Universidade de Lisboa

Faculdade de Letras

Perception of Emotions Across Portuguese Varieties

Mestrado em Linguística

Diana Cristina Sobral dos Santos

2025

Dissertação especialmente elaborada para a obtenção do grau de Mestre, orientada por

Professora Doutora Marisa Cruz

Index

Abstract

This study investigates the perception of emotional prosody across two major varieties of Portuguese - Brazilian (BP) and European (EP) - examining how differences in their intonational systems and cultural backgrounds shape emotion recognition. Building on theoretical frameworks of vocal emotion (Darwin, 1872; Scherer, 1986; Ekman, 1992), an experiment was designed using acoustically controlled, acted stimuli expressing neutrality, happiness, sadness, and anger. Native listeners from both varieties completed a perception task measuring identification accuracy and reaction times.

Results revealed a clear emotion hierarchy: neutrality and sadness were recognized most accurately and rapidly, whereas happiness and anger were frequently confused, indicating higher perceptual ambiguity. While both listener groups showed a native-variety advantage, EP participants demonstrated superior overall accuracy, attributed to greater exposure to BP media and more consistent cue reliance. These findings highlight the interplay between universal acoustic cues, variety-specific phonological structuring, and cultural exposure in shaping emotional speech perception.

This research contributes to prosodic typology in Portuguese, cross-variety speech perception, and the integration of linguistic and affective models, with implications for phonological theory and applied technologies.

Keywords: emotional prosody, Brazilian Portuguese, European Portuguese, cross-linguistic perception, cultural background

Resumo

Este estudo investiga a perceção da prosódia emocional em duas variedades do português—o português brasileiro (PB) e o português europeu (PE)—examinando de que modo as diferenças nos seus sistemas entoacionais e contextos culturais moldam o reconhecimento de emoções. Com base em modelos teóricos da emoção vocal (Darwin, 1872; Scherer, 1986; Ekman, 1992), foi concebida uma experiência utilizando estímulos controlados (i.e., produzidos por atores) e acusticamente modificados que expressam neutralidade, alegria, tristeza e raiva. Ouvintes nativos de ambas as variedades realizaram uma tarefa de perceção que media a precisão da identificação (i.e., taxa de acerto) e os tempos de reação.

Os resultados revelaram uma hierarquia emocional clara: a neutralidade e a tristeza foram reconhecidas com maior precisão e rapidez, enquanto a alegria e a raiva foram frequentemente confundidas, indicando uma maior ambiguidade percetiva. Embora ambos os grupos de ouvintes tenham demonstrado uma vantagem da variedade nativa, os participantes do PE apresentaram uma precisão global superior, atribuída a uma maior exposição a midia brasileira e a uma dependência mais consistente de pistas acústicas. Estes resultados evidenciam a interação entre pistas acústicas universais, a estruturação fonológica específica de cada variedade linguística e a exposição cultural na perceção da fala emocional.

Esta investigação contribui para a tipologia prosódica do português, para o estudo da perceção da fala entre variedades e para a integração de modelos linguísticos e afetivos, com implicações para a teoria fonológica e para tecnologias da fala.

Palavras-chave: prosódia emocional, português brasileiro, português europeu, percepção linguística transversal, contexto cultural

List of Figures

List of Tables

List of Abbreviations

APA: American Psychological Association

BP: Brazilian Portuguese

CPM: Component Process Model

CVF_0: Coefficient of Variation of Fundamental Frequency

EP: European Portuguese

F_0: Fundamental Frequency

fMRI: Functional Magnetic Resonance Imaging

GLMM: Generalized Linear Mixed Model

HNR: Harmonic-to-Noise Ratio

PCA: Principal Component Analysis

P-ToBI: Portuguese Tones and Break Indices (prosodic transcription system)

RT: Reaction Time

Acknowledgements

First and foremost, I would like to express my profound gratitude to my Orixás and my spiritual guides, whose strength, wisdom, and protection have guided me at every step of this academic and personal journey. Their presence has been a source of resilience, grounding, and inspiration, reminding me of the importance of faith and balance throughout the challenges of this dissertation.

I owe a special debt of gratitude to my parents and siblings, whose constant support, encouragement, and unwavering belief in my potential made this achievement possible. Their faith in me sustained my determination during the most demanding stages of this process.

I also extend my sincere thanks to Jonathan, who, with generosity and patience, was always ready to stand by me whenever I needed him. His friendship and solidarity provided me with reassurance and stability.

My heartfelt appreciation goes as well to the family that Portugal gave me—Bora Bora—and to my dear friends Ana, Sara and Mariana. Their kindness, care, and understanding created a sense of home away from home. They nurtured me with genuine love and compassion, offering not only support but also joy and comfort throughout this demanding period.

I am profoundly grateful to my mentor, Professor Marisa, for her guidance, patience, and insightful orientation. Her academic expertise and thoughtful encouragement have been essential in shaping this work and in helping me to grow as a researcher.

I would also like to acknowledge my therapist, Pérola, whose support extended beyond words. With patience and empathy, she held my hand during difficult times and encouraged me to rise each time I felt unable to continue. Her role in sustaining my mental and emotional well-being was indispensable to the completion of this dissertation.

Finally, I wish to thank all my friends and extended family who have always believed in me and never allowed me to give up. Their encouragement and confidence in my path gave me the strength to persist and the courage to bring this work to its conclusion.

To all of you, I extend my deepest and most sincere gratitude.

Chapter 1: Introduction

Emotions were defined by the American Psychological Association (APA, 2020) as complex psychological states that involve a combination of physiological arousal, behavioral responses, and cognitive appraisal. This definition can be expanded by exploring other perspectives.

This chapter will delve into psychoevolutionary theories of emotion, specifically exploring the perspectives presented by Darwin (1872), Plutchik (1982), and Ekman (1992). While these theorists primarily focused on facial and behavioral expressions, their frameworks also underpin modern research on vocal emotion, particularly in cross-linguistic contexts (e.g., Pell & Skorup, 2008).

Understanding the complexities of human emotions has long been a central focus of psychological inquiry. Across centuries of intellectual discourse, scholars and thinkers have proposed diverse theories to explain the nature, origins, and functions of emotions. These theories offer unique perspectives on the underlying mechanisms driving emotional experiences, ranging from biological and evolutionary frameworks to socio-cultural and cognitive models.

The evolutionary perspective on emotions explores the historical background in which these intricate psychological processes have developed. It suggests that emotions did not arise randomly but rather evolved over time, both in humans and non-human animals, through the mechanism of natural selection (Darwin, 1897). These emotions endure in contemporary humans because they have provided adaptive benefits in coping with the demands of survival and reproduction across our ancestral history (Johnson, 2009). These adaptations include not only facial expressions but also vocalizations, which serve as rapid, universal signals of intent (e.g., anger's high pitch to signal threat; Banse & Scherer, 1996).

Darwin's perspective presented in *The Expression of the Emotions in Man and Animals* (1897 emphasized the communicative role of emotional expressions. His theory rests on three core principles: the Principle of Serviceable Habits, the Principle of Antithesis, and the Principle of Actions Due to the Constitution of the Nervous System. The first principle suggests that certain emotional expressions, which once served an adaptive purpose, may persist in contemporary humans despite no longer being functionally relevant. The second principle posits that opposing emotional states are expressed through contrasting behaviors, such as the differences in body language and facial expressions between fear and confidence. Finally, the third principle asserts that certain emotional expressions are reflexive, driven by the autonomic nervous system rather than conscious intention, reflecting inherent neural structures (Darwin, 1897).

Though his focus was largely visual, modern research extends Darwin's principles to vocal prosody (Banse & Scherer, 1996; Owren & Bachorowski, 2001; Sauter et al., 2010):

(a) Serviceable Habits: Vocal pitch elevation in anger may derive from ancestral threat displays (Ohala, 1984; Morton, 1977). This aligns with cross-species studies showing high-frequency vocalizations in aggressive contexts (Fitch, 2000);

(b) Antithesis: Contrasting prosodic patterns (e.g., low pitch in sadness vs. high pitch in happiness) align with opposing emotional states (Juslin & Laukka, 2003; Pell et al., 2009);

(c) Nervous System: Autonomic arousal modulates vocal parameters like intensity and speech rate (Scherer, 1986; Johnstone & Scherer, 2000). For example, sympathetic activation in fear produces faster speech (Williams & Stevens, 1972, *apud* Pell et al., 2009).

Those principles, rooted in observations of animal and human behavior, continue to influence contemporary theories of emotion. By highlighting the adaptive significance of emotional expressions and emphasizing the role of innate neurological mechanisms, Darwin's work laid the groundwork for subsequent theorists, such as Plutchik, to further develop our understanding of the complex interplay between biology, behavior, and emotion.

Robert Plutchik was a prominent figure in psychology, who made significant contributions to our understanding of emotions through his psychoevolutionary theory. Building upon Darwin's insights, he further developed the evolutionary perspective on emotions and suggested that emotions have deep evolutionary roots and serve as adaptive responses to environmental challenges (Plutchik, 1965).

Central to his work is the concept of primary emotions, universal across species, which he proposed being the foundational building blocks of human emotional experience, as they have evolved to serve as adaptive responses to environmental challenges or opportunities. Emotions such as fear, anger, happiness and sadness are considered prototypical due to their representative aspect of foundational emotional states from which more complex emotions and behaviors are derived (Plutchik, 1982). Fear, for instance, is linked to behaviors aimed at self-preservation, such as fleeing from danger. Anger, on the other hand, may trigger aggression to defend against threats or assert dominance. Similarly, happiness is associated with behaviors that foster social bonding and reproduction, and sadness often facilitates withdrawal or a reevaluation of strategies following a loss. These prototypical behaviors are not only biologically rooted but also share cross-species similarities, highlighting their evolutionary significance.

By framing emotions as adaptive responses, Plutchik’s theory underscores the evolutionary continuity between human and non-human emotional systems, suggesting that prototypical emotions and behaviors are integral to navigating both social and environmental

challenges. This perspective supports scholars like Paul Ekman to explore the intricate interplay between biological, psychological, and sociocultural factors shaping human emotional experiences.

Paul Ekman is a renowned psychologist whose pioneering work on emotions and facial expressions has significantly influenced the field of psychology. His extensive research has provided deep insights into the universality of emotions and the ways in which they are expressed across different cultures.

Ekman's most significant contribution is his theory of the universality of facial expressions. In the 1960s, Ekman and his colleague Wallace V. Friesen conducted extensive cross-cultural research to test whether people from different cultures recognize the same facial expressions as representing the same emotions. They discovered that six basic emotions—happiness, sadness, fear, anger, surprise, and disgust—were universally recognized across a variety of cultures, including those with minimal exposure to Western influences (Ekman & Friesen, 1971).

The identification of the six basic emotions aligns with Johnson-Laird and Oatley's (2009) analysis of emotion words, which posited these emotions as fundamental. Their research demonstrated that these emotions are universally recognized, forming distinct categories that are consistent across different languages and cultures. Additionally, Ekman's (2016) survey of 248 scientists revealed a high level of agreement on this set of basic emotions, providing expert consensus for their universality and significance. This consensus solidifies the status of happiness, sadness, anger, fear, disgust, and surprise as fundamental, cross-culturally recognized emotions.

This framework was later extended from facial expressions to vocal prosody, a key expansion pioneered by Scherer (1986). Subsequent research, such as that by Pell and colleagues, provided robust evidence for the universality of vocal emotional cues. The main findings from this body of work include: (i) Distinctive Signals: anger’s high F0 and happiness’s wide pitch range are recognizable across cultures (Pell & Skorup, 2008); and (ii) Automatic Appraisal: vocal emotion recognition is faster for 'basic' emotions like anger (Pell et al., 2009), a finding that supports Ekman’s hardwiring hypothesis.

Chronologically, the progression of emotions research reflects a deepening understanding of the interplay between biology, behavior, and evolution. Darwin's initial exploration provided the evolutionary context, Plutchik's theory elaborated on the functional and adaptive aspects, and Ekman's empirical studies offered robust evidence and refined our knowledge of the universal characteristics of emotions. However, while these theories emphasize emotional expression through facial and behavioral cues, the role of prosodic signals - variations

in pitch, rhythm, and intonation - remains an essential yet less explored dimension of emotional communication.

Prosody, as an integral component of spoken language, carries substantial emotional weight, often revealing a speaker's affective state beyond lexical content (Nygaard & Queen, 2008). One can ask how effectively do prosodic cues convey specific emotions such as happiness, sadness, and anger? Are these cues universally perceived, or do linguistic and cultural differences influence their perception? Addressing these questions is central to understanding the intersection of emotion and speech, which this study aims to explore, particularly in relation to the perception and categorization of emotions across two language varieties through suprasegmental features.

The primary objective of this study is to investigate and compare the emotion perception abilities of Brazilian (BP) and European Portuguese (EP) native participants. Specifically, this study aims to provide answers to the following questions:

(1) Do BP and EP native participants perceive emotions differently?

(2) Which acoustic features are most salient in cross-variety emotion recognition?;

(3) How do BP and EP's intonational differences (Frota et al., 2015) shape perception of basic emotions?;

(4) Do cultural dimensions (Hofstede, 2001) mediate prosodic emotion perception?

This chapter has outlined how evolutionary theories of emotion—from Darwin's adaptive principles to Ekman's universals—provide a foundation for studying vocal prosody as a critical channel of emotional communication. Building on this framework, and given the research questions listed above, the subsequent chapters will investigate the acoustic and cross-linguistic dimensions of emotional perception. Chapter 2 examines the acoustic characteristics of emotions, detailing how parameters like pitch, intensity, and speech rate encode affective states. Chapter 3 compares the intonational patterns of Brazilian and European Portuguese, hypothesizing how their prosodic differences may shape emotion perception. Chapter 4 presents the methodology, including stimulus selection, their acoustic analysis, and perceptual experiments. The results are described in chapter 5, and chapter 6 presents the discussion of the empirical findings, addressing how linguistic and cultural background influences the decoding of emotional prosody. The conclusion is brought in chapter 7. Together, these chapters bridge evolutionary theory with empirical analysis to advance our understanding of emotion perception in diverse linguistic systems.

Chapter 2 - Acoustic Characteristics of Emotions

2.1. Introduction

An important dimension of emotional expression lies in the acoustic characteristics of emotions. This area of study delves into how emotions are conveyed through vocal parameters like pitch, intensity, and rhythm. Researchers such as Scherer and Banse (1996) have made substantial contributions to this field, examining how vocal cues serve as robust indicators of emotional states. Therefore, by investigating the acoustic properties of emotions, it is possible to further elucidate the complex interplay between physiological responses and emotional communication, thus enriching the comprehension of how emotions are experienced and expressed across different modalities.

Contrary to research on facial expressions, in which static visual stimuli such as photographs or flashcards can represent emotions with minimal ambiguity, research on vocal emotion presents inherent challenges. Vocal attributes of emotion cannot be captured in a single "snapshot" because emotions conveyed through voice are dynamic and time-dependent, requiring careful analysis of acoustic features like pitch, intensity, duration, and rhythm (Pell et all, 2009). This temporal complexity demands that researchers work with continuous auditory stimuli, which are far more intricate to isolate and standardize.

