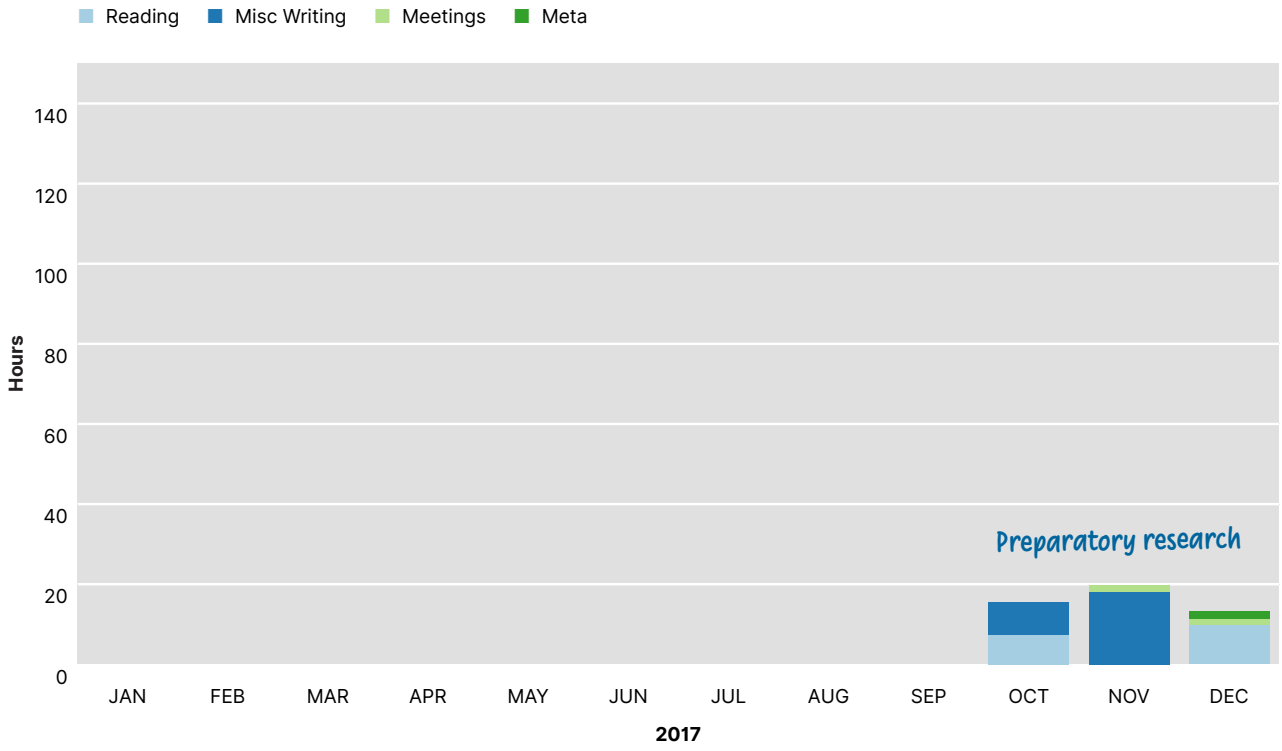


Fig. E.1: Hours spent in 2017.

