Alma Mater Studiorum · Università di Bologna

DIPARTIMENTO DI INTERPRETAZIONE E TRADUZIONE
Corso di Laurea magistrale in Specialized Translation (classe LM-94)

TESI DI LAUREA
in PROFESSION-BASED RESEARCH

The Fate of Interaction after Whisper Decoding Optimization: a Case Study on the KIParla Corpus

CANDIDATA:
MARTINA SIMONOTTI

Relatore:
Adriano Ferraresi

Correlatrici:
Caterina Mauri
Ludovica Pannitto
Maja Miličević Petrović

Anno Accademico 2023/2024
Appello marzo 2026

Acknowledgements

First and foremost, I would like to express my sincere gratitude to Professor Caterina Mauri for the trust she placed in me from day one, for introducing me to the extraordinary KIParla project, and for paving the way to so many opportunities. I am also deeply thankful to Professor Adriano Ferraresi for being a lighthouse not only for this thesis, but throughout the entire TraTec journey, and to Professor Maja Miličević Petrović for her constant availability and kindness.

A heartfelt thank you goes to Ludovica. It would be impossible to list everything I have learned thanks to your guidance throughout this computational adventure together. You have truly been a mentor in every sense of the word. I will never be able to fully express how much you have helped me grow academically and how much confidence you have instilled in me. I feel incredibly lucky to have had such an exceptional professional like you to guide me over the last year and a half.

I would also like to thank all those who, in different ways, contributed to this journey: Eleonora Zucchini, for her constant support and kindness during my internship; Professor Silvia Ballarè, who also contributed in shaping my internship with precious feedback; Jaka Čibej, who kindly introduced Ludovica and me to INCEpTION and provided essential advice on organizing the annotation process; and Gabriele Carioli, for introducing me to his tool, as well as for his patience and availability throughout the entire experimental phase.

A huge thank you goes to my family. To my mom Simona, my dad Roberto and my brother Alessio: these past few years have not been the easiest of our lives, but I feel that we are getting out of all the difficulties stronger than ever. Thank you for allowing me to come this far in my academic journey and for always taking care of me even when we are physically apart. A heartfelt thank you also goes to my dog, Lilli, the joy of my life.

A special mention goes to my wonderful grandparents, Antonia, Bruno and Pasqualina, to my uncles Massimo and Giovanni and to my aunt Cristina, for their unconditional love and constant support. I am so lucky to have such an amazing family.

I would like to dedicate this thesis to my beloved aunt Cristina. I still struggle to accept that we will not celebrate this milestone together. You left too early and unexpectedly, but I hope that, wherever you are, you are proud of me, as you always were.

I am also deeply grateful to my uncle Giorgio for his never-ending support and for introducing me, when I was only six years old, to two things that somehow summarize this TraTec journey: English and computers. Thank you to my aunt Simona, who taught me that even the strongest and most difficult battles can be faced and overcome.

Thank you to all the people who have been part of my life during these five and a half years in Forlì. To my TraTec friends and colleagues Jennifer, Davide, and Angela, with whom I shared both joy and tears. To all the Stragnocchi, for bringing happiness and lightness even in the most challenging moments. To my housemates Anna and Marta, who have truly been my home over the past few years.

Thank you to Davide, the love of my life and my best friend. You have been my anchor and my rock. Your support and love during these two years has been essential. Your presence has supported me not only in my academic journey, but also through the unexpected hardships that life has placed upon me during the past two years. I can see our future together more clearly every day, and little by little I feel that we are beginning to shape it. I love you.

A special thank you also goes to Raffaella, Oreste, and Morena, for welcoming me into the di Biase family and always making me feel loved and at home.

Last but not least, perhaps a little cliché, I would like to thank myself for the strength I have shown throughout this journey. I persevered even as two chronic illnesses completely reshaped my life. Today, I am proud of myself and deeply grateful for the resilience I discovered within me.

Thank you, sincerely, to everyone who has been part of this journey.

The sun will rise, and we will try again.
Truce – Twenty One Pilots

Abstract

End-to-end Automatic Speech Recognition (ASR) systems such as Whisper achieve high transcription accuracy; however, they are designed to prioritize semantically informative content and consequently suppress short interactional phenomena such as backchannels, repair sequences and filled pauses. These elements are brief, prosodically subtle, may contain truncations or repeated elements and are frequently produced in overlap, making them especially vulnerable to omission or normalization. This thesis examines how spontaneous conversational features are treated in automatic transcriptions of Italian speech, and whether different system configurations affect their representation in the resulting output. To this end, the acoustic model is kept fixed, while decoding parameters are optimized using two objective functions: a standard Word Error Rate (WER)-based pipeline and an Interaction-aware pipeline incorporating event-level weights. A subset of the KIParla corpus was manually annotated to create a gold standard, and ASR outputs were evaluated in terms of global WER, event-level match ratios, substitution and omission patterns, as well as overlap effects. Results show that decoding optimization exerts only a limited influence on overall accuracy. The Interaction-aware configuration does not substantially increase event preservation, but it does not degrade global performance and slightly stabilizes error dispersion in some cases. Recognition patterns emerge as strongly phenomenon-dependent: self-repairs are often preserved in linearized form, whereas backchannels are particularly vulnerable to overlap. Overlapping speech consistently reduces recognition probability across configurations, suggesting that parameter adjustments alone are insufficient to counteract the normalization bias of large-scale ASR systems.

Keywords: Automatic Speech Recognition, Whisper, KIParla, Conversation Analysis, spoken Italian, decoding optimization.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 General and Specific Objectives

Automatic Speech Recognition (ASR) systems based on large end-to-end neural models have achieved remarkable levels of performance in recent years. Models such as Whisper (Radford et al., 2023) are widely adopted in both research and commercial contexts due to their robustness across languages, recording condition and speaker variability. Despite these advances, most ASR systems are trained on monological and non-spontaneous speech and tend to normalize conversational input, often treating interactional phenomena as noise (Lopez et al., 2022; Yamasaki et al., 2023). In this respect, literature has highlighted systematic limitations, particularly in regard to elements such as short conversational words, interjections, disfluencies, overlapping speech, interruptions and paralinguistic cues (Liesenfeld et al., 2023; Lopez et al., 2022; Umair et al., 2022; Zayats et al., 2019).

From a linguistic perspective, these phenomena are central not only to interactional dynamics, but also to the broader analysis of spoken language, including monologic discourse. Backchannels (Section 2.2) are short listener responses (e.g., *mhmh*, *eh*, *okay*) that signal attention, alignment, or understanding. Repair mechanisms (Section 2.3) address problems in speaking, hearing, or understanding and can be initiated by the current speaker of the trouble source (self-repair) or by the listener (other-initiated repair). Filled pauses (Section 2.4), instead, reflect processes of speech planning and hesitation (e.g., *ehm*). In naturally occurring conversation, backchannels, repair initiators and filled pauses elements are typically brief, prosodically subtle

and often produced in overlap. Self-repairs, by contrast, unfold over multiple tokens and are frequently characterized by hesitation, reformulation or repetition. Although in different ways, these properties make such phenomena especially vulnerable to omission, normalization or reinterpretation in automatic transcriptions (Lopez et al., 2022).

Previous research showed that the limited representation of interactional phenomena in ASR output does not necessarily imply that relevant information is absent from the acoustic signal or from the model's internal representations. Rather, these phenomena tend to be selectively suppressed or transformed at later stages of processing, particularly during decoding, when the model probabilistically decides which hypotheses should be rendered as text. This implies that ASR output should be treated as the result of configuration decisions (Vitale et al., 2024b,a; Dinkar et al., 2023). However, while the impact of ASR systems on conversational phenomena has been widely discussed, less attention has been paid to the role of decoding configurations in shaping the textual visibility of such phenomena.

Against this background, the general objective of the present study is to investigate how decoding parameter choices influence the representation of interactional phenomena in automatic transcriptions of spontaneous Italian speech. The ASR model itself is kept fixed; what varies are the parameters that regulate the model's sensitivity to weak acoustic events, short utterances, and alternative hypotheses. More specifically, this study aims to: (1) quantify the extent to which different categories of interactional phenomena (backchannels, filled pauses, self-repairs and other-initiated repairs) are preserved or suppressed in ASR output; (2) compare two optimization strategies and assess whether decoding adjustments can reduce the systematic suppression of such phenomena; (3) examine the role of overlap in influencing recognition outcomes.

1.2 Workflow

This thesis follows a multi-stage workflow that combines manual annotation, automatic transcription, parameter optimization, and quantitative as well as qualitative analysis.

After establishing and providing clear conceptual definitions and linguistic framework for these phenomena, a subset of conversations from the KIParla corpus (Mauri et al., 2019a) was manually annotated to identify the interac-

tional phenomena under investigation, providing a gold standard for evaluation. Automatic transcriptions were then generated under different segmentation and decoding configurations. Decoding parameters were optimized using two alternative objective functions: a standard WER-based criterion and an Interaction-aware criterion designed to account for event-level preservation. The best-performing configurations were subsequently evaluated on unseen data. Finally, transcription outputs were aligned with the gold standard and analyzed at multiple levels, including global accuracy, event-level preservation, and overlap-sensitive recognition patterns.