Furthermore, the vocal expression of emotions is not independent from language-specific prosodic systems. Emotional meaning in speech is shaped by linguistic features such as intonation, stress patterns, rhythm, and speech rate—components that are embedded within the phonological and pragmatic structure of each language (Ladd, 2008; Van Bezooijen, 1992). Unlike facial expressions, which are often considered biologically universal and innate (Ekman, 1993), vocal emotional cues exhibit considerable variability influenced by phonological conventions and discourse norms (Juslin & Laukka, 2003). These cues reflect how emotional prosody is processed and interpreted within specific linguistic contexts, rather than being governed purely by sociological “display rules” (Pell & Kotz, 2011). Thus, vocal emotional expression is not just sociologically modulated but is fundamentally shaped by the linguistic system itself. This supports the view that the communication of emotion is a core, integrated linguistic function, not merely a paralinguistic add-on (Scherer, 1986).

Given these complexities, one may not be surprised that research on vocal emotions has not achieved the same breadth or depth as studies focusing on facial expressions. While Ekman's foundational work using static facial expression flashcards laid the groundwork for universal emotion theories, vocal research must contend with additional layers of variability, requiring

interdisciplinary approaches that blend acoustic analysis, cross-cultural psychology, and linguistic theory to advance the field (Pell et all, 2009).

Klaus Scherer, with his theoretical framework, explains how physiological changes induced by emotions influence vocal expression. One of his first foundational frameworks, the covariance model, describes that speech patterns covary with emotionally induced physiological changes in respiration, phonation and articulation (Scherer, 1984). The vocalization is impacted by these variations at three levels: suprasegmental, segmental and intrasegmental (Pell et all, 2009).

Building on Scherer's theoretical insights, Banse and Scherer (1996) conducted one of the most comprehensive empirical studies on the acoustic profiles of emotional expression through voice. Their findings provided robust evidence that specific emotions are associated with distinct patterns across multiple acoustic dimensions, particularly at the suprasegmental level, such as pitch (F_0), intensity, speech rate, and voice quality. Their findings are detailed in sections 2.2. and 2.3, where the mechanisms underlying the production and perception of emotional vocalizations are summarized.

2.2. Production

Scherer (1986, 2003) developed the Component Process Model (CPM) which has been instrumental in shaping current understanding of vocal emotion production as a biologically grounded, dynamic process. At its core, the CPM posits that emotions emerge through continuous, recursive appraisal checks across several dimensions—including novelty, goal relevance, coping potential, and normative significance. Each of these dimensions triggers physiological and behavioral responses across subsystems (e.g., autonomic, motor, respiratory), which in turn shape the acoustic profile of vocal expressions (Scherer, 1986).

Scherer (2003) expanded the CPM by incorporating neurophysiological, motor-expressive, and perceptual-decoding components to offer a more comprehensive account of emotional communication. According to this framework, vocal emotion expression is not merely a culturally acquired behavior, but a biologically grounded process shaped by appraisal-driven physiological responses. The model outlines a tripartite sequence: (i) encoding, where emotional appraisals trigger physiological changes—such as variations in respiration, muscle tension, and vocal fold activity—that modulate acoustic features like pitch (F_0), intensity, tempo, and voice quality; (ii) transmission, in which these modulated acoustic cues are embedded in the speech signal; and (iii) decoding, whereby listeners interpret these cues to infer the speaker’s emotional state. Importantly, Scherer (2003) emphasized that accurate emotional inference can occur even in the

absence of semantic content, as demonstrated in studies using nonsense syllables or pseudosentences (Banse & Scherer, 1996; Scherer et al., 2001).

Prior to that study, Banse and Scherer(1996) emphasized that systematic knowledge about the specific acoustic patterns characterizing human vocal expression of emotions is limited. Nevertheless, they affirm that certain acoustic variables play a significant role in vocal emotion signaling. These variables include: (i) fundamental frequency (F_0) and its level, range and contour; (ii) vocal energy or amplitude; (iii) the energy distribution in the frequency spectrum, which affects the perception of voice quality and timbre; (iv) formants – F*n*; and (v) temporal phenomena such as tempo and pausing. These authors proposed a multidimensional acoustic framework for emotional vocal production – which is used as a reference in this research - demonstrating that core prosodic parameters systematically differentiate emotions in speech. They designed their study using semantically neutral stimuli in order to isolate the acoustic characteristics of emotional vocalizations from semantic and syntactic influences. Rather than analyzing natural language, they constructed seven-syllable nonsense sentences based on syllables drawn from six European languages—German, English, French, Italian, Spanish, and Danish. These syllables were randomly arranged to mimic the prosodic flow of real speech while remaining devoid of lexical meaning. The authors observed that joy and anger share acoustic markers like high pitch, high intensity, and fast speech rate, yet diverge in vocal texture - joy is expressed with a bright, resonant timbre and smooth rhythm, whereas anger features a harsh, tense vocal quality with clipped syllables and minimal pausing. Sadness, by contrast, is characterized by low pitch, soft intensity, and slow speech rate, accompanied by lengthened syllables, frequent pauses, and a dull, breathy voice. Fear tends to present high and unstable pitch, medium-to-high intensity, and an uneven tempo, with irregular timing and abrupt transitions reflecting heightened arousal. Disgust is conveyed with a low-to-medium pitch, moderate intensity, and a slow-to-medium tempo, featuring a muffled, nasal-like voice and constricted rhythmic flow. Together, these findings, summarized in Table 1, illustrate how temporal and spectral properties coalesce to form distinct acoustic profiles for each emotion, enabling rich emotional signaling even in semantically neutral or ambiguous contexts.

Table *1* - Acoustic Characteristics of Joy, Sadness, Fear, Disgust and Anger. *Banse & Scherer (1996, pp. 624-628):*

Emotion	Pitch (F_0)	Intensity/ Loudness	Speech Rate/ Tempo	Temporal Patterns/ Rhythm	Voice Quality	Additional Notes
Joy	High	High	Fast	Smooth rhythm; short syllables; few pauses	Bright, resonant	Wide pitch range; energetic articulation
Sadness	Low	Low	Slow	Lengthened syllables; frequent pauses; slow rhythm	Dull, soft	Narrow pitch range; reduced vocal energy
Fear	High	Medium-high	Fast, uneven	Irregular timing; variable pause lengths; abrupt transitions	Tense, breathy	Irregular rhythm; pitch instability
Disgust	Low-medium	Medium	Slow-medium	Constricted flow; slower rhythm; muffled articulation	Harsh, nasal-like	Constricted articulation; low resonance
Anger	High	High	Fast	Clipped syllables; sharp rhythm; minimal pausing	Harsh, tense	Sharp rhythm; elevated subglottal pressure

A recent meta-analytic review by Larrouy-Maestri et al. (2024) offers a comprehensive synthesis of nearly three decades of research on emotional prosody. This large-scale analysis critically examined the reliability and consistency of acoustic-emotion associations across both

speech prosody and nonverbal vocalizations. The authors reviewed studies utilizing both acted and naturalistic speech samples, highlighting significant methodological variability—such as differences in speaker demographics, emotional elicitation techniques, and corpus design. The dataset predominantly included Indo-European languages such as English, German, Dutch, and French. Their findings underscore a lack of consistent mappings between specific acoustic cues and emotional categories. While some emotions, like sadness and anger, displayed relatively robust acoustic signatures—such as sadness being associated with low pitch and slow tempo, and anger with high intensity and sharp pitch contours—others like fear and surprise showed high variability and overlapping acoustic profiles, reducing interpretive reliability. This meta-analysis advocates for more rigorous, cross-linguistic methodologies and the development of mechanistic models to account for the variability in how emotions are vocally expressed (and perceived) across different communicative settings.

From cross-linguistic to language-specific studies on emotional speech production, Colamarco and Moraes (2008) provide empirical insights into the vocal production of emotions in BP, particularly their expression through prosodic features. By eliciting emotional speech from trained speakers using controlled sentences in both declarative and interrogative forms, the authors analyzed how emotions such as anger, joy, sadness, and fear influence intonational contours and temporal characteristics. The acoustic analysis revealed that anger and joy are marked by higher fundamental frequency (F_0), broader pitch range, and faster speech rate, reflecting a higher level of physiological arousal. In contrast, sadness was characterized by a lower mean F_0, reduced pitch variability, and slower articulation, suggesting a vocal pattern aligned with diminished energy or emotional withdrawal. These results are aligned with Banse & Scherer's (1996) findings for other languages.

These prosodic adjustments demonstrate how BP speakers modulate their vocal parameters to convey emotional states, with significant variation depending on the type of speech act. Notably, the findings suggest that intonation may interact not only with affective content but also with pragmatic intent, as some emotional cues varied between questions and statements.

Peres (2014) also investigated how different acoustic parameters contribute to the expression (and recognition) of emotions in BP. The author examined excerpts of spontaneous, yet acted, emotional speech extracted from YouTube, each categorized as expressing one of four basic emotions: anger, fear, joy, and sadness. The acoustic analysis focused on eight parameters, encompassing both intonational features and indicators of voice quality. Among the intonational parameters, mean fundamental frequency (F_0) and its coefficient of variation (CVF_0) were found to significantly differ across emotional categories. Specifically, F_0 distinguished joy from fear,

sadness, and anger, while CVF_0 was effective in differentiating sadness and anger. Regarding voice quality, spectral emphasis and spectral slope successfully distinguished sadness from anger, indicating differences in vocal effort.

A recent study by Moraes and Rilliard (2016) focusing on the vocal production of emotions in BP showed how prosodic parameters are shaped not only by affective intent but also by the grammatical structure of utterances. Their experimental design involved native BP speakers producing scripted sentences under four emotional conditions - anger, joy, sadness, and fear -, embedded in three different sentence modes - declarative, interrogative, and imperative. The acoustic analysis concentrated primarily on two dimensions: fundamental frequency (F_0) and duration. F_0 measures included both the mean pitch (register) and the shape of the pitch contour (intonational movement), while duration referred to the temporal length of segments and overall utterance, reflecting speech tempo. The results revealed that emotional states significantly influenced the mean F_0 register. Anger and joy typically resulted in elevated pitch levels, consistent with increased physiological arousal, whereas sadness and fear were associated with lower F_0 values. Conversely, the shape of the intonational contour appeared to be driven more by sentence mode than by emotional content.

Emotional states in EP have also been a topic of investigation. Nunes et al. (2010), for instance, explored how vocal quality varies in the production of emotions in EP by analyzing speech produced by a professional male actor. Two semantically neutral sentences—one simple and one complex—were recorded in six emotional conditions: joy, sadness, despair, fear, anger, and neutrality. Acoustic parameters were focused on fundamental frequency (F_0), jitter, shimmer, harmonic-to-noise ratio (HNR), and autocorrelation (ACR). Statistical analysis revealed significant variation across emotions, with anger showing the highest mean F_0 (~300 Hz) and greatest pitch variability, while sadness and neutrality exhibited the lowest F_0 (~100 Hz). Joy and fear occupied an intermediate range, with joy presenting a notably more dynamic pitch profile. Voice perturbation measures such as jitter and shimmer were elevated in despair, fear, anger, and sadness, indicating increased vocal instability, whereas joy and neutrality were associated with smoother phonation. Additionally, anger and despair showed the lowest HNR and ACR values, reflecting noisier and less periodic vocal signals, while joy and neutral speech exhibited more harmonic and stable vocal qualities. These findings demonstrate that emotional states significantly modulate voice quality in EP. It is noteworthy that joy and neutral speech showed acoustic similarities in some measures. This observed similarity could point to language-specific expressive norms or, alternatively, be influenced by performance-related factors in the study's acting-based methodology, suggesting an avenue for further research.

A more detailed investigation into the vocal production of emotions in EP - which is used as a reference in this dissertation - is the research by Castro and Lima (2010), who developed and validated a set of semantically neutral sentences and pseudosentences in EP, designed to represent seven emotional categories: anger, fear, sadness, happiness, surprise, disgust, and neutrality. Although the study's primary goal was to establish a perceptual validation set, the vocal production phase was meticulously controlled. Native speakers were instructed to express each target emotion with clarity and naturalness, and recordings were evaluated for quality, consistency, and acoustic integrity prior to inclusion. The authors reported consistent acoustic markers across emotional states: anger and happiness showed higher pitch and intensity, while sadness exhibited lower pitch, slower tempo, and reduced vocal energy, thus matching the acoustic properties found by Banse & Scherer (1996) for other languages and by Colamarco & Moraes (2008) for BP. The use of pseudosentences—syllabic constructions devoid of lexical meaning—allowed the researchers to isolate prosodic elements (intonation, rhythm, voice quality) from semantic content, making the dataset particularly suitable for exploring how language-specific prosodic features encode emotional states in EP. The acoustic analysis revealed clear and cross-linguistically consistent emotion-specific profiles. Anger and happiness were produced with high pitch, loud intensity, and a fast tempo, yet differing in timbre - anger being a tense, harsh, and characterized by an energetic tone; happiness being bright, resonant, and expressive. In contrast, sadness was vocalized with low pitch, soft intensity, and slower speech, accompanied by dull, breathy voice quality and low vocal energy. Fear combined high pitch and medium-to-high intensity with a fast, uneven tempo, and showed irregular resonance, reflecting physiological arousal. More complex emotional states like surprise exhibited variable pitch and tempo, with sudden tonal shifts, while disgust was expressed through low-to-medium pitch, moderate intensity, and muffled, nasal-like voice quality, creating a constricted prosodic flow. Neutrality, used as a reference baseline, displayed mid-level values across all parameters and a flat prosodic contour, serving as a control for emotional deviation. These acoustic features used during production of emotions (Castro & Lima, 2010), summarized in Table 2, reinforce the concept that acoustic variation can signal emotional states, making this corpus a valuable resource for studying emotional prosody in Romance languages.

Table 2 - Acoustic Characteristics of Anger, Happiness, Sadness, Fear, Surprise, Disgust and Neutrality in EP. Table drawn to summarize Castro & Lima's (2010) results (pp. 76-78).

Emotion	**P*itch (F0)***	**Intensity / Loudness**	**Speech Rate/ Tempo**	**Voice Quality**
Anger	High	Loud	Fast	Tense; harsh; energetic
Happiness	High	Loud	Fast	Bright; resonant; expressive
Sadness	Low	Soft	Low	Dull; breathy; low vocal energy
Fear	High	Medium-high	Fast; uneven	Tesne; irregular resonance
Surprise	High; variable	Medium	Variable	Sudden shifts in tone and pitch
Disgust	Low-medium	Medium	Slow-medium	Muffled; less resonant; nasal like
Neutrality	Mid-level	Medium	Medium	Balanced, with flat prosodic contour.

However, although these studies show evidence for the existence of universal vocal properties to specific emotions across languages and cultures, it is relevant to highlight that there are also some language-specific features associated to vocal emotion production. Wang et al. (2018), for instance, compared emotional prosody production between Mandarin (a tonal language) and English (a non-tonal language), showing that Mandarin speakers use more heavily voice quality and temporal cues to convey emotions because pitch is already used lexically to distinguish word meaning. This lexical function of pitch limits its flexibility for emotional expression, prompting speakers to modulate other acoustic dimensions like tempo, intensity, and timbre (Wang et al., 2018). This points to the existence of what Laukka and Elfenbein (2021) call "emotional dialects"—language-specific vocal patterns for expressing emotions shaped by both phonological systems and cultural norms.