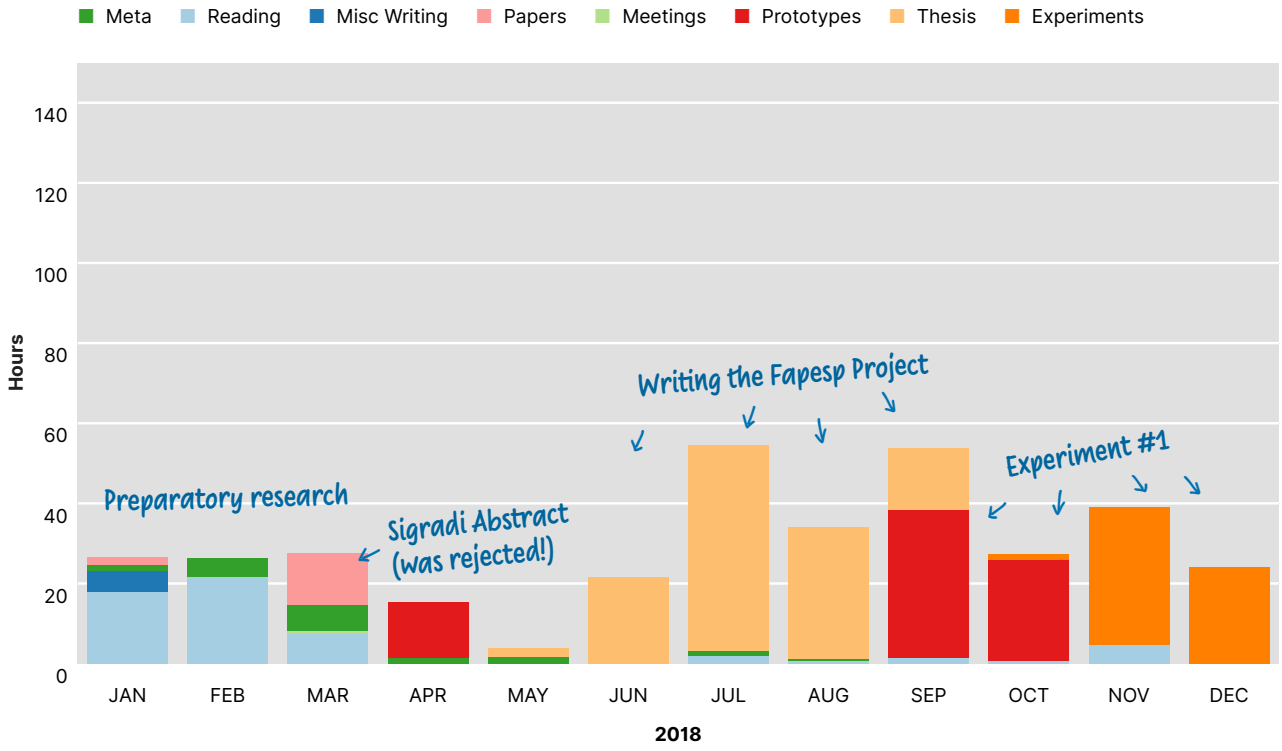


Fig. E.2: Hours spent in 2018.