1.3 Dissertation Structure

Chapter 2 provides the theoretical framework for the interactional phenomena investigated in this study, namely backchannels, repairs and filled pauses. Chapter 3 provides a comprehensive overview of Automatic Speech Recognition, tracing the historical evolution of modeling paradigms up to contemporary end-to-end neural approaches. Particular attention is dedicated to Whisper as a large-scale ASR model, as well as to the challenges posed by spontaneous conversational speech and to decoding optimization strategies relevant for the present study. Chapter 4 presents the KIParla corpus, describing its structure and the current resource building methodology. Chapter 5 details the methodological framework adopted in this thesis, including dataset selection, manual annotation procedures, ASR transcription pipeline, decoding configurations and evaluation metrics. Chapter 6 reports and discusses the quantitative and qualitative results of the analysis. Chapter 7 summarizes the main findings by answering to the research questions, outlines the limitations of the study and proposes directions for future research.

Chapter 2

Interactional Phenomena in Spoken Data

2.1 Chapter Overview

This chapter provides a theoretical background for the interactional phenomena under investigation. Unlike planned or monologic discourse, everyday interaction is shaped by continuous coordination between participants, who must manage turn-taking, understanding, engagement and repair in real time. Backchannels (2.2), conversational repair (2.3) and filled pauses (2.4) are three classes of interactional phenomena that contribute to such coordination and represent a direct consequence of the communicative spontaneity of conversational speech (Pernas and Borreguero Zuloaga, 2010).

As far as the analysis is concerned, backchannels will be analyzed from the hearer's perspective, focusing on how they employ strategies to display their ongoing analysis and engagement with the current talk. For clarity, the party producing the main turn will be referred to as the *current speaker*, while the party producing feedback will be defined as the *addressee* or *recipient*, recognizing that this latter role also produces talk (Goodwin, 1986).

This thesis makes a crucial distinction between backchannels and disfluencies. Although some vocalizations may be formally similar (e.g., *mh*, *eh*), they do not necessarily perform the same interactional function. While backchannels are produced by recipients as responses, filled pauses reflect the speaker's own speech production process. Previous work has shown that these phenomena may interact, for instance when speaker hesitations elicit

recipient feedback Ward and Tsukahara (2000), but their functions remain analytically distinct. This differentiation is essential both for the manual annotation and for evaluating how Whisper represents different types of interactionally relevant vocalizations.

All the examples discussed are extracted from the KIParla corpus (for a detailed description, see Chapter 4), specifically from the modules KIP (Goria and Mauri, 2018; Mauri et al., 2019b), ParlaTO (Cerruti and Ballarè, 2020a,b), KIPasti (Mauri et al., 2024b) and ParlaBO (Mauri et al., 2024a). All the phenomena under investigation provided in the examples will be highlighted in bold. Corpus-specific Jefferson annotation conventions can be consulted in Section 4.3.

2.2 Backchannels

Backchannels are verbal or non-verbal responses that listeners provide to the speaker to display recipiency and support for their ongoing turn (Dideriksen et al., 2019; Blomsma et al., 2024; Mereu et al., 2024). Although they are often described as brief responses, they are not necessarily restricted to single-unit forms: they may also be multi-unit sequences or short sentences (Mereu et al., 2024). In general, their function is to reinforce phatic contact and contribute to maintaining interactional organization (Pernas and Borreguero Zuloaga, 2010).

Despite their apparent simplicity, backchannels resist a single, unified definition. They have been variously denominated and defined as *reactive tokens* (Clancy et al., 1996), *acknowledgement tokens* (Jefferson, 1984) and *response tokens* (Gardner, 2001), however the most neutral and general term with respect to discourse function is *backchannel* (Ward and Tsukahara, 2000). An influential definition is provided by Ward and Tsukahara (2000), designed from the addressee's perspective. To be defined as such, a backchannel must follow three criteria: (1) responds directly to the content of an utterance of the other, (2) is optional, (3) does not require acknowledgement by the other. This definition has been widely adopted in empirical research to facilitate backchannel identification. However, the authors acknowledge that alternative definitions may be more appropriate depending on specific analytical goals. For corpus-based work such as a research based on KIParla, a clear operational definition is crucial to ensure consistency in transcription and annotation tasks.

Backchannels emerge from the co-presence and temporal coordination of participants, enabling immediate feedback to the speaker (Pernas and Borreguero Zuloaga, 2010). By signaling involvement, they establish trust and cooperation between participants and, in addition to improving the fluency and effectiveness of the communication, backchannels can also play a role in persuasion of the assertions of the other (Gratch et al., 2006). They can be described as a form of cooperative overlap or, from a turn-taking perspective, as a turn-yielding cue (Blomsma et al., 2024). In this sense, backchanneling represents the sign of shared common ground between speaker and recipient (Dideriksen et al., 2019) and is widely regarded as essential also for successful communication and the establishment of interpersonal relationships (Mereu et al., 2024).

They are typically produced at specific moments in the interaction, defined as *backchannel opportunity points (BOPs)*, where recipients recognize cues inviting feedback (Blomsma et al., 2024). People are unconsciously able to produce backchannels at interactionally appropriate moments and are equally sensitive to their interpretation, both in terms of their verbal and non-verbal realization and their absence, which equally carries interactional meaning (Gratch et al., 2006). Through these signals, listeners display interest, attention, understanding and agreement, and also contribute to manage the turn-taking process. When a recipient produces a backchannel, they are acknowledging that the current speaker is still in charge of the turn (Dideriksen et al., 2019; Blomsma et al., 2024; Mereu et al., 2024).

2.2.1 Backchannels in Conversational Organization

The foundational role of backchannels in conversational organization was first articulated by Yngve (1970), who challenged the traditional speaker-listener dichotomy by showing that conversation is not organized around a strict alternation between speaking and listening roles. Instead, participants are simultaneously engaged in both activities: while one interlocutor holds the turn, the other primarily listens but may still produce short vocalizations without taking the floor. Yngve therefore introduces the notion of *backchannel*, defining it as a secondary channel through which recipients produce brief vocalizations such as *yes* or *uh-huh* while the current speaker retains the turn. These elements serve to support the ongoing talk, allowing speakers to moni-

tor the quality of communication and adjust their production accordingly. In this sense, backchannels constitute a fundamental mechanism through which conversational coordination and mutual understanding are maintained (Yngve, 1970).

Later work in Conversation Analysis (CA) refined and contextualized the role of backchannels. Schegloff (1982), for example, stressed the importance of analyzing backchannels within their sequential and interactional environment, not as standalone phenomena. In CA, backchannels are treated as *continuers*: they display the recipient's understanding that a multi-turn unit is still in progress and that no speaker change is projected. However, beyond merely signaling attention or understanding, backchannels actively collaborate in the production of extended turns by explicitly passing up opportunities to take the floor or initiate repair. Therefore, extracting them from the specific context risks losing the interactivity that gives them functional significance.

This view was further developed by Ward and Tsukahara (2000), who argued against a uniform interpretation of backchannels as straightforward indicators of attention or understanding. While it is true that some display engagement or agreement, others may also signal boredom, skepticism or minimal recipiency, and in some cases there is little or nothing to be understood in the prior talk at all. In particular, it was observed that certain vocalizations, which are acoustically similar to backchannels, may instead function as disfluencies or post-completion vocalizations, therefore reflecting the speaker's own speech production process. More generally, Tsukahara and Ward emphasize that the meaning and function of backchannels cannot be determined in isolation or reduced to a fixed semantic value, as observed by Schegloff (1982), but emerge from their sequential placement and form the interactional environment in which they occur.

2.2.2 Sequential Placement and Functional Types

According to Blomsma et al. (2024), speakers provide backchannel-preceding cues at specific points in the conversation, through a variety of verbal and non-verbal behaviors. Verbal cues include prosodic patterns such as rising and falling intonation contours, as well as pitch lowering: for example, recipients are more likely to produce a backchannel after the speaker has

speaker	transcription
BO145	uno è abituato a ascoltare il cantautorato Italian o quello anche quello estero
BO139	**mh mh**
BO145	di un certo periodo vedrà che è sicuramente diverso da quello di adesso [però] >cioè< adesso la trap sfera ebbasta chi cazzo è altri
BO139	**[mh mh]**
BO145	>cioè< sono comunque tutti cantautori tra virgolette no?
BO139	**mh mh**
BO145	anche se uno dice eh mamma mia ma tu stai bestemmiando non è vero e invece sì
BO139	**mh [mh mh]**
BO145	[>cioè<]

Table 2.1: Continuers in BOA3018 (KIP).

lowered their pitch for at least 110 ms. Speaker pauses are also predictive of backchannels production and tend to occur after syntactically complete sentences. Backchannel opportunity points (BOPs) are not uniform events either, as they differ in their sequential and prosodic properties. This variability affects, in turn, the form and function of the backchannels they elicit.

Researchers have proposed different ways of classifying backchannels, depending on whether the focus lies on their informal properties or on their interactional functions. Within CA, however, priority is given to classifications grounded in sequential organization and turn-taking, as these best capture how backchannels operate in interactional context.