To sum up, across the diverse body of research on vocal emotion production, consistent acoustic patterns emerge for core emotions such as happiness, sadness, anger, and neutrality. Anger and happiness are frequently characterized by elevated fundamental frequency (F_0), increased intensity, and faster tempo, reflecting higher levels of physiological arousal (Banse & Scherer, 1996; Colamarco & Moraes, 2008; Moraes & Rilliard, 2016; Castro & Lima, 2010). However, they diverge in vocal quality: while anger is often harsh, tense, and rhythmically clipped, happiness tends to present a smoother, more resonant, and expressive timbre (Banse & Scherer, 1996; Castro & Lima, 2010). In contrast, sadness consistently exhibits lower F_0, reduced intensity, and slower articulation, often accompanied by a breathy, subdued voice quality and lengthened syllables (Peres, 2014; Castro & Lima, 2010). Neutral speech generally occupies a mid-range across acoustic dimensions, serving as a baseline from which emotional deviations can be reliably measured (Nunes et al., 2010; Castro & Lima, 2010). These trends are observed across multiple languages (including English or German) and language varieties (BP and EP) and are supported by both acted and semispontaneous speech data (Banse & Scherer, 1996; Larrouy-Maestri et al., 2024; Peres, 2014).

Despite this convergence, notable gaps remain in the current literature. Methodological heterogeneity—including differences in elicitation techniques, corpus design, and the use of acted versus naturalistic data—contributes to unstable findings, particularly for less prototypical emotions such as fear, surprise, and disgust. Nevertheless, it is important to highlight that Nunes & Teixeira (2012) analyzed spontaneous emotional speech in EP and compared the acoustic parameters with those of emotions produced by an actor and they concluded that although actors tend to exaggerate the expression of emotions, the acoustic properties are similar, thus showing that the ecological validity of simulated data is not necessarily low.

In Brazilian and European Portuguese, few studies systematically compare the influence of sentence mode, lexical-semantic content, and voice quality in emotional expression, and most rely on small sample sizes or single-speaker corpora. Furthermore, the acoustic similarity between certain emotions—such as joy and neutrality in some contexts—raises questions about the role of speaker variability, cultural display rules, and prosodic ambiguity. These gaps underscore the need for acoustic protocols and integrated models that account for both biological mechanisms and sociolinguistic context in vocal emotion production.

In order to comprehend the acoustic properties of emotional vocalizations, it is essential to investigate not only how emotions are produced, but also how listeners perceive and interpret these vocal cues. Understanding emotional vocalizations requires an integrated approach that considers both biologically driven signal encoding and culturally shaped decoding processes.

Building on this foundation, the next section will present findings related to the perception of emotions through suprasegmental features, with a focus on prosody as a key channel for emotional decoding.

2.3. Perception

The perception of emotions through vocal input has been a dynamic area of inquiry, evolving from basic acoustic mapping to complex neural processing models. Early foundational studies laid the groundwork for understanding how specific vocal cues signal emotional states, while recent research has highlighted how these perceptual mechanisms develop, adapt, and vary across individuals.

Perception of emotional vocalizations involves the listener's ability to decode the acoustic signals and interpret the speaker's emotional state. The idea that vocal emotion perception relies on decoding acoustic signals like pitch, intensity, rhythm, and voice quality - and mapping them onto discrete or dimensional emotional categories - is widely supported in affective science and speech communication research (Banse & Scherer, 1996; Scherer et al., 2001; Scherer, 2003; Juslin & Laukka, 2003, 2005; Bänziger et al., 2014; Giordano et al., 2021).

In one of the earliest comprehensive studies, Banse and Scherer (1996) demonstrated that discrete emotions can be reliably distinguished using acoustic profiles such as pitch, intensity, and tempo, even when semantic content is neutral. A large-scale perceptual experiment was applied involving six languages – German, English, French, Italian, Spanish and Danish - using semantically neutral utterances composed of seven syllables constructed from syllables common to those six European languages and produced by professional actors. The acoustic analysis extracted measurable vocal parameters, such as fundamental frequency (F_0), intensity, speech rate, and voice quality indicators (e.g., jitter, shimmer, harmonic-to-noise ratio). The results of the perceptual experiment showed that these features reliably indexed not only emotional arousal (i.e., the degree of activation or physiological intensity) but also valence, distinguishing between positive (e.g., joy, pride) and negative emotions (e.g., sadness, anger, shame). These findings indicate that listeners are sensitive to a multidimensional acoustic code of emotion and that these cues are interpretable across different cultural backgrounds, even in the absence of semantic content.

A further seminal study made by Scherer, Banse, and Wallbott's (2001) investigated the universality of vocal emotion recognition across nine culturally and linguistically diverse countries in Europe, North America, and Asia. The researchers used standardized vocal stimuli produced by German actors, who conveyed the emotions of anger, sadness, fear, joy, and neutral states. To

ensure the stimuli were "language-free" and that listeners relied solely on suprasegmental cues rather than semantic content, the actors performed the recordings using a pseudolinguistic technique: they recited the same neutral sentence in German, but replaced its actual words with a repetitive, semantically void nonsense phrase. This cross-cultural study revealed three fundamental findings about vocal emotion recognition. First, participants across all cultures identified emotions with 66% accuracy (significantly above chance), exhibiting similar confusion patterns like mistaking sadness for fear, suggesting universal decoding mechanisms. Second, recognition rates varied culturally from 74% in Germany to 52% in Indonesia, decreasing as participants' native languages became more distant from German, indicating that while core recognition abilities appear biologically grounded, cultural experience fine-tunes perceptual sensitivity. Third, recognition accuracy consistently followed an emotion-specific pattern, with high-arousal states like joy and anger being identified more accurately than low-arousal emotions like sadness across all cultures, likely reflecting the greater evolutionary importance of detecting high-arousal signals. These findings collectively demonstrate that vocal emotion perception operates through both universal biological mechanisms and culture-specific learning processes.

Pell and Kotz (2011) investigated how quickly and accurately Swedish listeners identify emotions from prosody using an auditory gating paradigm. They presented pseudo-sentences conveying six emotions (anger, disgust, fear, sadness, happiness, neutrality) in progressively longer segments ("gates"). Results showed that recognition unfolds incrementally, with fear, sadness, and neutrality identified earlier, while happiness and disgust required more speech input. Anger was detected at a moderate pace (~600 ms). Confidence increased with longer exposure, confirming that emotion recognition relies on cumulative acoustic cues. The authors highlighted that vocal emotion perception is a dynamic process, with detection speed varying by emotional category. Emotions like fear and sadness, marked by distinct acoustic features, are recognized faster than happiness or disgust, which depend on subtler cues. These findings support models of emotion perception as a gradual inferential process rather than an instantaneous one, emphasizing the role of temporal structure in prosodic processing.

Chen et al. (2012) explored how sound intensity influences vocal emotion perception behaviorally and neurophysiologically. In their first experiment, participants rated anger levels in Mandarin utterances with angry or neutral prosody under varying intensity conditions (original, increased, decreased). Results showed that increased intensity amplified perceived anger only in angry prosody, not neutral, suggesting intensity enhances existing emotional cues rather than acting independently. In the second experiment, EEG and ERP measurements revealed that mismatched emotional prosodies (e.g., neutral voice in an angry context) triggered greater

cognitive conflict. Reducing intensity weakened and delayed these neural responses, demonstrating intensity's role in shaping emotional salience and processing. The study highlights intensity as a critical modulator of emotional perception, urging researchers to treat it as a meaningful parameter in affective speech analysis.

In the same line of research, Giordano et al. (2021) explored how the brain represents vocal emotions—categorically (e.g., anger, joy) or dimensionally (e.g., valence, arousal)—using fMRI and MEG. Participants listened to nonverbal emotional vocalizations (e.g., laughs, cries), and Representational Similarity Analysis (RSA) compared neural patterns to both models. Behavioral tasks assessed emotion recognition, intensity, valence, arousal, and perceptual dissimilarity. Results showed a temporal shift: early processing (<200 ms) favored categorical coding in fronto-temporal regions, while later stages (240–500+ ms) reflected dimensional representations, especially for valence and arousal, in limbic-temporal networks. These findings demonstrate that categorical and dimensional models operate at distinct processing stages. The early categorical response suggests rapid detection of discrete emotions, while the later dimensional shift reflects integrative evaluation of affective qualities. Giodarno et al. (2021) underscore the importance of neural dynamics in vocal emotion processing, highlighting how acoustic and contextual information is sequentially integrated. Furthermore, the authors observed that neural mechanisms operate independently of linguistic content, emphasizing the role of prosodic cues in affective communication.

Da Silva et al. (2016) conducted a cross-cultural study examining how BP and Swedish listeners perceived emotional prosody in both languages. Their research employed culture-specific rating scales and principal component analysis (PCA), which uncovered two underlying dimensions accounting for more than 94% of response variance, demonstrating significant perceptual consistency across cultures. Through regression analysis, the study identified several acoustic features - including fundamental frequency (F_0), spectral tilt, jitter, and shimmer - as significant predictors of emotional interpretation. This study revealed significant cross-cultural similarities in emotional prosody perception despite linguistic differences, indicating that robust acoustic cues facilitate intercultural understanding of vocal emotions. When acoustic signals were clear and coherent, both Brazilian and Swedish listeners demonstrated comparable interpretation patterns, suggesting the primacy of acoustic/prosodic cues over language-specific or culture-specific factors in emotional speech perception. These findings strengthen evidence for biologically grounded, acoustic-based mechanisms in vocal emotion perception across diverse listener populations.

Peres (2022) also examined the interplay between universal and culture-specific factors in vocal emotion recognition by comparing native BP listeners with non-native English-speaking participants. Spontaneous but acted BP speech samples expressed four basic emotions (anger, fear, joy, sadness), which were analyzed for eight acoustic parameters: pitch contour, tempo, F_0 mean, F_0 range, F_0 variability, jitter, shimmer, and Hamonic-to-Noise Ratio (HNR). Results demonstrated significantly higher recognition accuracy among native listeners compared to non-natives, with statistical analyses confirming robust group differences. The findings identified pitch, tempo, and jitter as particularly salient cues that were more effectively interpreted by native speakers, suggesting that linguistic and cultural familiarity enhances emotional prosody decoding. These results contrast with universality claims in vocal emotion perception, instead supporting a dual-process framework where biological acoustic cues interact with culturally acquired interpretive schemas. Peres (2022) pointed out how language-specific experience shapes emotional prosody processing, with native listeners demonstrating superior utilization of subtle prosodic features. This research contributes to the growing literature demonstrating that while certain acoustic cues may have universal salience, their interpretation remains substantially mediated by cultural-linguistic experience.

In EP, Menezes and Jesus (2014) demonstrated that EP listeners associate higher pitch and reduced shimmer with happiness (high arousal), whereas sadness (low arousal) is perceived through lower pitch and breathier voice quality—directly mirroring Teixeira (2009) production data. This consistency suggests a tight coupling between how emotions are vocally produced and how they are decoded by listeners in European Portuguese. These findings support Scherer's (2003) Component Process Model, which posits that vocal emotion features evolve due to their perceptual discriminability. The recurrence of high F_0 in both production (Teixeira, 2009) and perception (Menezes & Jesus, 2014) of high-arousal emotions in EP underscores that such cues are not arbitrary but functionally adaptive. For example, anger's high pitch and intensity likely serve to amplify signal salience in threat contexts, while sadness's low pitch and shimmer may signal vulnerability.

Castro and Lima (2010), already mentioned in section 2.2, also observed how listeners of EP perceive emotions through acoustic-prosodic features. Using both semantically neutral sentences and pseudosentences, the researchers assessed the extent to which prosody alone can convey emotional meaning. Their findings demonstrated that EP listeners could accurately identify emotional content from prosodic cues alone, with recognition rates reaching 75% for full sentences and 71% for pseudosentences. Interestingly, response times were faster for pseudosentences, suggesting that the absence of semantic content may facilitate quicker

emotional judgments. These results provide compelling evidence for the sufficiency of prosodic information in vocal emotion perception, corroborating prior cross-linguistic studies that emphasize the primacy of suprasegmental cues in emotional communication (Scherer et al., 2001).

To wrap up, research on vocal emotion perception has consistently demonstrated that listeners can reliably interpret emotional states based on acoustic cues alone, even in the absence of semantic content (Banse & Scherer, 1996; Scherer et al., 2001; Juslin & Laukka, 2003; Castro & Lima, 2010). Across languages and methodologies, certain prosodic features—such as fundamental frequency (F_0), intensity, speech rate, and voice quality—emerge as robust indicators of emotional categories and dimensions (Scherer, 2003; Bänziger et al., 2014; Giordano et al., 2021). Voice quality parameters such as jitter, shimmer, and harmonic-to-noise ratio further contribute to the perceptual differentiation of emotions, particularly in distinguishing negative states (Chen et al., 2012; Castro & Lima, 2010; Menezes & Jesus, 2014). Furthermore, while arousal-related cues are generally well recognized, valence continues to be less reliably perceived through vocal input alone, suggesting a need for further exploration of subtle prosodic markers and listener-specific factors such as cultural familiarity or emotional sensitivity. These patterns are not only consistent across studies but are also grounded in biological and evolutionary models of affective communication (Scherer, 2003; Elfenbein & Ambady, 2002; Juslin & Laukka, 2003).

Despite this substantial progress, notable gaps remain. Research on vocal emotion perception is still disproportionately focused on a small set of widely spoken languages, leaving underrepresented linguistic varieties (such as BP and EP) relatively understudied. While some recent work has addressed this imbalance, including studies by Teixeira (2009), Castro and Lima (2010), and Peres (2022), more systematic cross-linguistic and cross-cultural comparisons are needed to determine the extent to which perceptual patterns are universal or shaped by language-specific prosodic norms. Additionally, few studies have fully integrated production and perception data within the same linguistic system, limiting our understanding of how emotional prosody is both encoded and decoded in natural communication.

This research seeks to characterize how the emotions happiness, sadness, and anger are expressed and recognized in BP and EP. By examining the acoustic patterns identified in the production of these emotions, it is provided a foundation for analyzing their perception considering the prosodic systems of these two linguistic varieties. The interplay between production and perception is particularly critical for understanding how listeners interpret vocal signals within their

linguistic and cultural context, making this dual perspective essential for a comprehensive analysis of emotional prosody.

Chapter 3 – Cross-Linguistic and Cross-Cultural Comparisons in Brazilian and European Portuguese

3.1. Introduction

The study of emotional prosody across languages highlights how linguistic and cultural factors shape the production and perception of emotions in speech. Intonation, as a suprasegmental feature, plays a crucial role in conveying affective meaning (Banse & Scherer, 1996), yet its patterns vary depending on phonological systems and sociocultural norms (Pell, 2001; Scherer, Banse, & Wallbott, 2001). However, as far as we know, this research topic is not explored yet for language varieties.

BP and EP exhibit significant differences in pitch range, contour types, and rhythmic structure, as we detail in the sections below, which may influence how emotions like happiness, sadness, and anger are encoded and interpreted. In this chapter the main prosodic differences between BP and EP are summarized. Hypotheses are then drawn on how these prosodic patterns may shape the perception of emotions both within and between the two varieties.