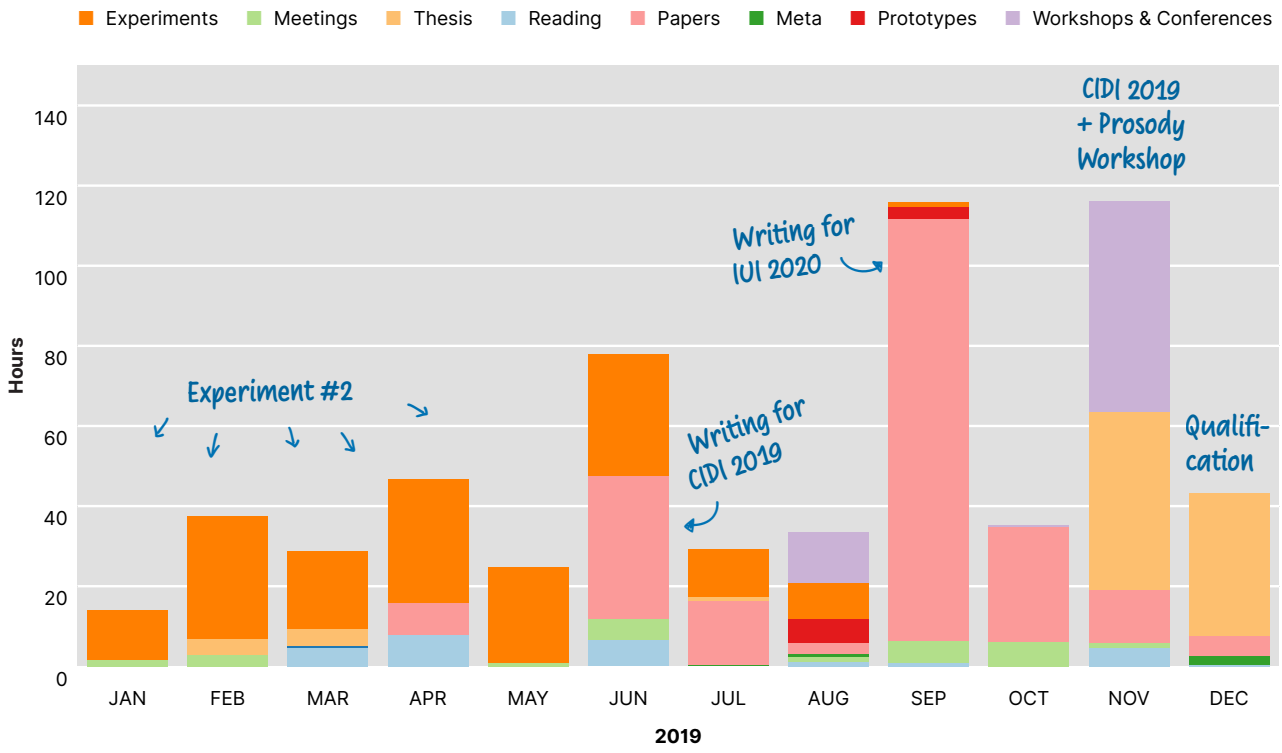


Fig. E.3: Hours spent in 2019.

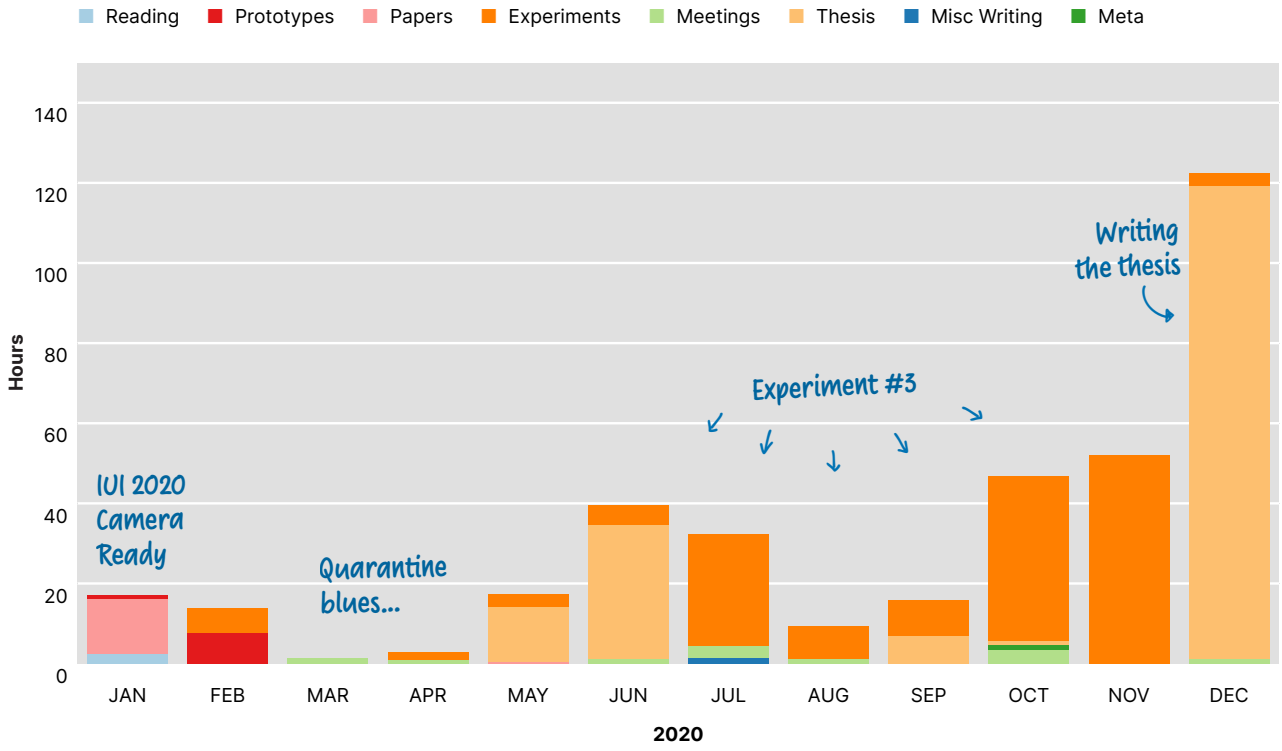


Fig. E.4: Hours spent in 2020.

This thesis' main text was set in John Hudson's Brill typeface,
with Rasmus Andersson's Inter used for notes, tables, page numbers, etc,
and Raph Levien's Inconsolata used for displaying computer code.
The document was composed by Donald Knuth's TEX system,
using the `tufte-book` *document class.*