Goodwin (1986), for instance, differentiated between continuers and assessments. Continuers display passive recipiency and typically occur at the boundaries of turn-constructional units and frequently overlap with the speaker's talk. Such overlap is interactionally unproblematic and provides structural evidence that the backchannel is a supportive signal bridging the end of one unit and the beginning of the next. An example is provided in Table 2.1.

speaker	transcription
BOI115	no vabbè poi anche il fatto che nel mio corso non ci sono le sessioni
BOR031	**ah**
BOI115	ma si studia duran c(io)è [gli esami ci son sempre infat[ti il primo] esame l'ho dato:
BOR031	**[oddio]**
BOI115	dopo: (.) un mese,
BOR031	**ah**
BOI115	sì sì sì e ci sono anche oggi ho dato un'esame ci sono esami sem[pre:] e si studia anche durante il tirocinio anzi per forza
BOR031	**((ride)) madonna ((ride))** **okay [(.) wow]**

Table 2.2: Assessments in PBA030 (ParlaBO).

Assessments, by contrast, display an evaluative reaction to the specific content of the current unit, such as surprise, appreciation, disbelief or moral judgment. For this reason, they are interactionally treated as actions that require completion before the conversation moves forward: speakers commonly delay the initiation of subsequent units until the assessment has been completed. This behavior indicates that assessments are locally tied to what is being said (Goodwin, 1986). While Gardner (2001) did not include assessments among its definition of backchannel, it is also true that they align with the definition provided by Ward and Tsukahara (2000), that does not make explicit whether an evaluative stance can or cannot be considered a backchannel. Table 2.2 illustrates an instance of assessment occurring during another speaker's extended turn. The recipient's responses (*ah*, *oddio*, *madonna*, *wow*) display an affective and evaluative stance toward the speaker's talk and engage directly with its content. Consistent with Goodwin's findings, the assessment is brought to completion within the boundaries of the current unit and requires recognition before the conversation proceeds. The expressive and multimodal nature of the assessment, combining laughter, stance-taking particles, and lexical items, further distinguishes it from minimal continuers and highlights its role in displaying active participation.

Within this functional perspective, Jefferson (1984) demonstrated that

speaker	transcription
BOR009	ma mi dicono che non ce l'ho proprio forte cioè forse è difficile da collocare perché non è
BOI091	sì
BOI091	super eh ehm
BOR030	non lo so
BOI091	identificabile perché non è troppo marcato
BOI057	**eh** si capisce che non sei di bologna però non si capisce di dove
BOI091	senti mi sa che io ormai devo andare

Table 2.3: An incipient speakership in PBB019 (from the ParlaBO module).

backchannels also differ in their implications for turn-taking. While *mm hm* functions as a continuer displaying passive recipiency, *yeah* is regularly associated with incipient speakership (an example is provided in Table 2.3), projecting a possible move from recipiency to taking the floor. Their placement may also have consequences for the course of the conversation, as their use is sensitive to sequential position, especially at moments where a transition or shift becomes relevant, and the systematic nature of the distinction allows it to be exploited or subverted. For instance, *mm hm* produced at points where speakership is expected may create interactional trouble. These observations underscore, once again, that backchannels cannot be considered only as passive reflections of understanding or attention.

Backchannels may also be produced as short sentences, often made of multiple, single backchannels (Mereu et al., 2024) or can consist of a repetition of elements from the prior turn, as noted by Dideriksen et al. (2019). The example in Table 2.4 illustrates this case: the repetition of the exam date *dieci luglio* functions as a backchannel, acknowledging the information provided by the speaker while allowing the current turn to continue.

Comparable functional distinctions have been shown to hold in Italian, although realized through language-specific resources. In particular, Italian relies on the multifunctional token *sì*, whose interactional interpretation depends on its prosodic realization. Savino and Refice (2013) illustrated that *sì* can be associated with at least three distinct pragmatic functions: (1) acknowledgment with passive recipiency, thus functioning as a continuer; (2) acknowledgment with incipient speakership and (3) a simple positive answer

speaker	transcription
BO028	io io ho un problema perché dovrei subire un'operazione al ginocchio e dovevo in-
BO026	e dunque il secondo appello sarebbe
BO030	è il dieci luglio
BO026	**dieci luglio**
BO030	lo scritto
BO030	noi avremmo bisogno di sapere
BO029	esatto
BO028	orientativamente la data

Table 2.4: An example of repetition during a student reception in BOA1002 (KIP).

to a yes/no question. Based on Map Task dialogues[1], the study demonstrates that these functions are systematically distinguished by prosodic cues, in particular by the terminal fundamental frequency (F0) contour [2]. Specifically, *sì* produced as a continuer is typically characterized by a rising contour, whereas *sì* associated with incipient speakership and with yes/no answers is produced with falling contours. This pattern was confirmed in a following work, which extended the analysis to a broader set of backchannel tokens (Savino, 2014). These findings are consistent with results reported by Sbranna et al. (2022), showing that rising contours tend to mark passive recipiency, while falling contours are associated with incipient speakership. More generally, systematic differences in melodic configuration (falling, rising and level contours) have been shown to correlate with distinct dialogue functions in Italian (Cerrato and D'Imperio, 2003). All these studies on Italian therefore suggest that prosody plays a key role for distinguishing backchannels that project continuation from those that project speakership.

[1] A Map Task dialogue is an experimental dialogue paradigm in which one participant (the information giver) describes a route on a map to another participant (the information follower), who must follow the instructions to reach a specified destination. The task is designed to elicit natural, goal-oriented interaction, typically involving instructions, clarifications, confirmations, and feedback, while allowing researchers to examine turn-taking, grounding, and coordination processes in dialogue (Meena et al., 2013).

[2] The terminal fundamental frequency contour is a sequence of linguistically motivated tone switches: this includes major transitions of the F0 trajectory associated with accented syllables, as well as so-called boundary tones occurring before prosodic boundaries (Isačenko and Schädlich, 1964; Stock and Zacharias, 1973; Mixdorff and Pfitzinger, 2005).

speaker	transcription
TOR001	infatti di solito io adesso vengo la domenica
TOR001	come i vecchi
TOI048	la domenica è vuoto
TOI048	**esatto**

Table 2.5: An agreement in PTD003 (ParlaTO).

In addition to prosodic cues, lexical choice has been shown to pattern with interactional function in Italian, as part of a complex interaction between backchannel type, function and intonation. In their cross-linguistic study, Sbranna et al. (2022) reported that Italian speakers display function-specific preferences in backchannel selection: *okay*, *sì* and *mhmh* are predominantly used for passive recipiency, while *okay* is strongly favored in contexts of incipient speakership.

More recent work integrates functional distinctions with insights into multimodality. Blomsma et al. (2024) distinguished between verbal and non-verbal backchannels, including vocalizations (laughs, sighs, etc.), paraverbal items (*mm-mh*, *uh-huh*, etc.), short utterances (*really*, *yeah*, *okay*), facial expressions, nodding, eye gaze and gestures. They further differentiated between feedback functioning as go-on cues, which confirm that the listener is following and understanding the talk, and feedback conveying a communication problem. In line with Dideriksen et al. (2019), the latter are treated in this thesis as instances of conversational repair and are discussed separately.

Focusing specifically on Italian, Mereu et al. (2024) proposed a multilayer classification. Although backchannels may differ in their realization (lexical vs. non-lexical, vocal vs. gestural, single- vs. multi-unit), they are systematically associated with five core functions: continuer, incipient speakership, agreement, assessment and evoking. Importantly, Mereu et al. (2024) demonstrate that multi-unit backchannels account for a substantial portion of Italian backchannel tokens and frequently combine more than one interactional function. Finally, agreement backchannels express acceptance of the content of the prior turn, often through lexical items such as *esatto*, as illustrated in Table 2.5

Despite the existence of systematic classifications and recurrent functional patterns, Blomsma et al. (2024) showed that backchannel behavior is idiosyncratic, revealing two main variabilities: between-addressee variability

and within-addressee variability. Between-addressee variability reflects stable individual differences: some listeners consistently produce more frequent and more expressive backchannels than others. This pattern was linked to personality traits: this was observed also in other studies such as Vinciarelli et al. (2015). Within-addressee variability, instead, shows that the same listener may respond differently across BOPs, depending on their sequential relevance. In particular, *last backchannels of round* (LBRs) elicit stronger multimodal feedback than continuer BOPs, suggesting that some backchannels function not only as signals of attention or understanding, but also as markers of local completion.

In short, while backchannels are organized by interactional principles and functional distinctions, their actual deployment in conversation seems to remain sensitive to individual differences and local interactional contingencies.

2.3 Conversational Repairs

In everyday conversation, speakers regularly encounter problems in understanding. To deal with these disruptions that naturally arise in the flow of talk, they employ what CA defines as *repair* (Schegloff et al., 1977; Fele, 2007; Clark, 2020). Repair constitutes a fundamental resource through which interlocutors restore mutual understanding and maintain the progressivity of talk when something goes wrong (Fele, 2007). More generally, this phenomena reflects the social nature of language use and the joint management of meaning construction in conversational settings (Dingemanse and Enfield, 2015). As stated by (Schegloff et al., 1977, p. 361), 'an organization of repair operates in conversation, addressed to recurrent problems in speaking, hearing and understanding'.

Repair must be distinguished from correction. While correction typically only targets grammatical errors, repair is a broader phenomenon that encompasses any practice oriented to resolving trouble during an interaction, regardless of whether the problem has a linguistic, pragmatic or interactional nature (Fele, 2007).