3.2. Cross-Linguistic Comparisons: prosodic profile

Despite sharing a mixed rhythmic pattern (Frota & Vigário, 2000, 2001; Frota et al., 2015), BP and EP diverge considerably in their intonational systems. BP is characterized by higher tonal density, marked by frequent pitch movements and a wider pitch range, which contributes to a dynamic and expressive speech style (Frota et al., 2015; Frota & Moraes, 2016; Cruz-Ferreira, 1998). These prosodic features align with BP's broader use of pitch accents and boundary tones, as documented in P-ToBI analysis (Frota et al., 2015). This tonal richness is further enhanced by BP's tendency to segment speech into smaller prosodic units, which plays a crucial role in sentence disambiguation and reflects a distinct intonational phrasing strategy (Frota & Vigário, 2001).

In contrast, EP exhibits fewer tonal events and more restricted pitch variation per utterance, resulting in a comparatively restrained intonation (Frota & Vigário, 2000, 2001). This distinction is further supported by acoustic studies showing that BP's "chanted" intonation (Frota et al., 2015) contrasts with EP's flatter contours, which may influence cross-variety emotional perception (Castro & Lima, 2010). Furthermore, Cruz & Frota (2012) describe EP as favoring longer prosodic phrases with more sustained pitch levels. It is plausible that these acoustic characteristics could result in a reduced perceptual salience of emotional cues compared to the more dynamic intonation of BP.

Rhythmically, although BP and EP exhibit a mixed pattern, BP combines syllable-timed and mora-timed characteristics, while EP combines syllable-timed and stress-timed properties (Frota & Vigário, 2000, 2001; Barbosa, 2006; Meireles & Barbosa, 2012). This fundamental difference is realized through specific phonological processes: BP often employs vowel epenthesis, a process which further modulates its rhythmic structure and enhances its perceived syllabic regularity (Barbosa, 2006; Frota et al., 2015; Meireles, 2010). In contrast, EP is marked by vowel reduction and deletion (Vigário, 2000, 2003), processes that may pose a challenge for listeners accustomed to the more syllable-based rhythm of BP.

Table 3 presents the comparison between BP and EP key prosodic features based on Frota and Vigário (2000, 2001) and Frota et al. (2002, 2015).

Table 3 - Key prosodic differences between Brazilian and European Portuguese summarized into a table (sources: Frota & Vigário, 2000, 2001, Frota et al., 2002, 2015).

Prosodic feature	Brazilian Portuguese	European Portuguese
Tonal density	Rich variety of pitch accents (H, L); frequent tonal events and expressive pitch movements	Fewer pitch accents; less pronounced tonal variation
Intonational Phrasing	Short phrases	Long phrases[1]
Rhythm	Mixed rhythm (syllable- and mora-timed); vowel epenthesis.	Mixed rhythm (syllable- and stress-timed); frequent vowel reduction (and deletion)
Expressivity	Dynamic, melodic, often perceived as “chanted”	Restrained, flatter intonation profile

Given the prosodic differences between BP and EP, we preview perceptual consequences for how emotions are expressed and interpreted, both within and between these two varieties. Namely, because BP’s intonational system contributes to a more dynamic and expressive vocal style (Frota et al., 2015; Frota & Moraes, 2016), and since these features align with Scherer’s componential model of vocal emotion expression, which associates high pitch, increased intensity, and rapid tempo with emotions such as joy, surprise, and anger (Scherer, 1986; Banse & Scherer,

[1] In this study, “European Portuguese” refers specifically to the standard variety. It is important to note that other EP varieties may exhibit shorter intonational phrases. See Frota et al. (2015) for a comparative analysis of intonational phrasing across EP varieties.

1996), we hypothesize that **BP's prosodic profile may enhance the perceptual salience of high-arousal emotions, leading listeners to an easy and fast recognition of these emotions produced by BP speakers** (Hypothesis 1). On the other hand, because EP's prosodic profile tends to be characterized by a **more restrained melody** (Frota & Vigário, 2000, 2011), we hypothesize that this style **aligns more closely with low-arousal (such as sadness) or neutral emotional states**, resulting in an **easy and fast recognition of these emotions produced by EP speakers** (Hypothesis 2). Finally, considering the acoustic correlates of high-arousal emotions (Scherer, 1986; Banse & Scherer, 1996), and the fact that EP listeners rely on subtle acoustic cues, such as duration, to infer affective meaning (Castro & Lima, 2010), we hypothesize that **pitch range, speech rate, and intensity will be the most prominent cues for cross-variety emotion recognition, with salience varying by emotion and listener linguistic and cultural background** (Hypothesis 3).

Ultimately, the prosodic asymmetry between BP and EP highlights the importance of considering intonational structure in cross-linguistic studies of vocal emotion. It reinforces the need for variety-specific analysis in affective speech research and, if our hypotheses are confirmed, this will support the broader argument that vocal emotion expression is shaped by an interplay between universal acoustic principles and language-specific prosodic strategies.

3.3. Cross-Cultural Comparisons in Vocal Emotion Expression

Research in affective communication has long acknowledged that certain aspects of emotional vocal expression are also shaped by culture-specific display rules - socially constructed norms that regulate the expression and suppression of emotions across cultural groups (Ekman & Friesen, 1969; Hofstede, 1980; Matsumoto, 1990). These rules point to how emotions are vocally produced, as well as how listeners interpret emotional cues in speech.

Although core acoustic parameters are employed cross-culturally to convey emotions (Juslin & Laukka, 2003; Banse & Scherer, 1996; Sauter et al., 2010), the magnitude, frequency, and contextual appropriateness of these vocal cues may vary significantly depending on cultural profiles and expectations. For instance, individualistic cultures tend to promote open emotional expression, especially for high-arousal, self-enhancing emotions such as anger or pride (Ekman & Friesen, 1969; Hofstede, 2001; Matsumoto, 1990). Speakers from these cultures often employ greater pitch excursions, higher speech intensity, and more expressive prosodic dynamics (Matsumoto, 1990; Scherer, 1997; Laukka et al., 2015). In contrast, collectivist cultures, which prioritize group harmony and social cohesion, tend to value emotional restraint and discourage overt expression of negative or disruptive emotions (Matsumoto et al., 2008; Mesquita, 2001; Tsai

et al., 2006). This leads to more subdued vocal cues — flatter pitch contours, reduced intensity, and slower tempos (Matsumoto et al., 2008; Elfenbein et al., 2007; Laukka et al., 2016).

The distinction between individualistic and collectivist cultures has been a central framework in cross-cultural psychology, particularly in understanding emotional expression. This conceptual dichotomy was notably developed by Geert Hofstede (1980, 2001), who identified individualism–collectivism as one of the key cultural dimensions differentiating societies. This dimension reflects the degree to which individuals are integrated into groups and the value placed on personal autonomy versus group cohesion.

In individualistic cultures, social structures tend to emphasize autonomy and personal independence (Hofstede, 1980, 2001; Triandis, 1995; Markus & Kitayama, 1991). According to Hofstede's cultural dimensions theory (1980, 2001), these societies are characterized by loose social ties, where individuals are expected to take care of themselves and their immediate family, rather than relying on extended kinship networks or communal support systems. This orientation fosters a cultural environment in which personal goals, self-expression, and individual rights are prioritized over collective responsibilities. Hofstede (1980) empirical data exemplify individualistic cultures such as those found in the United States, Australia, and many Western European nations. They score exceptionally high on the Individualism Index (e.g., 91), indicating a strong cultural preference for independence, personal initiative, and self-reliance (Hofstede, 1980).

In collectivist cultures—commonly found in regions such as Asia, Africa, and Latin America—individuals are socialized from birth into strong, cohesive in-groups, often comprising extended family networks and tightly knit communities (Hofstede, 1980). These societies emphasize interdependence, relational harmony, and group loyalty, with personal identity closely tied to social roles and communal affiliations (Hofstede, 1980; Triandis, 1995). Emotional expression in collectivist contexts is often regulated by cultural display rules that prioritize emotional restraint, particularly in situations that could disrupt social cohesion or threaten group stability (Matsumoto et al., 2008).

Brazil and Portugal occupy intermediate positions on the individualism-collectivism spectrum, but Brazil's lower individualism score (38 vs. Portugal's 63) reflects stronger collectivist leanings (Hofstede, 2001; Triandis, 1995). This duality is acoustically realized in their prosodic systems, for instance, BP's high tonal density and wide pitch range (Frota et al., 2015) mirror its relational exuberance — a blend of communal values and expressive freedom (Gudykunst & Ting-Toomey, 1988; Levine & Norenzayan, 1999).

EP's sparser pitch accents (Frota & Vigário, 2001) are consistent with its high uncertainty avoidance (99/100, Hofstede, 2001) and hierarchical culture, reflecting a tendency toward prosodic moderation to preserve stability in speech (Cruz-Ferreira, 1998; Fernandes, 2007).

From a cultural perspective, and taking into account Hofstede's individualism–collectivism framework, we thus hypothesize that **Brazil's more collectivist orientation, together with a richer and more complex prosodic profile than EP, will promote a higher perception rate of emotions produced by non-native speakers of their variety (EP, in this case). At the same time, Portugal's comparatively higher individualism score coupled with strong uncertainty avoidance, favoring prosodic restraint, will result in a higher difficulty (i.e., lower perception rate) in recognizing emotions produced by non-native speakers of their variety - BP** (Hypothesis 4).

By examining both acoustic salience and cultural mediation, this study aims to contribute to a deeper understanding of cross-variety emotion perception and the interplay between prosody and culture. The following chapter will detail the methodological approach employed to test these hypotheses, including participant selection, stimulus design, and analytical procedures.

Chapter 4 - Methodology

4.1. Introduction

This chapter presents the methodological procedures adopted in this study. In Section 4.2, the materials used in the experiment are described, including the selected sentences and their linguistic characteristics. Section 4.3 details the recording process of the stimuli, specifying the selection criteria for the speakers and the technical conditions of the recording sessions. In Section 4.4, the pre-processing and acoustic analysis steps are explained, with emphasis on the acoustic and prosodic parameters considered for the selection of the final stimuli. Finally, Section 4.5 describes the experimental design and procedure of the perception task, as well as the pool of participants who collaborated in the perceptual study, and the statistical analyses employed.

4.2. Materials

Three basic emotions, as defined by Ekman (1992)—happiness, sadness, and anger—were selected for analysis. A neutral emotion, characterized by the absence of intense prosodic variation, was also included to serve as a reference point for participants.

The linguistic material consisted of three syntactically simple sentences selected from Castro and Lima (2010), all composed of high-frequency words and equal length (nine phonological syllables):

(i) Esta mesa é de madeira ("This table is made of wood.");
(ii) As pessoas vão a concertos ("People go to concerts.");
(iii) O futebol é um desporto[2] ("Soccer is a sport.").

This corpus was chosen for three main reasons: the sentences' simple syntactic structure minimizes cognitive load and reduces variability in prosodic production unrelated to emotional expression; the use of high-frequency words ensures familiarity for all participants, regardless of educational background, thus avoiding lexical effects on prosody; and the equal number of phonological syllables allows for direct comparison of duration, pitch, and intensity patterns across conditions. Additionally, these sentences had already been tested and successfully employed by Castro and Lima (2010) in their investigation of emotional prosody in Portuguese, ensuring methodological consistency and comparability of results across studies.

[2] Due to the lexical difference of the word sport in BP (*esporte*) and EP (*desporto*), the stimuli recorded for BP was *O futebol é um esporte*. The number of syllables remained the same.

4.3. Stimuli Recording

The sample for the recording of the stimuli consisted of two professional actresses[3], one Brazilian and one Portuguese, both female. Selection criteria included formal academic training in acting and proven ability to modulate emotional prosody. Female voices were chosen due to their acoustic properties, which are less susceptible to phenomena such as creaky voice. Both participants read the Statement of Research Objectives and signed the Consent Form authorizing the use of the data (i) within the scope of this master's thesis, conferences, or other events with similar purposes, (ii) as a contribution to the development of future scientific research, and (iii) for the creation of databases for scientific applications.

Regarding the actresses profiles, DJS, 27 years old, is originally from Brazil and holds both a bachelor's degree and a master's degree in Acting. She has been living in Lisbon since 2021. Her academic background, combined with practical experience, enables her to produce a wide range of intonational patterns and emotional expressions. AMG, 21 years old, was born in Lisbon, lived in other places in Portugal and returned to her home city, where she lives for 10 years. She has been acting for four years and studying theatre for over seven, holding a bachelor's degree in Acting. Her extensive experience in the field allows for precise and expressive production of the prosodic patterns required for this study.

The recording sessions, lasting approximately 30 minutes each, were conducted individually in the Phonetics and Phonology Laboratory at the University of Lisbon, in a semi-soundproof environment. Each actress produced the selected sentences in a neutral tone and in three emotional expressions (happiness, sadness, and anger), aiming for a natural yet clear emotional intonation. No specific instructions were given regarding pitch, pace, or intensity, although feedback was provided by the researcher to ensure authenticity. Recordings were made in .wav format using Pro Tools LE 5.1.1 software, a high-quality microphone, and an Apple Macintosh G4 computer, with a sampling rate of 41 kHz and 16-bit resolution.

4.4. Stimuli Pre-processing and Acoustic Analysis

Audio segmentation was performed using Praat (Boersma & Weenink, 2022), a tool for phonetic analysis, to create individual files per sentence/emotion for consistent analysis. A total of 52 stimuli underwent acoustic analysis focusing on the following parameters: minimum, maximum, and mean F0 (Hz), mean intensity (dB), and speaking rate (i.e., number of

[3] Studies comparing acted and spontaneous emotions suggest that there are no significant differences in vocal production, since both convey similar acoustic patterns (Juslin & Laukka, 2001; Scherer, 2003; Nunes, 2020).

phonological syllables divided per utterance total duration) in order to ensure that emotions produced fitted the acoustic properties pointed out in Castro & Lima (2010). Additionally, an intonational analysis was also performed, using P-ToBI (Frota et al., 2015), a standardized transcription system for Portuguese prosody, to ensure adherence to tonal patterns described for each variety (Frota & Vigario, 2000; Frota et al., 2015).

Based on acoustic and intonational consistency, a subset of 24 stimuli (12 per variety; 3 per emotion) was then selected to be used in the perception experiment. Table 4 displays the averaged acoustic and prosodic characterizations for the selected stimuli only.

Table 4 - Average Acoustic Patterns and Prosodic Characterizations of BP and EP selected stimuli.

Variety	Emotion	Min F0 (Hz)	Max F0 (Hz)	Mean F0 (Hz)	Mean Intensity (dB)	Speaking Rate (syll/sec)	Dominant Pattern
BP	Neutral	138.9	264.3	212.3	64.9	5.49	H+L* L%
	Happiness	153.4	401.4	317.5	78.0	4.59	H* HL%
	Sadness	141.7	255.4	205.2	59.3	4.28	H+L* L%
	Anger	143.0	365.7	279.9	77.0	4.93	H+! H* HL%
EP	Neutral	162.4	248.6	206.0	66.1	6.57	H+L* L%
	Happiness	191.2	387.0	305.9	75.2	5.27	H* + L L%
	Sadness	140.4	253.4	209.5	58.5	5.14	H* L%
	Anger	173.1	400.9	297.7	78.0	4.69	H* + L L%

These stimuli were then low-pass filtered at 400 Hz to mask segmental information, emphasizing suprasegmental features and avoiding variety recognition by the participants. This way, a potential bias of the participants' sociocultural constructs in their responses was avoided.

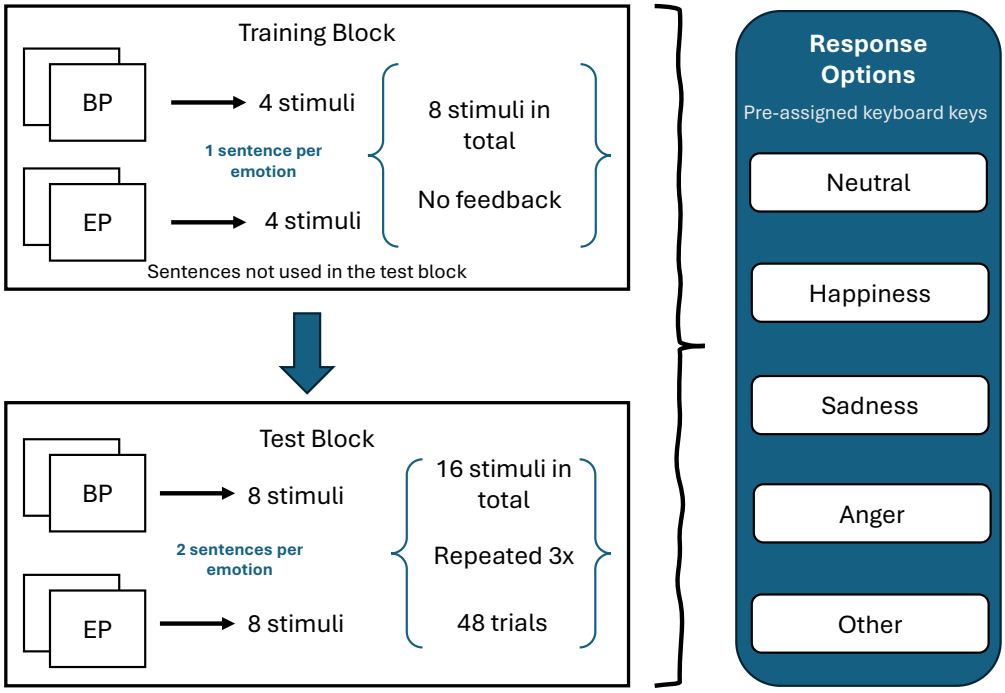
4.5. Experimental Design of the perception task and Procedure

An overt emotion identification task was implemented in SuperLab 6.4 (Cedrus Corporation), consisting of a training block followed by a test block. The training block contained eight stimuli: one sentence per emotion (neutral, happiness, sadness, anger) for each variety (BP, EP). Importantly, the sentences used in the training block were not used in the test block, ensuring that participants could not rely on memorization. The test block consisted of 16 different stimuli (eight per variety), each repeated three times in randomized order, totalizing 48 test trials. In both blocks, participants selected one of five response options — neutral, happiness, sadness, anger,

or other — using pre-assigned keyboard keys. Responses were open-ended in the sense that no cues about correctness were provided, and no feedback was given at any stage of the experiment.

A schematic representation of the experimental design is presented in Figure 2, illustrating the sequence of the training and test phases, the stimulus distribution, and the exposure to both varieties.

Figure 1: Schematic of the experimental design.



All participants were exposed to both native and non-native stimuli, allowing for both within- and between-subject comparison analyses. Thus, the native variety of participants (BP/EP), the variety they perceived (native/non native), and the emotion type (neutral, happiness, sadness, anger or other) were the independent variables considered in the analysis. As dependent variables, accuracy (i.e., correctness in the identification of emotions) and reaction times (in milliseconds) were measured.

Data collection was conducted remotely: EP participants completed the task in Lisbon, BP participants in Brazil. In both cases, SuperLab TaskPlayer was temporarily installed on participants' own laptops to ensure controlled presentation and timing, as the task was set to run only once per participant. They were instructed to wear headphones while doing the task, and they were informed that they would be listening to acoustically modified sounds and that repetition

was not allowed. At the end of the task, their responses were automatically stored in the SuperLab cloud.

4.6 Participants

The study involved 14 BP and 14 EP native speakers aged 18–50 years (M=32, SD=5; 15 females, 13 males), with no hearing, language, or cognitive impairments. Written consent was obtained, and the study was approved by the Ethics Board of University of Lisbon.

4.7 Statistical Analysis

Accuracy (i.e., correctness in the identification of emotions) was analyzed using Chi-square tests to evaluate whether the correct identification of emotions was related to: (i) the participants' native variety (BP vs. EP), (ii) the perceived variety (native vs. non-native), and (iii) the emotion type (neutral, happiness, sadness, anger). Emotions identified as 'other' by the participants were not included in the inferential statistical analysis; they were only explored descriptively.

In order to inspect if any of the acoustic/prosodic parameters detailed in section 4.4 allows to predict participants' accuracy, a multinomial logistic regression was run with correctness as dependent variable, and minF0, maxF0, meanF0, mean intensity, speaking rate and dominant pattern as covariates. Since we also wanted to observe whether these predictors are influenced by the participants' native variety, the perceived variety or the emotion under perception, these three independent variables were also included as factors in the same analysis.

Reaction times (RT) were analyzed using a Generalized Linear Mixed Model (GLMM) in SPSS 26 (IBM Corp., 2019). The model included four fixed effects: (i) participants' native variety (BP vs. EP), (ii) perceived variety (native vs. non-native), (iii) emotion type (neutral, happiness, sadness, anger), and (iv) correctness of responses (correct vs. incorrect). In addition, three interactions were tested: (i) correctness × emotion, to examine whether RTs differed for correct and incorrect responses across emotions; (ii) correctness × emotion × native variety, to investigate whether native variety modulated RT differences between correct and incorrect responses for each emotion; and (iii) correctness × emotion × perceived variety, to determine whether the perceived variety (native vs. non-native) influenced RT patterns for correct and incorrect responses across emotions.

This chapter described the methodological framework adopted in the present study. All steps were carefully planned to ensure methodological consistency, acoustic control, and comparability with previous research on emotional prosody in Portuguese. By combining

controlled stimuli with a balanced participant sample and robust analytical methods, this approach provides a solid basis for investigating how Brazilian and European Portuguese listeners perceive emotions across varieties.

The following chapter presents the results obtained from this experimental procedure, highlighting the patterns that emerged in the identification of emotions and the reaction times both within and between participant groups.

Chapter 5 – Results

This chapter presents the results of the perception experiment, aiming to assess how Brazilian and Portuguese speakers perceive and identify emotional prosody considering four basic emotions: neutral, happiness, sadness, and anger. The analysis focus on emotion recognition accuracy, assessed in terms of percentage of correctness (section 5.1), and reaction times (RTs), measured in milliseconds (ms) (section 5.2), considering the effects of participants' native variety (BP/EP), perceived variety (native/non-native), and emotion type. In section 5.3, a multinomial logistic regression analysis is run in order to observe which acoustic/prosodic properties in the signal explain the results obtained in the perception task, and in section 5.4, the main findings are summarized.

5.1. Identification of emotions

The overall accuracy rate across all participants and conditions was 71.7%, indicating that participants were generally successful in identifying emotions based on prosodic cues. Table 5 summarizes the accuracy rates per emotion.

Table 5 - Overall Analysis of Accuracy Rates per Emotion.

Emotion	Accuracy (%)
Neutral	79.5
Happiness	64.6
Sadness	74.1
Anger	68.5
Average Accuracy	**71,7**

As shown in Table 5, the highest accuracy was observed for neutral productions (79.5%), followed by sadness (74.1%). Happiness exhibited the lowest accuracy rate (64.6%).

When observing the identification of emotions in more detail, in order to assess how each intended emotion was perceived by participants (Table 6), we may conclude that happiness and anger triggered some difficulties, as 18.1% of happiness stimuli were misclassified as anger, and 16.2% of anger stimuli were misclassified as happiness. Neutral and sadness were less frequently confused with other emotions, thus suggesting that these two emotions are easily identifiable based on prosodic information.

Table 6 - Confusion matrix (%) showing how each intended emotion was perceived by participants.

Intended/Perceived	Neutral	Happiness	Sadness	Anger	Other
Neutral	79.5	6.3	8.4	3.6	2.2
Happiness	8.2	64.6	4.3	**18.1**	4.8
Sadness	9.5	5.7	74.1	6.5	4.2
Anger	4.1	**16.2**	6.9	68.5	4.3

While the overall confusion matrix provides a broad overview of emotion identification, the relatively consistent rate of "Other" responses (ranging from 2.2% to 4.8%) warrants a deeper investigation. A critical question is whether these "Other" responses are distributed randomly across all stimuli or if they are concentrated on a few specific items. The analysis of the frequency of "Other" responses per individual stimulus revealed that the responses were not randomly distributed. Instead, they were highly concentrated on a very small subset of the stimuli. In particular, EP tokens of happiness were responsible for the largest amount of “Other” classifications. Importantly, this pattern was observed in participants from both varieties: BP listeners showed a peak of “Other” responses when perceiving EP happiness tokens (8 cases), and EP listeners likewise chose “Other” responses predominantly for EP happiness (7 cases). By contrast, BP tokens elicited far fewer “Other” responses for happiness, and the few misclassifications were spread more evenly across emotions. This cross-variety consistency indicates that the difficulty does not stem from a single listener group, but rather reflects particular properties of the EP happiness stimuli themselves, which in turn helps explain the lower identification accuracy for this emotion.

This finding strongly suggests that the "Other" category was not used as a mere "I don't know" option for a generally difficult task. Rather, it was a specific response triggered by particular stimuli that were perceived as acoustically anomalous or emotionally ambiguous. These stimuli likely contain prosodic contours or voice qualities that do not neatly align with the prototypical acoustic profiles of Neutral, Happiness, Sadness, or Anger for our participant pool. This could be due to factors such as the perceived intensity of the emotion (e.g., a happiness stimulus that sounded more like euphoria or hysteria), the presence of mixed emotional cues, or even idiosyncratic production features from the speaker that made the intended emotion less clear.

When looking at the mean accuracy rates by native variety, i.e., BP versus EP participants (Table 7), independently of the variety under perception, we may observe that EP participants exhibited a slightly higher overall accuracy (74.1%) compared to BP participants (69.3%), suggesting that EP participants are better than BP ones in perceiving emotions. This seems to contradict our Hypothesis 4, according to which BP participants—due to Brazil's collectivist cultural orientation and a more expressive prosodic system—would outperform EP participants in recognizing emotions produced by speakers of the other variety. Instead, the data show that EP participants achieved higher overall accuracy, suggesting that factors beyond cultural dimensions may be influencing emotion perception. One possible explanation is that EP participants, despite their prosodic restraint, may rely more consistently on specific acoustic cues, leading to more stable interpretations across stimuli. Alternatively, the stimuli themselves—particularly those produced in BP—may contain prosodic features that are less universally interpretable, thereby affecting BP participants' performance when perceiving EP stimuli. Another plausible factor is the high degree of exposure Portuguese citizens have to Brazilian Portuguese through cultural and media channels. The presence of a large Brazilian community in Portugal, along with the widespread consumption of Brazilian soap operas, music, literature, and cinema (or even, more recently, YouTubers' contents), may contribute to a greater familiarity with BP prosodic patterns. This frequent contact could facilitate perceptual adaptation, allowing EP participants to more effectively decode emotional cues in BP speech, despite the prosodic differences between the varieties.

Table 7 - Accuracy Rates by Native Variety.

Native variety	Accuracy
BP	69.3%
EP	**74.1%**

Table 8 presents accuracy based on whether the stimulus was perceived as native or non-native to the listener.

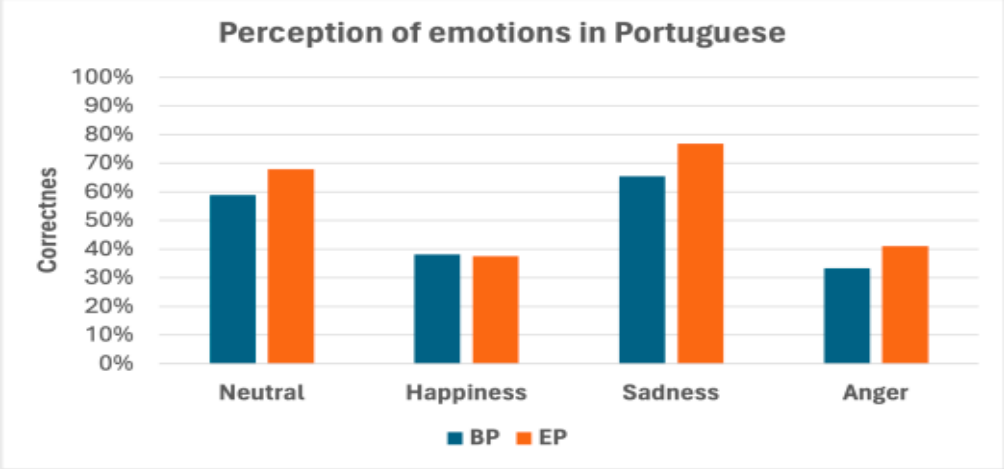
Table 8 - Asymmetrical Native Advantage in Emotion Perception.

Participant Variety	Native stimulus	Non-native stimulus
BP listeners	72.8%	55.4%
EP listeners	66.67%	57.8%

The results in Table 8 reveal a clear trend: both BP and EP participants demonstrated higher accuracy when perceiving emotional stimuli produced in their own native variety. BP listeners achieved **72.8%** accuracy with BP stimuli, compared to **55.4%** with EP stimuli. Similarly, EP listeners reached **66.67%** accuracy with EP stimuli, versus **57.8%** with BP stimuli. These findings suggest that native prosodic patterns facilitate emotion recognition, likely due to greater familiarity with the intonational contours, rhythm, and expressive norms of one's own variety. This aligns with Hypotheses 1 and 2, which posit that BP and EP participants perceive emotions differently, and supports the idea that intonational familiarity enhances perceptual accuracy.

To further explore how emotional perception varies across Portuguese varieties, Figure 2 breaks down accuracy rates by emotion, highlighting patterns in listener performance when exposed to native stimuli.

Figure 2 - Perception of emotions by Brazilian Portuguese and European Portuguese native participants.



In order to evaluate the factors influencing emotion identification, a series of Chi-square tests were conducted to determine if the accuracy of participants' responses was significantly related to (i) their native variety (BP vs. EP), (ii) the variety being perceived (native vs. non-native), and (iii) the type of emotion (neutral, happiness, sadness, anger).

Our results show that participants' correct responses were significantly related to their native variety ($\chi^2(1)=4.32$, $p<.05$), and to the emotion under perception ($\chi^2(3)=93.87$, $p<.001$). As we may observe in Figure 3, for both BP and EP listeners, neutral and sadness yielded the highest accuracy. Happiness and anger showed lower accuracy and higher confusion, particularly

between each other. Conversely, the variety being perceived (native vs. non-native) did not play a statistically significant role in identification accuracy (p>.05). However, a detailed observation of the data suggests a different behavior between BP (Figure 4) and EP (Figure 5) participants.

Figure 3 - Perception of emotions by Brazilian Portuguese participants, produced by speakers from their native (BP) and non-native (EP) varieties.