From a sequential perspective, repair is closely tied to turn-taking organization. As Fele (2007) argues, repair initiation may occur in different positions with respect to the trouble source: within the same turn (turn 1), in the following turn (turn 2), or in later turns, depending on how and when the problem is detected. Positions are illustrated in Table 2.6.

speaker	turns
speaker A	first (1) position
speaker B	second (2) position
speaker A	third (3) position
speaker B	fourth (4) position

Table 2.6: Repair initiation positions, adapted from (Fele, 2007, p. 46).

Repair sequences consists of two analytically distinct sequential events: an initiation and an outcome. The outcome may be a modification of the trouble source (the repairable) or it may simply repeat the original item without alteration (Fele, 2007). Furthermore, attempts at repair may fail: not all repair initiations result in a successful resolution. Treating initiation and outcome separately allows researchers to account for unsuccessful repair as a specific interactional possibility (Schegloff et al., 1977).

A fundamental distinction concerns who initiates the repair. Repair may be *self-initiated*, when speakers themselves address a problem in their own talk, or *other-initiated*, when a recipient signals a problem and prompts repair (Schegloff et al., 1977; Fele, 2007; Dingemanse and Enfield, 2015; Clark, 2020). The distinction is not symmetrical: self- and other-repair are hierarchically organized within the repair system and thus are not independent or equivalent options (Schegloff et al., 1977). The organization of repair displays a strong preference for self-correction: even when recipients detect a problem, they often delay intervention to allow the original speaker to resolve it independently. When recipients do intervene, they typically initiate repair rather than providing a direct correction, thereby creating an opportunity for the speaker to self-repair (Schegloff et al., 1977; Fele, 2007).

Importantly, repair should not be automatically equated with successful communication. As argued by Galantucci et al. (2020), achieving faithful mutual understanding is a substantially demanding task that interlocutors frequently fail to perform or actively avoid. Repair initiations may correlate with lower success in communicative tasks, further underscoring the need to distinguish between repair initiation and repair outcome.

Repair practices emerge from early childhood and play a central role in the process of language acquisition. From their first year, children begin to monitor and repair their own speech, relying on memory representations of adult language. By their second year, they also respond to others' re-

pair initiations, showing sensitivity to conversational feedback (Clark, 2020). In family interaction, repair functions not only as a mechanism employed to correct conversation, but also as a powerful tool of socialization into shared systems of knowledge and cultural norms (Maroni and Arcidiacono, 2010). Other-initiated repair is particularly prominent in asymmetrical interactional contexts, such as adult-child interaction, teacher-student exchanges and native-L2 speakers conversations (Fele, 2007).

2.3.1 Other-initiated Repair

Other-initiated repair refers to practices through which recipients signal problems in understanding and prompt speakers to resolve them (Dingemanse and Enfield, 2024). Crucially, other-initiated repair is not to be confused with 'other-correction': while recipients frequently initiate repair, they rarely carry out the correction themselves (Schegloff et al., 1977). Other-initiations constitute a form of negative metacommunicative feedback, as they explicitly mark a breakdown in understanding and the need to re-establish common ground between participants. Compared to backchannels, they occur less frequently (Dideriksen et al., 2019). Nevertheless, they are a regular feature of interaction, occurring on average every 1.4 minutes across languages (Clark, 2020).

From a sequential perspective, it typically occurs in the next turn following the trouble-source turn. Listeners typically withhold repair initiation while the trouble-source turn is in progress, rarely interrupting. They often wait until slightly past the possible completion of the turn to give the speaker an opportunity to self-repair first (Schegloff et al., 1977).

Other-initiated repairs are organized as side sequences, consisting of a repair initiation followed by a repair solution. These side sequences momentarily suspend the ongoing interaction without terminating it, as it is expected to resume later (Jefferson, 1972; Dingemanse and Enfield, 2015). When side sequences function specifically as conversational repair, they temporarily shift attention away from the main course of discussion in order to deal with a problematic element. Structurally, a side sequence includes three phases: (1) the ongoing sequence, (2) the side sequence itself and (3) the return to the original activity (Jefferson, 1972). The structure of other-initiated repair mirrors this structure, consisting of a trouble source (T-1), a repair initiation (T0) and a candidate repair solution (T+1) (Schegloff,

2007).

Early work by Schegloff et al. (1977) proposed a set of five specific turn-constructional devices to initiate other-repair, ordered according to their relative strength, understood as their capacity to locate the repairable. These include, in increasing order of specificity:

- *huh?, what?*;
- wh-question words *who, where, when*;
- a partial repeat of T-1 and a question word;
- a partial repeat of the trouble-source turn;
- *you mean* accompanied by a possible understanding of T-1.

The ordering type is based on the relative strength intended as the capacity to locate a repairable. Stronger initiators are preferred over weaker ones, sometimes interrupting this latter type. Also, if more than one other-initiated sequence is needed, the other-initiators are used in order of increasing strength Schegloff et al. (1977).

While this classification was influential for the following studies, later work has argued that they do not 'fully specify all relevant features of the formats' because they are language-specific and thus not valid cross-linguistically (Dingemanse and Enfield, 2015). Repair practices are indeed shared by all languages, although they may differ in the specific forms and expressions used to initiate them. For this reason, Dingemanse and Enfield (2015) identified canonical formats based on sequential properties, that is, on how repair initiations relate to their sequential environment: how they target prior talk (T-1) and what type of response they make relevant next (T+1). Because turn-taking and sequential organization are universal properties of interaction, these dimensions allow for cross-linguistic comparison. Specifically, formats of other-initiated repair can be analyzed along two sequential dimensions: the retrospective dimension and the prospective dimension. The retrospective dimension concerns how the repair initiation targets the the trouble source in the prior turn (T-1), distinguishing between *open* formats, which indicate a problem without specifying its location or nature, and *restricted* formats, which instead point to the specific element of trouble. The prospective dimension concerns the type of response made relevant in the next turn, further distinguishing between *request* formats, which seek

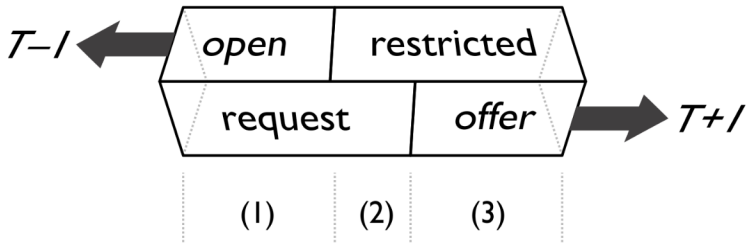

Figure 2.1: Two formats of sequential aspects (Dingemanse and Enfield, 2015, p. 105).

clarification or specification, and *offer* formats, which propose a candidate understanding to be accepted or rejected Schegloff (2007); Dingemanse and Enfield (2015); Rossi (2015). Taken together, the two dimensions build three basic formats of repair initiation: (1) open request, (2) restricted request and (3) restricted offer, as shown in shown in Figure 2.1, (Dingemanse and Enfield, 2015). Rossi (2015)'s work on other-initiated repair in Italian also adopted this classification.

Beyond its local sequential organization, other-initiated repair plays a central role in sustaining the robustness of interaction. As argued by Dingemanse and Enfield (2024), interactive repair should not be viewed as a response to communicative failure, but rather as an integral fallback mechanism that enables the flexibility and complexity of human language use. From this perspective, trouble is not treated as a defect of linguistic performance, but as a routine and expected possibility in interaction.

It is also worth mentioning that repair has two core functions. The first concerns informational robustness, ensuring that messages are correctly transmitted and understood. The second concerns social accountability, whereby participants negotiate and recalibrate the social commitments generated through talk. Repair practices may thus address not only problems of understanding, but also issues of appropriateness and normativity, as when a speaker is prompted to reformulate a request in a more socially acceptable manner. This view aligns with the conversation-analytic conception of repair as a jointly managed process grounded in participants' shared responsibility for maintaining mutual understanding. Repair becomes relevant precisely because interlocutors are publicly accountable for the intelligibility and progressivity of the interaction and its initiation marks a moment in which this

accountability is made explicit (Dingemanse and Enfield, 2024).

2.3.1.1 Open Initiators

Open type initiators offer no explicit account of the nature of the trouble or the specific item in the prior turn, leaving the repairable underspecified (Drew, 1997; Dingemanse and Enfield, 2015). As a result, generally, their use prompts the speaker of the problematic turn to repeat it in full. In some cases, the original utterance may also be modified; this behavior reflects the speaker's orientation to potential problems of appropriateness or clarity (Rossi, 2015).

While open initiators have only often been associated with problems of understanding, Drew (1997) argues that the difficulty may be related to the sequential relationship between the prior turn and the ongoing interactional context, rather than a single, specific trouble source. From this perspective, open repair initiators do not necessarily indicate mishearing or non-understanding, but may signal problems of sequential fit. Two primary environments were identified: (1) instances where a speaker introduces an abrupt, unmarked topic shift and (2) turns that are sequentially inapposite with respect to the prior course of action. It is important to underline that these two environments are not exhaustive: they do not exclude canonical cases of mishearing nor instances in which the open repair initiator is used to prompt the correction of a speaker's utterance or its pragmatic appropriateness, as frequently observed in in child-adult interaction (Drew, 1997; Clark, 2020).

Within this framework, open repair initiations can be produced through interjections (Table 2.7), question-word formats (Table 2.8) or formulaic expressions (Tables 2.9 (Dingemanse and Enfield, 2015).