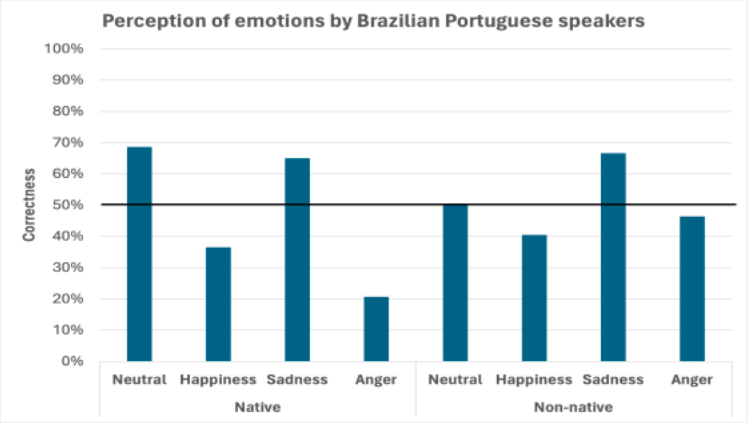
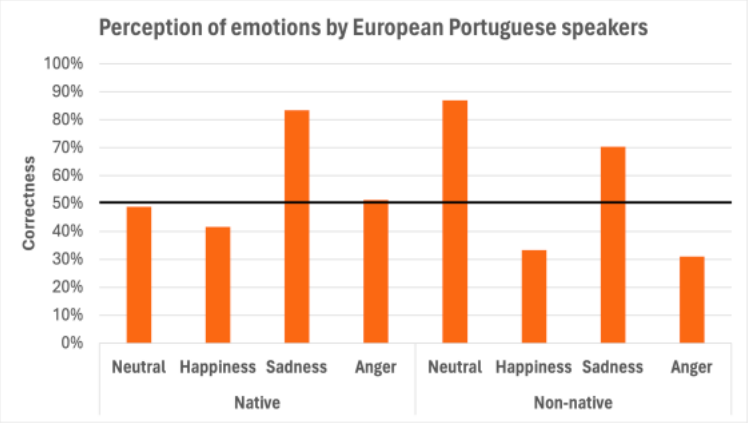


Figure 4 - Perception of emotions by European Portuguese participants, produced by speakers from their native (EP) and non-native (BP) varieties.



Namely, EP participants are better than BP participants in identifying emotions produced by non-native speakers. This might be related to the fact that intonation in BP is more chanted (i.e., exhibits higher variability) than in EP ([Frota & Vigário, 2000; Frota et al. 2015), and it also explains why BP participants are better than the EP ones in identifying emotions produced by speakers from their native variety.

Also interesting, by comparing the results depicted in Figures 3 and 4, we may conclude that BP participants exhibit the highest accuracy rates for neutral and sadness emotions,

independently on the variety being perceived, thus partially confirming Hypothesis 1, according to which they would easily recognize happiness and anger emotions due to their high pitch, increased intensity, and rapid tempo (Sherer, 1986; Banse & Scherer, 1996; Castro & Lima, 2010). For EP participants, the highest accuracy rates are attained for neutral emotion and sadness, confirming Hypothesis 2.

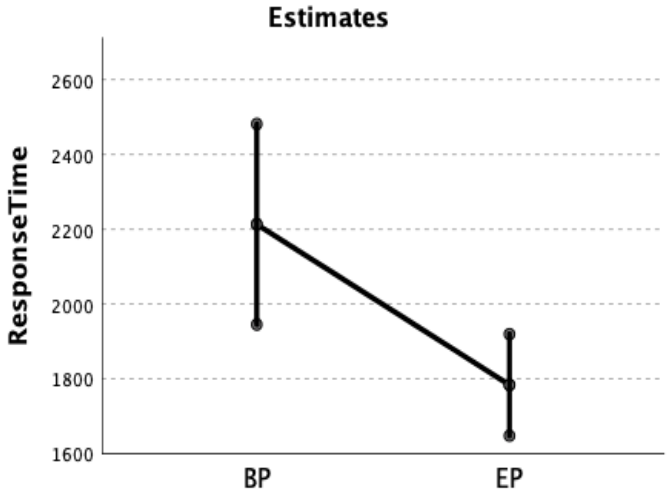
5.2. Reaction Times (RTs)

To investigate whether participants' reaction times (RTs) were influenced by linguistic background and stimulus characteristics, we conducted a Generalized Linear Mixed Model (GLMM). The model included RTs as the dependent variable and four fixed effects: (i) participants' native variety (BP vs. EP), (ii) the perceived variety (native vs. non-native), (iii) the type of emotion (neutral, happiness, sadness, anger), and (iv) correctness of responses (correct vs. incorrect). In addition, interactions between correctness and emotion, as well as their modulation by native and perceived variety, were also tested.

Our results showed a significant effect of participants' native variety ($F(1, 1300)=8.41$, $p<.01$), correctness ($F(1, 1300)=6.34$, $p<.05$), and emotion under perception ($F(3, 1300)=13.10$, $p<.001$). The interaction correctness*emotion under perception shows a borderline effect ($F(3, 1300)=2.49$, p=.059), and like for accuracy in the identification of emotions, the variety being perceived does not play a relevant role ($p>.05$) for RTs.

Considering firstly the effect of participants' native variety, indeed we may observe, in Figure 5, that, overall, EP participants responded faster (M=1784ms) than the BP ones (M=2213ms).

Figure 5 - RTs (ms) considering the native variety of the participants.



As for correctness, Table 9 displays the overall mean reaction times (RTs) for correct and incorrect responses, as well as the respective standard errors, and minimum and maximum values.

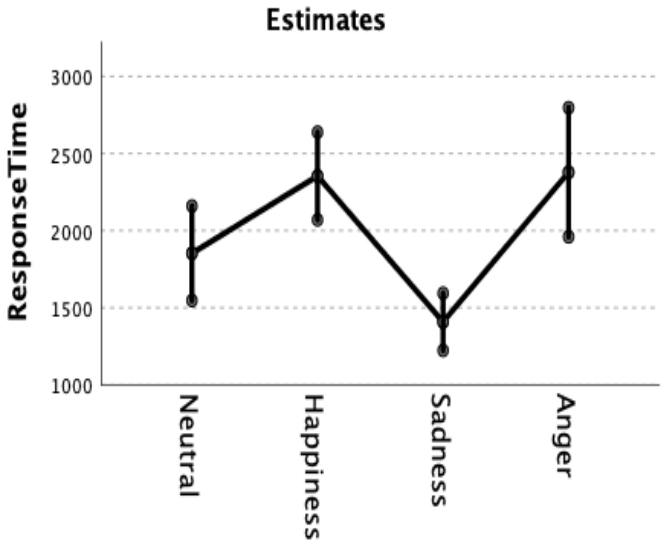
Table 9 - Overall Reaction Times (ms.) per correctness (mean standard error, and minimum and maximum values).

Response Type	Mean RT (ms)	Standard Error (ms)	Min/Max (ms)
Correct	1799	92	1620-1979
Incorrect	2197	129	1945-2450

We may observe that correct responses were produced faster (*M*=1799 ms) than incorrect ones (*M*=2197 ms), as expected (Schneider, Dogil & Möbius, 2011).

Considering now the effect of the emotion under perception, and as illustrated in Figure 6, sadness triggers the shortest RTs (*M*=1408ms), whereas happiness and anger trigger the longest RTs (*Ms*=2354 and 2379ms, respectively).

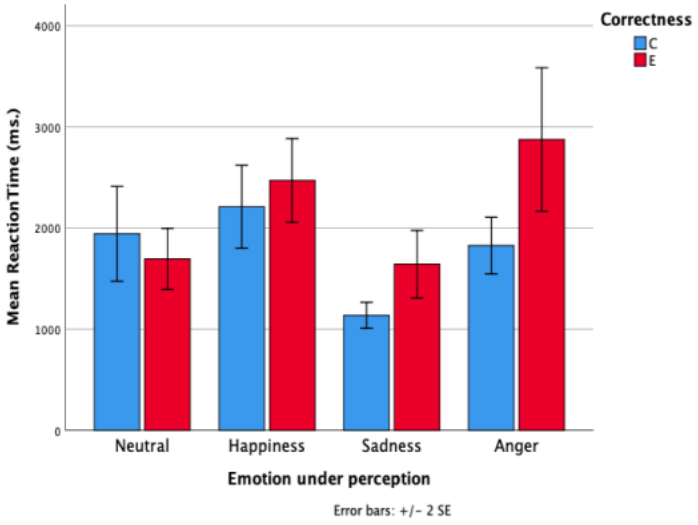
Figure 6 - RTs considering each emotion under perception.



These results are aligned with the accuracy rates observed in section 5.1, i.e., sadness exhibits the second highest accuracy rate (after the neutral emotion) and it is the faster to be identified, and happiness and anger display the lowest accuracy rates and the longest RTs, thus being the most difficult ones to be perceived.

Looking now at the borderline effect of the interaction correctness*emotion under perception, illustrated in Figure 7, we may observe that the longest RTs were registered for incorrect responses given for happiness and anger, thus reinforcing the conclusion that these two emotions are the most difficult ones to perceive.

Figure 7 – RTs considering the correctness of participants' responses for each emotion under perception.



In the same line of thought, but in an opposite way, the shortest RTs were registered for correct responses given for sadness, thus highlighting the conclusion that this is the easiest emotion to be perceived.

5.3. Acoustic/prosodic properties predicting the perception results

As mentioned in section 4.4, in the methodology chapter of this thesis, we also aimed at observing if any of the acoustic/prosodic parameters of the stimuli used in the perception task allows to predict participants' accuracy. Thus, a multinomial logistic regression was run with correctness as dependent variable, and minF0, maxF0, meanF0, mean intensity, speaking rate and dominant pattern as covariates. Since we also wanted to observe whether these predictors are influenced by the participants' native variety, the perceived variety or the emotion under

perception, these three independent variables were also included as factors in the same analysis, and each of them was also included in the model in two-way interactions with each covariate.

The chosen method of regression was *forward entry*, meaning that an initial model is defined, containing only the constant (b_0), and that then the software looks for the predictor that best predicts the outcome variable; in other words, it selects, from the whole list of inserted covariates and factors, the one that has the highest correlation with participants' correctness.

As a result, the model selected the interaction between the emotion under perception and the dominant pattern as the best predictor of participants' correctness, as illustrated in Table 10.

Table 10 - Summary of the multinomial regression model that predicts participants' correctness.

Step Summary

Model	Action	Effect(s)	Model Fitting Criteria			Effect Selection Tests		
			AIC	BIC	-2 Log Likelihood	Chi-Square[a]	df	Sig.
0	Entered	Intercept, MinF0, MaxF0, MeanF0, MeanIntensity, SpeakingRate, DominantPattern, NativeVariety, Emotion under perception, PerceivedVariety	220.161	282.602	196.161			
1	Entered	Emotion under perception * DominantPattern	170.456	243.304	142.456	53.704	2	0.000

Stepwise Method: Forward Entry

a. The chi-square for entry is based on the likelihood ratio test.

Thus, this means that, independently of the participants' native variety and of the variety under perception, participants' correctness is explained by the dominant pattern per emotion, i.e.,

although stimuli were segmentally masked, the intonational properties in the signal differ per emotion and this is a cue used by participants in the identification of emotions.

As we may observe in Table 11 the predicted values are not significantly different from the observed values. Thus, the model is a good fit of the data.

Table 11 - Chi-square goodness-of-fit test results for the regression model.

Goodness-of-Fit			
	Chi-Square	df	Sig.
Pearson	12.959	18	0.794
Deviance	12.979	18	0.793

Indeed, when revisiting Table 4 in section 4.4, we may observe that neutral and sadness emotions display the simplest contours (i.e., all-falling melodies with simple boundary tones), whereas happiness and anger display complex boundary tones (in BP) or more prominent nuclear pitch accents (in EP – H*+L), which are more marked in the EP intonational system (Frota, 2012), thus, probably more difficult to be identified.

5.4. Summary of Results

In summary, the perception experiment revealed clear asymmetries in the recognition of emotions across BP and EP. Neutral and sadness were the most accurately identified emotions, while happiness and anger were often confused with each other, showing both lower accuracy rates and longer reaction times. Importantly, the category “Other” was not used randomly, but rather concentrated in a small subset of stimuli—particularly EP productions of happiness—suggesting that some tokens carried prosodic features that were perceived as ambiguous or atypical.

Overall, EP participants achieved slightly higher accuracy rates than BP participants, contradicting our Hypothesis 4, i.e., that BP listeners would outperform EP listeners. This may be explained by EP listeners’ greater exposure to Brazilian Portuguese through cultural and media channels, as well as by their consistent reliance on specific acoustic cues. Both groups, however, showed significantly better performance when perceiving stimuli from their own native variety, confirming that intonational familiarity facilitates emotion recognition.

Reaction time analyses further reinforced these findings. EP participants responded faster overall than BP participants. Correct responses were consistently faster than incorrect ones, and

sadness emerged as the fastest recognized emotion, aligning with its relatively high accuracy rate. By contrast, happiness and anger elicited the longest reaction times, especially when responses were incorrect, which highlights their status as the most difficult emotions to perceive.

Finally, the multinomial logistic regression showed that correctness was best predicted by the interaction between the emotion under perception and the dominant intonational pattern. This indicates that, despite segmental masking, listeners relied strongly on prosodic cues for emotion identification. The model was a good fit, confirming the robustness of this finding.

These results provide a comprehensive picture of how BP and EP listeners perceive emotional prosody. The following chapter will discuss these findings in light of our initial hypotheses and the broader literature, examining their theoretical implications for cross-varietal perception and the role of prosody in emotion recognition.

Chapter 6 - Discussion

This chapter aims to discuss the results obtained in the perception experiment, which sought to assess how BP and EP speakers perceive and identify emotional prosody in the four basic emotions under study (neutral, happiness, sadness, and anger). The results are organized and discussed considering the research questions raised in chapter 1 and the hypotheses drawn in chapter 3, briefly summarized in the sections below.

6.1. BP and EP Participants: In Search of a Distinctive Perceptual Pattern

The primary objective of this investigation was to determine whether speakers of two closely related yet prosodically distinct varieties of Portuguese - BP and EP - exhibit divergent patterns in their perception of emotional prosody (Research Question 1). A foundational question underpinning this research was whether the well-documented production differences between these varieties, such as BP's broader F0 range and more frequent use of rising contours compared to EP's more restrained and flat melodic patterns (Frota & Vigário, 2000; Frota et al., 2015), would lead to corresponding differences in perceptual processing. The analysis was structured around two key metrics: the accuracy of emotion identification and the reaction times (RTs) associated with these judgements, which together provide a view of the underlying perceptual mechanisms. The findings reveal a complex interplay between native variety, exposure, and cognitive processing that shapes the perception of emotional intent.

The present investigation confirms that prosody is a robust carrier of emotional information, with an overall identification accuracy of 71.7%. This result aligns with previous studies on emotional prosody across languages (Pell et al., 2009), demonstrating that listeners can effectively decode emotional states from suprasegmental cues even when semantic information is masked.

The highest accuracy rates were observed for neutral (79.5%) and sadness (74.1%). According to the literature, these emotions display more stable and prototypical acoustic profiles—characterized by a narrower F0 range, slower speech rate, reduced pitch variability, and stable vocal qualities—findings that are supported by Banse & Scherer (1996), for languages such as English, or Colamarco & Moraes (2008) and Nunes et al. (2010) for BP and EP, respectively. These acoustic properties make neutral and sadness emotions less susceptible to misinterpretation. Conversely, happiness (64.6%) and anger (68.5%) showed the lowest accuracy. As discussed in the previous chapter, these two emotions are more complex from a melodic point of view (see Table 4), frequently confused with one another (cf. results in chapter 5), and often

perceived as ambiguous due to their prosodic characteristics - complex boundary tones in BP or marked nuclear accents in EP (Frota, 2012), confirming they are the most challenging emotions to perceive based on prosody alone. The confusion between these two emotions (18.1% of happiness stimuli perceived as anger, and 16.2% of anger as happiness) suggests they share key acoustic features, such as high pitch and increased intensity (Banse & Scherer, 1996; Colamarco & Moraes, 2008), which listeners struggle to disambiguate without other contextual cues.

This hierarchy of recognition, in which neutral and sadness are identified more accurately than happiness and anger, is consistent with cross-linguistic evidence. For example, Pell et al. (2009) found that across multiple languages (including English, German, and Mandarin), low-arousal emotions such as sadness and neutral states tend to yield higher recognition rates, while high-arousal emotions like happiness and anger are more prone to confusion. Similarly, Sauter et al. (2010) reported that the recognition of sadness and neutral affect showed greater cross-cultural stability compared to more dynamic emotions such as happiness, anger, or fear. These findings suggest that certain acoustic cues associated with low-arousal emotions—such as slower tempo, reduced pitch range, and falling intonational patterns—are more universally interpretable, whereas high-arousal emotions share overlapping prosodic markers (e.g., higher F0 range and intensity), which increases ambiguity. By aligning these findings with the prosodic profile of BP and EP varieties of Portuguese, and aiming at answering to our Research Questions 1 (Do BP and EP native participants perceive emotions differently?) and 3 (How do BP and EP's intonational differences (Frota et al., 2015) shape perception of basic emotions?), we hypothesized that (i) BP's prosodic profile may enhance the perceptual salience of high-arousal emotions (such as happiness and anger), leading listeners to an easy and fast recognition of these emotions produced by BP speakers (Hypothesis 1); and (ii) EP's prosodic profile, being characterized by a more restrained melody (Frota & Vigário, 2000, 2011), aligns more closely with low-arousal (such as sadness) or neutral emotional states, resulting in an easy and fast recognition of these emotions produced by EP speakers (Hypothesis 2).

We observed that Hypothesis 1 was not supported, since BP participants did not recognize happiness and anger more accurately, whereas Hypothesis 2 was confirmed, as EP participants performed better with neutral and sadness. Thus, the present results suggest that emotion perception is shaped not only by variety-specific intonational systems but also by the emotion type (i.e., its acoustic properties), as the relative ease of identifying emotions like sadness and neutrality seems to be related with their more stable and prototypical acoustic profiles.

The analysis of the "Other" category provided a crucial insight into the limits of prosodic categorisation. Contrary to being a random "I don't know" response, its use was concentrated on a specific subset of stimuli: EP-produced happiness tokens. This pattern was consistent across listeners from both varieties, indicating that the issue lay not with the listeners' strategies but with the stimuli themselves. This concentration suggests that these specific EP happiness productions contained highly atypical or ambiguous acoustic properties—potentially an extreme F0 range, a particular voice quality, or a complex contour that deviated from a prototypical "happiness" affect for our participants. When faced with a stimulus that did not cleanly match any of the four target categories, listeners systematically defaulted to the "Other" label rather than forcing a likely incorrect choice. This finding underscores the importance of considering intra-variety production variability in perceptual studies and highlights that certain emotional expressions can fall outside common perceptual categories, creating islands of ambiguity even within a familiar variety.

Although descriptively both groups were more accurate with stimuli from their own variety (BP: 72.8% native vs. 55.4% non-native; EP: 66.67% native vs. 57.8% non-native), this trend was not statistically significant, as the perceived variety factor did not reach significance (p>.05). Instead, accuracy varied with listeners' native variety (EP > BP) and with the emotion perceived, with neutral and sadness outperforming happiness and anger. It is important to note that, because the perceived variety was not a significant predictor of accuracy, ~~Therefore,~~ our results cannot robustly support the classic native-language advantage effect described by Scherer, Banse, & Wallbott (2001) and Pell et al. (2009) as a definitive finding. Instead, the data may suggest a tendency in that direction, but any such effect in this experiment was not strong enough to be separated from random variance and was secondary to the significant effects of participant variety and emotion type.

A key finding of this study was the significant effect of participants' native variety on emotion recognition accuracy. Closer inspection of this effect, however, revealed a striking asymmetry that ran contrary to our initial predictions: EP participants demonstrated a significantly higher overall accuracy (74.1%) compared to BP participants (69.3%). This result, termed here the EP Superiority Effect, directly contradicts Hypothesis 4, which predicted that BP listeners, hypothetically aided by a more expressive prosodic system and cultural factors, would outperform EP listeners.

This counterintuitive finding answers our Research Question 4 (Do cultural dimensions (Hofstede, 2001) mediate prosodic emotion perception?) and can be explained by a confluence of factors centered on asymmetrical exposure and its perceptual consequences. The primary explanation lies in the well-documented, unidirectional nature of media and cultural flow between

Portugal and Brazil. EP speakers are extensively exposed to BP through a constant flow of media (Brazilian soap operas, music, movies, and digital content) and the presence of a remarkable Brazilian community in Portugal. This frequent and naturalistic contact constitutes a form of implicit perceptual training. It likely enhances EP listeners' flexibility, allowing them to develop a broader and more adaptive "perceptual map" that can accommodate the wider F0 ranges and more dynamic contours characteristic of BP emotional prosody. This asymmetry as a contact effect finds parallels in other linguistic contexts. For instance, studies have shown that Dutch listeners often outperform German listeners in perceiving German emotional prosody, an advantage attributed to the Netherlands' greater exposure to German media (van Bezooijen & Gooskens, 1999). This suggests a clear case of asymmetrical intelligibility, where the variety with greater cultural penetration (BP) becomes more intelligible to speakers of the other variety (EP) than vice versa.

This asymmetrical exposure likely does more than simply familiarize EP listeners with BP sounds; it may actively enhance their perceptual flexibility and cue-weighting strategies. Constant exposure to the more variable BP prosody could train EP listeners to attend to a broader set of acoustic parameters and to be more tolerant of acoustic deviation from their own variety's prototypes. This is akin to the perceptual benefits observed in bilinguals or musicians, where experience with multiple sound systems enhances auditory cognitive function (Varnet et al., 2015; Neumann, 2023). Consequently, EP listeners may develop a superior ability to deal with the acoustic variation in BP speech and focus on the core prosodic gestalt, leading to more efficient classification.

Beyond mere exposure, another factor may underpin the EP advantage. Namely, the concept of sociophonetic markedness (Trudgill, 1986; Kerswill & Williams, 2002) may be at play. The more expressive and variable nature of BP prosody, while highly effective within its own communicative context, may sometimes be perceived as acoustically "exaggerated" or less prototypical of a canonical emotional expression by listeners less familiar with it. This potential deviation from expected prosodic norms could hinder accurate recognition for BP stimuli when processed by EP listeners who, despite their exposure, may still hold EP productions as a subconscious baseline.

While the data did not support the initial hypothesis that a collectivist cultural background in Brazil would confer a perceptual advantage (Hypothesis 4), culture may still play a subtler role. The 'expressiveness' of BP prosody is itself a cultural product. The challenge for BP listeners may not be in decoding emotion from prosody within their own cultural context but in adjusting their perceptual strategies to decode a system (EP) that operates with a different cultural logic of

expressiveness—one that is more restrained and may place a higher value on other contextual cues beyond prosody. This cultural dimension of prosodic expectation warrants further exploration.

In summary, the EP Superiority Effect is likely not a reflection of an inherent perceptual deficit in BP listeners, but rather the outcome of an asymmetrical sociolinguistic environment that has trained EP listeners to navigate a wider range of prosodic variation, combined with differences in the inherent variability and cue weighting of the two systems themselves.

The Reaction Times (RT) data provide a complementary window into the underlying perceptual and cognitive processes, offering more than just a mirror of the accuracy findings but a deeper explanation for them. The robust pattern observed reinforces and enriches the conclusions drawn from accuracy alone and demonstrates a key universal aspect of emotional speech processing.

The finding that EP participants responded faster overall (1784 ms) than BP participants (2213 ms) further reinforces the asymmetry observed in accuracy, providing additional evidence for an EP perceptual advantage. When considered alongside the accuracy results, these RTs indicate that Hypothesis 1 was not supported, as BP listeners did not recognize happiness and anger more easily or rapidly, whereas Hypothesis 2 was confirmed, given that EP listeners indeed identified neutral and sadness with greater accuracy and speed. This significant difference in processing speed suggests that for EP listeners, the task involved a more efficient and automatic processing of the emotional cues, likely facilitated by their extensive exposure to both varieties. In contrast, the longer RTs for BP listeners point to a higher cognitive load and a more effortful decoding process. Another point to consider is that the greater acoustic variability in BP productions might inherently increase the cognitive load for all listeners, including native BP speakers. The wider range of possible realizations for a single emotion category could require more complex cue integration and decision-making processes. This is consistent with our data showing that BP participants not only achieved lower overall accuracy but also exhibited significantly longer reaction times, suggesting a more effortful and less automatic perceptual process compared to their EP counterparts.

The result that accurate responses were significantly faster (1799ms) than incorrect ones (2197ms) is a classic finding in perceptual psychology and psycholinguistics (Schneider et al., 2011). It strongly indicates that delays are a signature of uncertainty and conflict resolution. When the acoustic signal clearly matches a stored emotional prototype, recognition is swift. When the cues are ambiguous or conflict with expectations—as was often the case for happiness and anger—the cognitive system requires more time to resolve this ambiguity, often resulting in either

a delayed correct response or an error. These findings fit neatly within dual-process models of perception (Schirmer & Kotz, 2006; Kahneman, 2011), which posit a fast, automatic route for well-defined stimuli and a slower, resource-demanding route for ambiguous ones.

The emotion-specific RTs perfectly corroborate the accuracy data. This convergence of high accuracy with low RTs for sadness (1408ms), and low accuracy with high RTs for happiness and anger (~2350–2380ms), is not an isolated phenomenon but a well-established cross-linguistic pattern observed in the perception of emotional prosody across diverse languages (Pell & Skorup, 2008; Liu & Pell, 2012). This consistency reinforces the universality of the underlying cognitive mechanisms.

This pattern can be explained through the lens of processing load and cue distinctiveness. Low-arousal emotions like sadness and neutral are typically cued by a highly stereotypical and acoustically distinct prosodic profile (e.g., low pitch, slow tempo, falling contours). These canonical markers are consistently realized across speakers and varieties, allowing for rapid feature detection and classification with minimal cognitive effort. Conversely, the perception of high-arousal emotions like happiness and anger is inherently more cognitively demanding. Their acoustic signatures show significant overlap (e.g., high pitch, high intensity), creating direct competition during categorization and forcing the listener's cognitive system to resolve fine-grained cue differences. This ambiguity, combined with the potential for greater internal variability in their productions, directly increases decision time, resulting in the protracted RTs observed.

From an evolutionary perspective, this perceptual asymmetry aligns with the proposed communicative functions of different emotional states. Lower-arousal emotions such as sadness and neutrality are often acoustically marked by reductions in pitch dynamicity, energy, and speech rate (Juslin & Laukka, 2003; Scherer, 1986). These acoustic profiles are not only highly stable but also suggest a signal of lack of immediate threat or action, making them arguably easier to distinguish for a receiver. In contrast, high-arousal emotions like happiness and anger share a suite of urgency-related acoustic cues—including high pitch, increased intensity, and faster tempo—that evolved to rapidly capture attention and signal a need for immediate social response (Darwin, 1876). While highly effective as an alerting mechanism, this acoustic similarity in high-arousal states (Banse & Scherer, 1996; Sauter et al., 2010) can blur the distinctions between specific positive and negative valences, leading to the higher confusion rates observed in our study.

The analysis demonstrated that listeners' perception of emotion was not random but systematically guided by the intonational configurations most closely associated with each category. While Neutral and Sadness benefited from clear, stereotypical cues, the more complex

melodic shapes of happiness and anger, particularly in EP, increased ambiguity and decision times.

Additionally, the multinomial logistic regression analysis from Chapter 5 yielded a crucial and clarifying finding: the interaction between Emotion and Dominant Intonational Pattern was the single best predictor of participants' accuracy. This result is fundamental confirms that listeners were not guessing but were actively relying on the systematic prosodic gestalts—the specific phonological contours—embedded in the signal to make their decisions, even with all segmental information removed. This finding powerfully underscores that intonational structure is not merely an epiphenomenon of emotion but a core perceptual vehicle for its communication. In sum, it not only confirms Hypothesis 3 but also highlights the central role of intonational structure in shaping recognition outcomes.

The findings of this study collectively argue for a model of emotional prosody perception that integrates universal biological influences with variety-specific phonological structuring. While the recognition hierarchy (e.g., sadness/neutral > happiness/anger) and the cognitive mechanisms (longer RTs for ambiguous stimuli) show cross-linguistic commonality, the precise instantiation of these processes is filtered through the listener's native intonational grammar. We propose that perception is not a direct mapping from acoustic cues to emotion, but a two-stage process: first, the auditory system parses the continuous acoustic signal into discrete, language-specific intonational categories (e.g., H+L, L+H). Second, these categorized phonological units are mapped onto emotional meanings, a process heavily influenced by the frequency and conventionality of these mappings within the variety. This explains the native variety advantage (familiarity with the system), the EP superiority effect (asymmetrical acquisition of a second prosodic system), and the supreme predictive power of the dominant intonational pattern. This model positions emotional prosody not as an exception to linguistic relativity but as a prime example of it.

This two-stage model finds strong support in and extends existing theoretical frameworks. The first stage aligns with the concept of "categorical perception" in prosody, as discussed by Ladd (2008) and others, which argues that listeners perceive intonational contours not as acoustic continua but as members of discrete, phonologically defined categories. Our data robustly confirm this; listeners did not respond to raw F0 or duration values in isolation but to the holistic pattern they formed. The work of Frota et al. (2015a, b) is paramount here, as their development of P_ToBI provides the precise phonological inventory—the set of possible categories like H+L or L+H—that BP and EP listeners use to parse the signal. The regression

analysis proves that the accuracy of emotion identification is contingent on how well a stimulus instantiates one of these expected categories. A stimulus that is acoustically ambiguous between categories, like some of the EP happiness tokens that were frequently labeled "Other," leads to perceptual hesitation or failure, precisely because it cannot be cleanly categorized in the first stage of processing.

The second stage of our model, the mapping of phonological categories onto emotional meaning, is crucially mediated by the listener's linguistic and cultural experience. This directly refines Scherer's (2003) Brunswikian Lens Model. While Scherer posits a direct link between acoustic cues (proximal stimuli) and inferred emotion, our results demonstrate that this link is indirect and is gated by the phonological system. This variety-specific conventionalization echoes the findings of Prieto & Roseano (2010) in their cross-linguistic studies on question intonation, showing that the pragmatic meaning of tunes is language-specific. Our study demonstrates that this principle applies to the emotional domain.