In standard varieties of Italian, speakers may initiate repair using the monosyllabic interjection *eh*, composed of 'an open-mid, near-front, slightly elongated vowel [ɛ:], optionally preceded by a glottal stop', typically pronounced with a low rising intonation contour (Rossi, 2015, p. 261). An example is provided in Table 2.7: in the trouble source turn (T-1), PKP057 asks whether the ingredients had been bought there (*li aveva presi lì o no gli ingredienti?*), treating the reference to the previously mentioned shop as shared. PKP060 responds by producing an oper repair initiation, highlighted in bold (*eh?*), not specifying the specific trouble. In response, PKP057Q repeats the prior turn in full, preserving the original formulation. The repair

speaker	transcription
PKP060	devo fare una capatina
	al negozio orientale
PKP057	negozio dove?
PKP060	a reggio quello da di fronte la stazione
PKP057	ah
	ci sei già stato lì?
PKP060	indovina chi c'è andato
PKP057	luca
	classico
PKP060	anche marco
PKP057	mh
PKP057	li aveva presi lì no gli ingredienti?
PKP060	**eh?**
PKP057	li aveva presi lì no gli ingredienti?
PKP057	sì sì

Table 2.7: Open repair, interjection in KPN014 (KIPasti).

sequence is brought to completion by PKP060's response *sì sì*, which closes the repair sequence and allows the interaction to return to the main course of action.

Question-word strategies in Italian display a considerable degree of formal and functional variation. The question word corresponding to English 'what' can be translated in three forms: the compound *che cosa?*, the single-word form *cosa?* and the reduced form *che*. All three variants can function as open repair initiators when produced with a low rising intonation contour, leaving the trouble unspecified. Italian also makes use of the question word *come?* ('how') as an open repair initiator. However, unlike *cosa* and its variants, it displays a distinct interactional profile: it is typically used in contexts where the trouble is not attributed to a defect in the prior turn, but rather to the recipient's lack of attention or access. In this sense, *come?* indicates recipient-responsibility for the trouble and can be more associated to 'sorry?' in an interactional setting, rather than 'what?'(Rossi, 2015). An example of open repair initiation using a question word is shown in Table 2.8 PKP083 makes an assessment concerning the ham (*questo prosciutto è buonissimo*, which is not sequentially tied to the immediately preceding turn.

speaker	transcription
PKP082	#stammattina avimm accattat i luci
	ecco fatto
PKP083	tra l'altro questo prosciutto è buonissimo
PKP081	**che cosa?**
PKP083	il prosciutto è buonissimo
PKP082	sì?
PKP083	mh
PKP082	a bologna ci sta
	più buono

Table 2.8: Open repair, question-word in KPS015 (KIPasti).

After PKP081 responds with an open repair initiator, PKP083 repeats the prior turn, resolving the trouble. This example also aligns with observations made by Drew (1997) that open repair initiators frequently occur in environments involving topic initiation or weak sequential integration.

Other open strategies are represented by formulaic initiators. Rossi (2015) includes *prego?* (pardon?), *scusa?* (excuse me?) and *non ho capito* (I don't understand) in his analysis of Italian other-initiated repair. However, he notes that *non ho capito* is less open than other formulaic initiators: although it does not identify a specific repairable, it explicitly frames the trouble as one of understanding rather than hearing. In this respect, formulaic initiators represent a gradient within the open class, differing in the degree to which they specify the type of trouble while still refraining from locating it in the prior turn (Rossi, 2015). An example is provided in Table 2.9. In the turns preceding the repair, the participants are engaged in a coherent sequence concerning Marco's exam preparation. The sequence is abruptly interrupted when PKP041 introduces a new topic (*ma' tu te lo sei visto sherlock?*, without any explicit transition. PKP040 responds with the formulaic initiator *prego?*. PKP041 then repeats the question adding new information (*te lo sei visto sherlock? della bbc?*). The sequence is resolved with PKP040's response.

2.3.1.2 Restricted Initiators

Restricted repair initiators explicitly locate the source of trouble within the prior turn (T-1). By targeting a specific element of the trouble-source turn, restricted initiators guide the speaker toward a particular type of repair so-

speaker	transcription
PKP040	quando c'ha l'esame marco?
PKP041	il dieci
PKP126	quindi mo sta chiuso a studiare
PKP041	sta ripetendo con un tizio
PKP039	fa bene
PKP040	un tizio che sta preparando l'esame con lui?
PKP041	e certo
PKP041	ma' tu te lo sei visto sherlock?
PKP040	**prego?**
PKP041	te lo sei visto sherlock? della bbc?
PKP040	no cioè se è quello li che facevano la serie sì

Table 2.9: Open repair, formulaic in KPS008 (KIPasti).

lution (Drew, 1997; Dingemanse and Enfield, 2015).

From a sequential perspective, restricted formats display a tighter coupling between repair initiation (T0) and repair outcome (T+1). Rather than prompting a full repetition of the prior turn, they make relevant a focused response addressing the identified repairable. In this sense, restricted initiators are both retrospective, by locating the trouble, and prospective, by projecting a specific kind of response (Dingemanse and Enfield, 2015; Rossi, 2015). This constraining effect on next actions is a central feature of repair organization, as emphasized by Schegloff (2007), who shows that repair initiators systematically shape the trajectory of subsequent turns.

Following Dingemanse and Enfield (2015) and Rossi (2015), restricted repair initiators can be broadly divided into restricted requests and restricted offers, depending on the type of response they make relevant.

Restricted request formats typically involve WH-questions that specify the domain of trouble, such as reference to a person (*chi?*, who?), an object (*cosa?*, what?), a place (*dove?*, where?), or a time (*quando?*, when?). By pinpointing the problematic element, these formats invite clarification or specification rather than repetition of the entire prior turn.

Table 2.12 illustrates an example. In the trouble-source turn, PKP125 refers to a character in vague terms (*questa qua era tipo cinese però*). PKP131 initiates repair with *ma chi?*, which explicitly targets the referential component of the prior turn. The repair solution unfolds incrementally: PKP125

provides a descriptive clarification (*questa coi codini*), which is followed by PKP131's candidate identification (*musa*). This candidate is then further specified (*che ha i capelli blu?*) and ultimately confirmed by PKP125 (*eh*).

speaker	transcription
PKP125	questa qua era tipo cinese però in realtà
PKP131	**ma chi?**
PKP125	se ci pensi bene questa coi codini
PKP131	musa
PKP125	musa?
PKP131	che ha i capelli blu?
PKP125	eh

Table 2.10: Restricted request in KPS024(KIPasti).

speaker	transcription
T0999	ma secondo te quali sono le cose a cui quando si condivide un appartamento una casa in situazioni temporanee tipo appunto tra studenti erasmsu così quali sono le cose che non possono mancare cioè cosa quali elementi ci devono essere?
TO082	**dici della casa?**
TO099	sì
TO082	mh
TO082	ma allora secondo me la la pulizia resta una cosa importante io poi ho sempre vissuto cioè ho sempre avuto esperienze positive in questo senso però poimh dai racconti mi è capitato di sentire persone che dicessero
TO082	[...]

Table 2.11: Restricted offer in TOD2001 (ParlaTO).

Restricted offers differ from requests, as they propose a candidate understanding of the trouble-source turn, which the speaker can confirm or reject. As noted by Rossi (2015), while candidate hearings typically involve partial repetition of T-1, candidate interpretations often rephrase or explicate

what has been said, thereby displaying an interpretive stance toward the prior talk. Table 2.13 displays a case of restricted offer. The trouble source arises from the breadth and ambiguity of TOD099's question, which does not clearly specify whether the reference concerns interpersonal dynamics or material features of shared housing. TO028 initiates repair by proposing a candidate interpretation (*dici della casa?*), thereby narrowing the scope of the prior turn. This repair initiation both locates the source of the trouble and projects a confirmation as relevant next action. The speaker's confirmation (*sì*) resolves the repair sequence, allowing the interaction to proceed with a focused answer.

2.3.2 Self-Repair

According to Schegloff et al. (1977), self-repair is the most common and preferred form of repair in conversation: conversational organization is systematically designed to favor the speaker's own resolution of problems by allocating repair opportunities in serial order that prioritize the current speaker. This preference is not just quantitative, but structural: opportunities for self-repair are made available before recipients are entitled to intervene.

Self-initiated repair can occur in three distinct positions, all of which precede the opportunity for others to intervene and initiate repair: (1) within the same turn as the trouble source, (2) in the transition space and (3) in the turn immediately following the next speaker's response. Evidence from Italian interaction confirms this organization, showing that self-initiation opportunities systematically precede those for other-initiation (Maroni and Arcidiacono, 2010).

Initiation is often signaled through non-lexical speech perturbations, such as cut-offs, sound prolongations, and fillers like *mh*, which indicate an impending repair. Table 2.12 illustrated the basic format of same-turn self-repair: the trouble source (*la compe-*, self-initiation with a non-lexical initiator (*mh*) and the candidate repair (*quello che deve misurare e nulla di più*). This is also the most common and successful format: the vast majority of repairs are both initiated and resolved by the speaker within the same turn as the trouble source (Schegloff et al., 1977).

speaker	transcription
PSB050	ti trovi meglio a bologna o pavia?
PSB049	pavia
	eh no no scusa bologna scusa
PSB050	ah era la risposta sbagliata
PSB049	eh sì sì bologna

Table 2.13: Self-repair in the transition space in SBIB006 (StraParlaBO).

speaker	transcription
BO104	riesci ad argomentarmi questo concetto in modo
BO144	sì allora eh per essere utile un test deve essere innanzitutto
BO104	preciso
BO144	valido
BO144	cioè deve mh mh misurare
BO144	**la compe- mh quello che deve misurare e nulla di più**

Table 2.12: Self-repair within the same turn in BOC1007 (KIP).

In self-repair occurring in transition-relevance places speakers retract or revise their prior response before the interaction progresses further. An example is shown in Table 2.13, where the speaker initially provides an answer (*pavia*), then immediately corrects it (*eh no scusa bologna scusa*), preventing the erroneous response from being taken up by the interlocutor.

Finally, the example in Table 2.14 illustrates a case of self-repair in third position, triggered by a problem that becomes visible in the recipient's turn. In BOI086's turn (*e poi bianche in casa mia le facciamo bianche quindi*), the adjective *bianche* is potentially ambiguous, as it may refer either to the absence of sauce or to the type of pasta. This ambiguity is highlighted in the following turn by BOR024, who produces a candidate understanding (*senza ragù?*), immediately followed by a rejection, which creates a repair-triggering environment for the current speaker. In the following turn, BOI086 produces a self-repair, explicitly reformulating and specifying what they meant with *bianche*. The current speaker therefore resolves the ambiguity, and the recipient confirms that they have understood.

speaker	transcription
BOR024	io anche sono sono team più besciamella meno ragù perchè la trovo più
BOI086	è vero più cremosa più morbida più
BOR024	sì sì esatto esatto esatto no più buone le preferisco anche io
BOI086	e poi bianche in casa mia le facciamo bianche quindi
BOR024	senza ragù?
BOR024	no
BOI086	**no scusa la pasta non verde con gli spinaci ma bianca la pasta come quella dei tortellini insomma**
BOR024	ah ah

Table 2.14: Self-repair in the third position in PBB022 (ParlaBO).

2.4 Filled Pauses

Filled pauses are language-specific disfluencies produced at moments of hesitation, defined as temporary interruptions in speech that do not contribute to propositional content. Rather than mere breakdowns of fluency, they function as tools of speech planning, signaling ongoing cognitive activity such as searching for information, resolving lexical retrieval difficulties, or organizing the continuation of an utterance. At the interactional level, filled pauses also serve a floor-keeping function, indicating that the speaker has more to say and intends to maintain the turn (Christenfeld et al., 1991; Spreafico, 2012; Schettino and Cataldo, 2019; Cossavella and Cevasco, 2021). An example is shown in Table 2.15.

speaker	transcription
BO104	riesci ad argomentarmi questo concetto in modo
BO144	sì allora **eh** per essere utile un test deve essere innanzitutto
BO104	preciso

Table 2.15: Filled pause in BOC1007 (KIP).

Although they are extremely frequent in spontaneous speech, they have long received relatively limited attention in general linguistics. Spreafico (2012) reviews earlier approaches in which filled pauses have often been treated as paraverbal phenomena or as involuntary symptoms of speech production processes, external to the linguistic system and therefore not fully relevant for linguistic analysis. However, Spreafico (2012) also discusses that pragmatics and lexicology have increasingly argued for recognizing filled pauses as linguistic elements in their own right. When granted word status, they are commonly referred to as *fillers* and are typically classified as primary interjections, that is, lexical items not derived from other words and not employed with alternative grammatical functions. In the Italian language, they have been analyzed as expositive interjections expressing the speaker's epistemic stance, such as hesitation or uncertainty.

From a psycholinguistic perspective, they are studied in both production and comprehension. As far as production is concerned, they may represent the product of cognitive load or serve as intentional communicative strategies to manage dialogue. In comprehension, while some theories suggest listeners consider them noise, more studies indicate that they function as informative cues assisting in the incremental processing of speech. Specifically, filled pauses can signal upcoming delays or novel information, as well as influence the listener's perception of the speaker's confidence and even facilitate better memory recall of the conversation content (Dinkar et al., 2023).

Filled pauses also play a social and cognitive role. For example, non-native speakers tend to produce them more frequently due to lower automaticity in a second language, whereas individuals with Autism Spectrum Disorder (ASD) produce fewer. They can facilitate comprehension by signaling delays, highlighting new or unpredictable information and prompting listeners to anticipate upcoming content. Studies also show that filled pauses improve memory for connected speech, indicating that they function as informative cues rather than mere pauses, unlike linguistic interruptions such as coughs (Cossavella and Cevasco, 2021).

Chapter 3

Automatic Speech Recognition

3.1 Chapter Overview

This chapter provides an overview of Automatic Speech Recognition (ASR), framing it both as a computational system and as a field of interdisciplinary research. After providing definitions based on different perspectives, Section 3.2 discusses the traditional modular architecture that dominated ASR research for several decades, highlighting the role of its components. Section 3.3 subsequently traces the historical development of ASR technology, illustrating how each modeling paradigm has progressively shaped modern speech end-to-end recognition systems, which challenge the traditional modular design by integrating multiple components within a single trainable model. The focus is then switched to Whisper, a large-scale end-to-end ASR system that represents a significant shift in recent and current research. Section 3.4 provides a critical review of the literature on ASR applied to conversational data, focusing on the challenges posed by spontaneous interaction. Finally, Section 3.5 introduces automatic optimization strategies for ASR decoding and motivates the adoption of Optuna (Akiba et al., 2019) as the optimization framework for this study.

3.2 What is ASR?

Automatic Speech Recognition (ASR), also known as computer speech recognition and speech-to-text (STT), is a pattern recognition task[1] that converts a speech signal to a sequence of words (i.e., spoken words to text) by means of an algorithm implemented as a computer program. ASR technologies aim to enable spoken interactions between humans and machines by treating speech as an input modality alongside text or graphical interfaces (Karpagavalli and Chandra, 2016; Wang et al., 2019).

From a signal processing perspective, ASR involves the transformation of continuous acoustic waveforms into a discrete symbolic representation, typically a sequence of words (Karpagavalli and Chandra, 2016; Yu and Deng, 2015).

From a theoretical perspective, ASR can refer to either a technological system (as described above) or as a disciplinary field (Gregori, 2021). As a field of research, it represents an interdisciplinary area at the intersection of linguistics, computer science and electrical engineering, concerned with the development of models and methods for enabling machines to process spoken language (Rabiner and Juang, 1993; Maffi, 2016).

3.2.1 The Traditional Architecture

The canonical architecture that dominated most ASR history is composed of a set of interconnected modules, each responsible for a specific stage of processing, from the raw acoustic signal to the final word hypothesis. The architecture specifically consists of an acoustic front-end, acoustic model, lexicon, language model and decoder, as illustrated in Figure 3.1 (Yu and Deng, 2015; Karpagavalli and Chandra, 2016). This modular design was

[1]Pattern recognition refers to the computational tasks in which structured patterns are identified in data, approximating the human ability to perceive and recognize structures despite noise, variability and incomplete information. Unlike data processing, pattern recognition operates on structured regularities rather than on exact symbolic input (Yegnanarayana, 1994; Zhang et al., 2020).

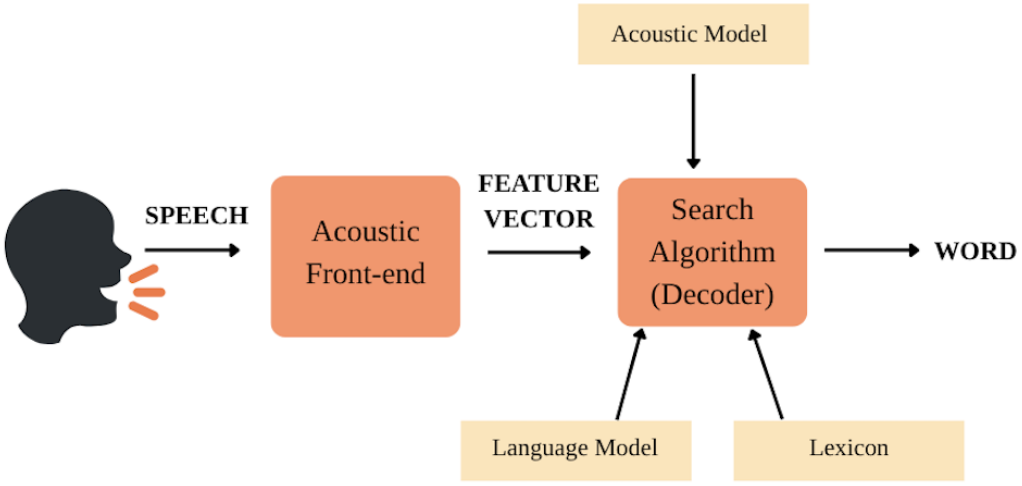

Figure 3.1: Speech recognition architecture, adapted from Karpagavalli and Chandra (2016).

developed in the context of large-vocabulary recognition systems (LVR)[2], where directly searching over all possible word sequences is computationally impractical. By separating acoustic, lexical and linguistic knowledge, the recognition problem can be factorized and efficiently approximated through probabilistic decoding, as expressed by the Bayesian formulation introduced in Equation 3.2 (Jelinek, 1976; Young, 1996; Rabiner and Jung, 2007).