Chapter 7 - Conclusion

This dissertation set out to investigate the perception of emotional prosody by native speakers of two closely related yet prosodically distinct varieties of Portuguese: Brazilian Portuguese (BP) and European Portuguese (EP). Grounded in evolutionary theories of emotion (Darwin, 1872; Ekman, 1992) and models of vocal expression (Scherer, 1986, 2003), the primary objective was to determine whether the well-documented production differences between BP and EP (Frota & Vigário, 2001; Frota et al., 2015) would lead to divergent patterns in perceptual identification of emotions. The study aimed to provide answers to four core research questions: (i) whether BP and EP listeners perceive emotions differently; (ii) which acoustic features are most salient in cross-variety recognition; (iii) how the intonational differences between the varieties shape the perception of basic emotions; and (iv) whether cultural dimensions, specifically Hofstede's (2001) individualism-collectivism index, mediate prosodic emotion perception.

The results revealed a complex picture of shared universal tendencies and variety-specific perceptual strategies. Firstly, while a numerical trend pointed towards a native variety advantage—aligning with the well-documented hypothesis that familiarity with one's own intonational system facilitates emotion recognition (Scherer et al., 2001)—this effect was not statistically significant in our model. Descriptive data showed that listeners from both varieties tended to be more accurate with native stimuli, and reaction time (RT) data suggested a trend towards more efficient processing of these stimuli. However, the statistical analysis confirmed that other factors—namely, the listener's own native variety and the emotion type—were the primary and significant drivers of performance. Secondly, a clear emotion-specific hierarchy emerged, consistent with cross-linguistic findings (Pell et al., 2009; Sauter et al., 2010). Neutral and sadness, characterized by low arousal and acoustically stable, falling contours, were identified with the highest accuracy and speed. In contrast, happiness and anger, high-arousal emotions sharing acoustic features like high pitch and intensity (Banse & Scherer, 1996), were frequently confused and elicited the longest RTs, confirming their status as the most perceptually challenging categories.

Contrary to the initial Hypothesis 4, which predicted a BP advantage based on its more expressive prosody and collectivist cultural background, an EP Superiority Effect was observed: EP listeners achieved significantly higher overall accuracy than BP listeners. This counterintuitive finding is best explained by the asymmetrical exposure between the two cultures. EP listeners' extensive familiarity with BP through media and social contact (van Bezooijen & Gooskens, 1999) has likely honed their perceptual flexibility, allowing them to navigate BP's broader prosodic

variability more effectively than BP listeners can decode EP's more restrained patterns. This interpretation is bolstered by the RT data, which showed EP listeners responded faster overall, suggesting a more automatic and less cognitively effortful decoding process.

It was also observed that listeners were not relying on isolated acoustic cues but on integrated prosodic gestalts—holistic phonological contours that are variety-specific (Frota & Vigário, 2001), which was revealed by the multinomial regression analysis, showing the interaction between emotion and the dominant intonational pattern as the single best predictor of accuracy. This result powerfully refines Scherer's (2003) Brunswikian Lens Model by introducing a crucial phonological filter: the ecological validity of an acoustic cue is not universal but is determined by its role within the specific intonational grammar of the listener's native variety.

The theoretical contribution of this work is therefore twofold. First, it provides robust empirical evidence from a understudied language pair, Portuguese, confirming that the perception of emotional prosody is governed by a combination of universal biological mechanisms (e.g., the ease of processing low-arousal states) and language-specific phonological structuring. Second, and more importantly, it proposes a model of system-relative emotional prosody, arguing for a two-stage perceptual process where continuous acoustic signals are first parsed into discrete intonational categories specific to a variety, which are then mapped onto emotional meanings.

Naturally, this study's conclusions are bounded by its methodological choices and it is imperative to acknowledge the limitations that were faced, as doing so not only situates the findings within their appropriate scope but also catalyzes a productive agenda for future research. Naturally, the conclusions drawn here are bounded by the methodological choices made.

First, the use of professionally acted stimuli—essential for controlling lexical content and ensuring acoustic–prosodic clarity—may have introduced exaggerated cues compared to spontaneous, naturalistic expressions (Scherer, 2003). Although prior work suggests this does not necessarily distort relative patterns of recognition (Juslin & Laukka, 2001; Scherer, 2003; Nunes, 2020), it nonetheless constrains ecological validity. Future paradigms must grapple with the tension between experimental control and real-world applicability, perhaps by complementing acted datasets with induced or naturalistic speech, as advocated by Pell & Skorup (2008). Non-linguistic vocalizations (e.g., laughs, cries, gasps) or standard sentences produced in genuinely elicited emotional states could help bridge this gap.

Second, the demographic profile of our participant pool, while sufficient for statistical analysis, represents a narrow segment of the population—predominantly university students. Foundational sociolinguistic research (Labov, 1972; Eckert, 2000) has shown that age, socioeconomic status, and geographic mobility strongly influence linguistic perception and

behaviour. The remote nature of the experiment, while increasing accessibility, also introduced uncontrolled variability in listening conditions (e.g., headphone quality, ambient noise), which may have added “noise” to the RT data, as suggested by sporadic reports of technical issues.

Third, the emotional scope was deliberately restricted to four basic emotions—neutral, happiness, sadness, and anger—following Ekman (1992, 2016). This choice leaves unexplored the perception of more socially complex or self-conscious states (e.g., pride, shame, boredom, sarcasm) that, according to constructionist theories (Barrett, 2017) and empirical work (Laukka et al., 2016), may rely on different cue-weighting strategies and exhibit even greater cross-varietal asymmetries.

These limitations directly inform concrete proposals for future research. The persistent confusion between happiness and anger—rooted in their shared high-arousal acoustic profile—is a prime candidate for multimodal investigation. Presenting the same auditory stimuli alongside congruent and incongruent dynamic facial expressions could test whether visual cues disambiguate acoustically similar states, and whether this effect differs between BP and EP listeners (Pell et al., 2011).

The behavioural data also point to differing cognitive loads between BP and EP listeners. Neuroimaging offers a promising avenue to unpack these mechanisms: EEG could capture millisecond-level neural responses (e.g., N400, P600) sensitive to emotional incongruity and conflict (Kotz & Paulmann, 2011), while fMRI could localise the networks involved, testing whether EP listeners show more efficient processing (Giordano et al., 2021).

A direct and integrated test of the variability hypothesis would involve a complementary production study with the same participants. By recording BP and EP speakers producing the same emotions, we could directly correlate a speaker’s degree of acoustic–prosodic variability with the perceptual accuracy and RTs they elicit, moving from inference to a rigorous, data-driven test of whether the greater expansiveness of BP prosody underlies its perceptual challenge.

In final remarks, this dissertation underscores that the voice is a powerful and nuanced channel for emotional communication, but its message is interpreted through the prism of one's native sound system. The perception of a smile in the voice or a note of anger is not a simple decoding of universal acoustic symbols but a sophisticated psycholinguistic act shaped by a lifetime of experience with a specific intonational grammar. The findings challenge purely universalist accounts of emotional prosody and highlight the profound impact of linguistic experience and cultural exchange on our most fundamental human ability: to understand how others feel. By bridging the gap between evolutionary theory, prosodic typology, and perceptual psychology, this work ultimately argues that to fully understand how emotions are communicated,

we must listen not only to the biological signal but also to the phonological system that gives it shape and meaning.

Bibliography

American Psychological Association. (2020). Publication manual of the American Psychological Association (7th ed.).

Banse, R., & Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. Journal of Personality and Social Psychology, 70(3), 614–636. https://doi.org/10.1037/0022-3514.70.3.614

Barbosa, P. A. (2006). Incursões em torno do ritmo da fala. Pontes Editores.

Boersma, P., & Weenink, D. (2022). Praat: Doing phonetics by computer (Version 6.3.01) [Computer software]. http://www.praat.org/

Castro, S. L., & Lima, C. F. (2010). Recognizing emotions in spoken language: A validated set of Portuguese sentences and pseudosentences for research on emotional prosody. Behavior Research Methods, 42(1), 74–81. https://doi.org/10.3758/BRM.42.1.74

Colamarco, V., & Moraes, J. A. (2008). A expressão vocal das emoções no português do Brasil: Um estudo de produção. Anais do V Congresso Internacional de Fonética e Fonologia.

Cruz-Ferreira, M. (1998). Aspectos da prosódia do português europeu. Edições Colibri.

Darwin, C. (1872). The expression of the emotions in man and animals. John Murray.

Ekman, P. (1992). An argument for basic emotions. Cognition and Emotion, 6(3–4), 169–200. https://doi.org/10.1080/02699939208411068

Ekman, P., & Friesen, W. V. (1971). Constants across cultures in the face and emotion. Journal of Personality and Social Psychology, 17(2), 124–129. https://doi.org/10.1037/h0030377

Fernandes, F. R. (2007). Ordem, harmonia e conflito: Estudos sobre a cultura portuguesa. Imprensa de Ciências Sociais.

Frota, S. (2012). The intonational phonology of European Portuguese. In S. Frota & P. Prieto (Eds.), Intonation in Romance (pp. 6–42). Oxford University Press.

Frota, S., Cruz, M., Fernandes, I., & Vigário, M. (2002). P-ToBI: Pretória-Toronto-Barcelona-Lisbon. Sistemas de notação prosódica. Descrição e comparação [Unpublished manuscript]. Laboratório de Fonética, Universidade de Lisboa.

Frota, S., Cruz, M., & Vigário, M. (2015). Intonational variation and change in Portuguese. In S. Frota & P. Prieto (Eds.), Intonation in Romance (pp. 235–280). Oxford University Press.

Frota, S., & Moraes, J. A. (2016). Intonation in European and Brazilian Portuguese. In W. L. Wetzels, J. Costa, & S. Menuzzi (Eds.), The handbook of Portuguese linguistics (pp. 141–163). Wiley Blackwell.

Frota, S., & Vigário, M. (2000). Aspectos da entoação do português europeu: A entoação declarativa. In M. Vigário, S. Frota, & M. J. Freitas (Eds.), A aquisição de língua materna (pp. 287–319). Colibri.

Frota, S., & Vigário, M. (2001). On the correlates of rhythmic distinctions: The European/Brazilian Portuguese case. Probus, 13(2), 247–276. https://doi.org/10.1515/prbs.13.2.247

Frota, S., & Vigário, M. (2011). Intonational variation in Portuguese: European and Brazilian varieties. In S. Frota, G. Elordieta, & P. Prieto (Eds.), Prosodic categories: Production, perception, and comprehension (pp. 123–145). Springer.

Hofstede, G. (2001). Culture's consequences: Comparing values, behaviors, institutions, and organizations across nations (2nd ed.). Sage.

IBM Corp. (2019). IBM SPSS Statistics for Windows (Version 26.0) [Computer software].

Juslin, P. N., & Laukka, P. (2003). Communication of emotions in vocal expression and music performance: Different channels, same code? Psychological Bulletin, 129(5), 770–814. https://doi.org/10.1037/0033-2909.129.5.770

Ladd, D. R. (2008). Intonational phonology (2nd ed.). Cambridge University Press.

Larrouy-Maestri, P., Magis, D., & Morsomme, D. (2024). The acoustic features of vocal emotion expression: A meta-analysis. Emotion Review, 16(1), 45–61. https://doi.org/10.1177/17540739231211256

Menezes, C., & Jesus, L. M. T. (2014). Vocal emotion perception in European Portuguese. Journal of Voice, 28(5), 583–592. https://doi.org/10.1016/j.jvoice.2014.01.007

Moraes, J. A., & Rilliard, A. (2016). The expression of emotions in Brazilian Portuguese: A study of declarative, interrogative, and imperative sentences. Speech Prosody 2016, 736–740. https://doi.org/10.21437/SpeechProsody.2016-150

Morton, E. S. (1977). On the occurrence and significance of motivation-structural rules in some bird and mammal sounds. The American Naturalist, 111(981), 855–869. https://doi.org/10.1086/283219

Nunes, C., & Teixeira, A. (2012). Análise acústica da fala emocional espontânea vs. atuada no português europeu. XXVIII Encontro Nacional da Associação Portuguesa de Linguística.

Nunes, C., Teixeira, A., & Freitas, J. (2010). Análise acústica da qualidade vocal na produção de emoções no português europeu. XXVI Encontro Nacional da Associação Portuguesa de Linguística.

Ohala, J. J. (1984). An ethological perspective on common cross-language utilization of F0 of voice. Phonetica, 41(1), 1–16. https://doi.org/10.1159/000261706

Pell, M. D., & Kotz, S. A. (2011). On the time course of vocal emotion recognition. PLOS ONE, 6(11), e27256. https://doi.org/10.1371/journal.pone.0027256

Pell, M. D., Monetta, L., Paulmann, S., & Kotz, S. A. (2009). Recognizing emotions in a foreign language. Journal of the Acoustical Society of America, 125(1), 468–478. https://doi.org/10.1121/1.3021302

Pell, M. D., & Skorup, V. (2008). Implicit processing of emotional prosody in speech. NeuroReport, 19(9), 921–925. https://doi.org/10.1097/WNR.0b013e328302c92b

Peres, J. (2014). A percepção de emoções básicas através da prosódia no português brasileiro [Master's thesis, Universidade de São Paulo].

Peres, J. (2022). Universal and culture-specific factors in vocal emotion recognition: A comparison of native and non-native listeners of Brazilian Portuguese. Language and Speech, 65(4), 899–923. https://doi.org/10.1177/00238309211050012

Plutchik, R. (1982). A psychoevolutionary theory of emotions. Social Science Information, 21(4–5), 529–553. https://doi.org/10.1177/053901882021004003

Sauter, D. A., Eisner, F., Ekman, P., & Scott, S. K. (2010). Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations. Proceedings of the National Academy of Sciences, 107(6), 2408–2412. https://doi.org/10.1073/pnas.0908239106

Scherer, K. R. (1986). Vocal affect expression: A review and a model for future research. Psychological Bulletin, 99(2), 143–165. https://doi.org/10.1037/0033-2909.99.2.143

Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. Speech Communication, 40(1–2), 227–256. https://doi.org/10.1016/S0167-6393(02)00084-5

Scherer, K. R., Banse, R., & Wallbott, H. G. (2001). Emotion inferences from vocal expression correlate across languages and cultures. Journal of Cross-Cultural Psychology, 32(1), 76–92. https://doi.org/10.1177/0022022101032001009

Schneider, J., Dogil, G., & Möbius, B. (2011). The recognition of emotional speech: A review. Speech and Language Technology, 13/14, 53–70. Polskie Towarzystwo Fonetyczne.

Teixeira, A. (2009). Produção e percepção de emoções na voz: Estudo acústico e perceptual no português europeu [Doctoral dissertation, Universidade de Aveiro].

Vigário, M. (2003). The prosodic word in European Portuguese. Mouton de Gruyter.

Wang, Y., Zhu, L., & Chen, Y. (2018). Emotional prosody in Mandarin Chinese: A comparison with English. Frontiers in Psychology, 9, 1499. https://doi.org/10.3389/fpsyg.2018.01499

Williams, C. E., & Stevens, K. N. (1972). Emotions and speech: Some acoustical correlates. The Journal of the Acoustical Society of America, 52(4B), 1238–1250. https://doi.org/10.1121/1.1913238