The acoustic front-end is responsible for converting the continuous speech signal into a sequence of feature vectors that provide a compact and at the same time informative representation of the input. This process, commonly referred to as *feature extraction*, involves segmenting the signal into short frames and computing spectral features that capture relevant acoustic properties of speech. The goal of feature extraction is to reduce the dimensionality of the signal while preserving information that is useful for distinguishing between speech sounds across different speakers and environments. Among the various feature representations proposed in the literature, Mel-Frequency

[2]Large-vocabulary recognition (LVR) systems refer to ASR systems designed to operate over vocabularies consisting of thousands to tens of thousands of words, typically in continuous speech conditions (i.e. speech produced without artificial pauses between words). Such systems emerged in the 1970s and 1980s in response to the need for speaker-independent recognition and real-world applications (Jelinek, 1976; Young, 1996; Rabiner and Jung, 2007).

Cepstral Coefficients (MFCCs)[3] have been the most widely adopted in traditional ASR systems (Karpagavalli and Chandra, 2016). Therefore, feature extraction constitutes a critical interface between the physical speech signal and higher-level linguistic modeling (Yu and Deng, 2015).

After the feature extraction stage, the task of an ASR system is to determine the most likely word sequence corresponding to the observed signal. As introduced by Jelinek (1976), given a word sequence $W = \{w_1, \ldots, w_U\}$ and an acoustic feature sequence $X = \{x_1, \ldots, x_U\}$, decoding can be formulated as a statistical decision problem, in which the goal is to select the word sequence W that maximizes the posterior probability according to the acoustic observation:

$$\hat{W} = \arg\max_{W} P(W \mid X) \tag{3.1}$$

In practice, this posterior probability is not estimated directly. Instead, in large-vocabulary speech recognition systems, the decoding objective is reformulated by introducing an intermediate sequence of subword units $S = \{s_1, \ldots, s_T\}$, which makes it possible to separate acoustic, lexical, and linguistic knowledge. Using Bayes' rule, the objective in Equation 3.1 can be factorized as follows:

$$\hat{W} = \arg\max_{W,S} P(X \mid S)P(S \mid W)P(W) \tag{3.2}$$

This reformulation, originally adopted by Jelinek (1976) in early statistical approaches to large-vocabulary continuous speech recognition, makes explicit the three core components of a conventional modular ASR system: the acoustic model $P(X \mid S)$, the lexicon $P(S \mid W)$ and the language model $P(W)$.

The acoustic model captures the relationship between acoustic feature vectors and the basic units of speech, such as phonemes. Given a hypothesized word or phoneme sequence, it estimates the likelihood of observing a given sequence of acoustic features. Traditionally, acoustic modeling has been performed using Hidden Markov Models (HMMs), often combined with Gaussian Mixture Models (GMMs) to represent the distribution of acoustic

[3]Mel-Frequency Cepstral Coefficients (MFCCs) are features obtained by analyzing short segments of the speech signal in the frequency domain, using the Fast Fourier Transform (FTT), and encoding the spectral information in a compact cepstral representation (Turrisi, 2021).

features. Each phonetic unit is modeled as a sequence of states, allowing the system to account for temporal variation in speech (Karpagavalli and Chandra, 2016).

While the acoustic model accounts for the signal-level realization of speech, the language model encodes linguistic constraints on word sequences. Its function is to estimate the probability of a word sequence occurring in a given language, thereby guiding recognition toward syntactically and semantically plausible hypotheses. Language models are generally trained on large text corpora and are often implemented as n-gram models, such as bigrams or trigrams, which approximate the probability of a word based on a limited span of preceding words (Karpagavalli and Chandra, 2016).

Finally, the decoder integrates information from the acoustic model, the language model and the lexicon in order to search for the most probable word sequence given the observed acoustic input. This search process is commonly implemented using dynamic programming[4] algorithms which efficiently explore the space of possible hypotheses. Rather than evaluating all possible word sequences, the decoder limits the search space, balancing recognition accuracy and computational efficiency(Karpagavalli and Chandra, 2016).

This modular architecture provides the conceptual basis against which the historical evolution of ASR methodologies can be understood. The following section traces how different modeling paradigms progressively emerged within, and eventually moved beyond, this framework.

3.3 Historical Development and State-of-the-Art

The historical development of Automatic Speech Recognition (ASR) can be described in terms of successive technological and methodological generations. In this respect, Rabiner and Jung (2007) identified four established

[4]Dynamic programming is an algorithmic paradigm for solving multistage decision problems by decomposing them into a sequence of interdependent subproblems (states), each corresponding to a partial solution. The overall optimal solution is obtained by recursively combining optimal solutions to these subproblems, according to Bellman's principle of plotting optimality, which states that any optimal solution is composed of optimal decisions at each stage of the process (Bautista and Pereira, 2009; de Souza et al., 2022).

generations and briefly outlined a fifth one that was emerging at the time of writing:

- Generation 1 (1930's-1950's): ad-hoc methods designed to recognize isolated sounds or very small vocabularies.
- Generation 2 (1950's-1960's): acoustic-phonetic approaches aimed at recognizing phonemes, phones or digit vocabularies.
- Generation 3 (1960's-1980's): introduction of pattern recognition approaches relying on increasingly formalized statistical and computational techniques.
- Generation 4 (1980's-2000's): statistical modeling of continuous speech using Hidden Markov Models (HMMs).
- Generation 5, (2000's-2020's): large-scale, data-driven approaches exploiting parallel computation and increasingly complex models to improve recognition robustness.

3.3.1 Early Acoustic-Phonemic Approaches

The early history of speech recognition dates back to the period between 1930's and 1950's. At that time, research followed the acoustic-phonemic approach, focusing on the relationship between the linguistic identity of speech sounds (such as phonemes or syllables) and their acoustic realizations, with particular attention to spectral properties of the speech signal, namely, the distribution of acoustic energy across frequencies (Rabiner, 2004). Pioneering contributions were made at the AT&T Bell Laboratories by Fletcher and Homer Dudley (Fletcher, 1922; Dudley, 1939; Dudley et al., 1939), whose work demonstrated that phonemic identity is closely linked to systematic patterns in the frequency content of the signal, rather than to its overall amplitude or duration alone. As a consequence, these studies established frequency analysis as a fundamental component of speech recognition (Rabiner, 2004; Rabiner and Jung, 2007).

During the decade from 1950's to 1960's, efforts were made to formalize the acoustic-phonemic approach by developing algorithmic methods capable of recognizing speech sounds on the basis of time-varying spectral features (Rabiner and Jung, 2007). A landmark achievement in this period took

place in 1952, when Bell Labs created Audrey (Davis et al., 1952), the first complete speech recognizer. This system was able to recognize a vocabulary of ten spoken digits (i.e., the spoken word forms of the digits 'zero' to 'nine'). Although very limited in scope, this work represented the first complete implementation of a speech recognition pipeline. Despite the construction of several working systems, progress remained limited: systems could only recognize isolated words or vowels and could not handle speaker variability (Wang et al., 2019).

3.3.2 Pattern-recognition Based Approaches

Significant technological advances occurred between the 1960's and 1980's, marking a transition toward pattern-recognition based approaches. In the United States, a major milestone was the launch of the ARPA Speech Understanding Research (SUR) Project in 1971, which challenged research centers to develop demos of working systems capable of recognizing continuous speech. One of the most successful outcomes was the HARPY system (Lowerre, 1990), developed by Bruce Lowerre at Carnegie Mellon University (CMU). HARPY integrated multiple levels of linguistic models, including acoustic models, lexical representations, syntactic rules and word boundaries, into a unified finite-state network (FSN), an architecture that allows efficient search by retaining only the most promising hypotheses during decoding, rather than exhaustively exploring all possible word sequences (Rabiner and Jung, 2007).

Major contributions were also made by the IBM Research Labs and AT&T Bell Laboratories, both of which played a decisive role in shaping modern speech recognition. At IBM, research led by Fred Jelinek developed Tangora (Das and Picheny, 1996), a voice typewriter, which was capable of converting spoken language into written text. Although it was initially limited to isolated words and was speaker-dependent, Tangora introduced several innovations, including very large vocabularies, statistically defined grammars and the formulation of speech recognition as a probabilistic decoding problem. In parallel, AT&T Bell Labs focused on the development of speaker-independent systems for telephone-based applications such as voice dialing and call routing. By clustering speech data from large populations, they developed generalized acoustic models capable of handling speaker and accent variability. Together, these efforts established a strong mathematical

and statistical foundation for modern ASR (Rabiner and Jung, 2007).

Japanese laboratories also made important contributions. It is worth mentioning the phoneme recognizer, developed at Kyoto University by Sakai and Doshita (1961), who introduced 'a speech segmenter for analysis and recognition of speech in different portions of the input utterance' (Wang et al., 2019, p. 2). This work can be seen as a precursor to later continuous speech recognition systems. Furthermore, during the 1970's, technologies such as linear prediction technology[5], dynamic programming and Linear Predictive Coding (LPC)[6] were introduced, contributing to performance improvements in speaker-dependent, isolated-word and small-vocabulary taks (Wang et al., 2019).

3.3.3 Statistical ASR and Hidden Markov Models

Despite the success of pattern-recognition approaches, their reliance on fixed templates and limited temporal modeling made them increasingly inadequate for handling continuous speech, speaker variability and large vocabularies. These limitations motivated a shift from deterministic and acoustic-phonetic techniques toward fully statistical approaches, particularly models trained on labeled data and based on Hidden Markov Models (HMMs) (Karpagavalli and Chandra, 2016; Deshmukh, 2020). The key advantage of HMM-based systems lies in their ability to model speech as a time-evolving process, rather than as a static acoustic pattern (Wang et al., 2019; Deshmukh, 2020). As a matter of fact, they represent speech as a sequence of short, latent *states* evolving over time, which generate the observed acoustic signal on a probabilistic basis (Rabiner and Jung, 2007). Once trained on labeled speech data, these models can evaluate the likelihood that an unseen utterance corresponds to a given word or sequence of words (Rabiner and Jung, 2007). Their ability to jointly model temporal dynamics and acoustic variability proved crucial for unlocking large-vocabulary continuous speech recognition

[5]Linear Prediction (LP) is a method for modeling speech signals by approximating each sample as a linear combination of previous samples, effectively capturing the spectral characteristics of the vocal tract over short time intervals (Benesty et al., 2007).

[6]Linear Predictive Coding (LPC) is an application of linear predictive analysis in which the speech signal is represented as a linear combination of past samples, estimating an all-pole model that captures the spectral characteristics of the vocal tract(Schafer, 2007).

(LVCSR) in speaker-independent settings (Wang et al., 2019; Deshmukh, 2020).

By the late 1980s, HMMs had become the dominant paradigm in speech recognition and remained so for more than three decades (Baker et al., 2009; Wang et al., 2019). During the 1990s and early 2000s, DARPA programs consolidated speech recognition as a large-scale, statistically driven field. Systems such as CMU Sphinx (Lee, 1988), BBN BYBLOS (Schwartz et al., 1989), and SRI DECIPHER (Murveit et al., 1989) demonstrated that HMMs–based approaches could scale to large vocabularies and diverse speaking conditions, while also highlighting the persistent challenges posed by spontaneous conversational speech. These efforts established data quantity, probabilistic modeling and rigorous evaluation as central drivers of progress, laying the conceptual and methodological groundwork for the later shift toward deep learning–based ASR systems (Rabiner and Jung, 2007; Wang et al., 2019).

3.3.4 From GMM-HMM to Deep Learning

Despite their success, HMMs proved to exhibit several limitations that motivated the shift toward neural and deep-learning based acoustic models(Karpagavalli and Chandra, 2016; Deshmukh, 2020). While they provided an effective mechanism for modeling temporal structure, they were often combined with Gaussian Mixture Models (GMMs)[7] to represent acoustic observations. This choice required extensive manual feature engineering and carefully designed training pipelines, and was limited in representing the complex, non-linear structure of speech data (Deshmukh, 2020).

Deep Neural Networks (DNNs) were introduced as a discriminative alternative capable of learning hierarchical representations directly from acoustic features (Wang et al., 2019). Unlike GMMs, neural networks do not assume a predefined parametric form for the distribution of acoustic features and can model complex relationships through multiple layers of non-linear transformations. This representational flexibility made them particularly suitable for acoustic modeling, where variability due to speaker characteristics, speaking style and environmental conditions poses a major challenge (Maffi, 2016).

[7]Gaussian Mixture Models (GMMs) are probabilistic models that represent data as a weighted combination of multiple Gaussian distributions (i.e., several simple statistical patterns), each capturing a portion of the overall variability in the data (IBM, 2024).

More generally, artificial neural networks are computational models inspired by biological neural systems and designed to learn input-output mappings from data. Although neural networks had already been explored since the 1980's, early architectures were shallow and limited in representational capacity. Technological advances in computing power, data availability and training algorithms eventually made deeper architectures feasible, enabling their effective application to large-scale continuous speech recognition tasks (Maffi, 2016).

The beginning of the widespread adoption of deep learning methods in modern ASR systems dates back to 2011, when the research team of Yu Dong and Deng Li at Microsoft research proposed a hybrid architecture combining HMMs with context-dependent DNN, known as CD-DNN-HMM (Yao et al., 2012; Wang et al., 2019). In these systems, deep neural networks replace GMMs as the acoustic modeling component. Specifically, they employ multiple hidden layers, enabling the modeling of complex and non-linear relationships in speech data. Therefore, while preserving the modular structure of traditional ASR systems, thus maintaining the lexicon, language model and the decoder, these systems are capable of improving robustness to speaker variability, acoustic conditions and pronunciation differences (Yao et al., 2012). CD-DNN-HMM achieved substantial performance improvements over traditional GMM-HMM systems on large-vocabulary continuous speech recognition tasks, demonstrating the effectiveness of deep learning for acoustic modeling (Wang et al., 2019).

3.3.5 End-to-end Models: The Present

While deep neural networks improved acoustic modeling within the traditional modular architecture, more recent end-to-end (E2E) approaches challenge the modular paradigm by learning a direct mapping from audio to text within a single model. Engineering process is replaced by learning process and needs no domain expertise. For this reason, an end-to-end model is simpler in terms of building and training (Vitale et al., 2024b; Wang et al., 2019). As shown in Figure 3.2, most end-to-end models consist of an encoder, an aligner and a decoder.

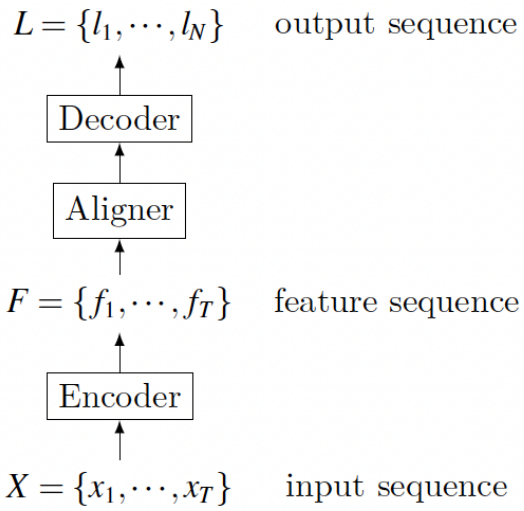

Figure 3.2: End-to-end model structure, from (Wang et al., 2019, p. 5).

The encoder transforms the input acoustic sequence into a sequence of higher-level representations; the aligner is responsible for modeling the temporal correspondence between acoustic frames and output symbols; finally, the decoder generates the final output sequence, typically characters, subword units or words (Prabhavalkar et al., 2023).

Depending on how alignment is handled, end-to-end ASR systems can be broadly categorized into three families: models based on Connectionist Temporal Classification (CTC), Attention-based models and Transducer-based architectures such as the Recurrent Neural Network Transducer (RNN-T). While these approaches differ in how they learn alignments and generate output sequences, they share the core principle of jointly optimizing acoustic, alignment and linguistic information within a single model (Wang et al., 2019; Prabhavalkar et al., 2023).

From an architectural perspective, end-to-end models depart from traditional ASR systems by eliminating explicit pronunciation lexicons and standalone language models, therefore learning these constraints implicitly from data (Rabiner and Jung, 2007; Prabhavalkar et al., 2023). End-to-end models replace this explicit structure with a single integrated network trained with a unified objective. As discussed by Prabhavalkar et al. (2023), this integration increases modeling modeling flexibility and robustness, especially in large-scale settings, where abundant training data are available. At the

same time, it reduces transparency and direct linguistic control, since pronunciation and language constraints are no longer separately represented but are embedded implicitly in the model parameters. This contrast highlights a key trade-off between modularity and interpretability, on the one hand, and integration and flexibility, on the other, which characterizes the shift from traditional to end-to-end ASR paradigms (Prabhavalkar et al., 2023).

3.3.6 Whisper

Whisper[8], introduced in 2023 by Radford et al. (2023), marks a departure point from recent approaches to ASR improvement by revealing that robust performance can be achieved through large-scale pretraining on weakly supervised data, where supervision is provided by imperfect or automatically generated transcriptions instead of manually verified labels. Trained end-to-end on a corpus of 680,000 hours of labeled audio, Whisper (Fig.3.3) addresses the poor cross-dataset generalization often observed in fine-tuned models, in which systems perform well on training data but degrade when applied to speech from different domains, recording conditions, speakers or languages, enabling robust zero-shot performance across tasks and languages without additional supervised adaptation (Radford et al., 2023).

Whisper is specifically based on an off-the-shelf encoder-decoder sequence-to-sequence Transformer[9] architecture, that is, a standard neural model that maps sequences of acoustic features directly to sequences of next tokens, trained jointly on basic transcription and on a wide range of speech processing tasks, including multilingual speech recognition, speech translation, spoken language identification and voice activity detection. These tasks are jointly formulated as *token prediction problems*, enabling a single model to replace several components of a traditional speech processing pipeline. As far as the functioning itself is concerned, the model maps raw audio input directly to text by first converting the input signal into a log-Mel spectrogram (i.e., a compact time-frequency representation of speech), which is then processed by multiple encoder blocks that extract acoustically relevant features

[8] https://openai.com/index/whisper/

[9] Transformer models are neural architectures that process a sequence by allowing the model to compare all parts of the input with each other and to determine which portions are most relevant for producing each output element. In speech recognition, this enables the model to relate distant parts of the acoustic signal without relying on step-by-step temporal processing (Vaswani et al., 2023).

while preserving temporal order of the signal. The encoded information is then passed to the decoder via cross-attention mechanism and, finally, the transcription is generated incrementally using a next-token prediction strategy, where each output token depends on the previously generated sequence (Radford et al., 2023).

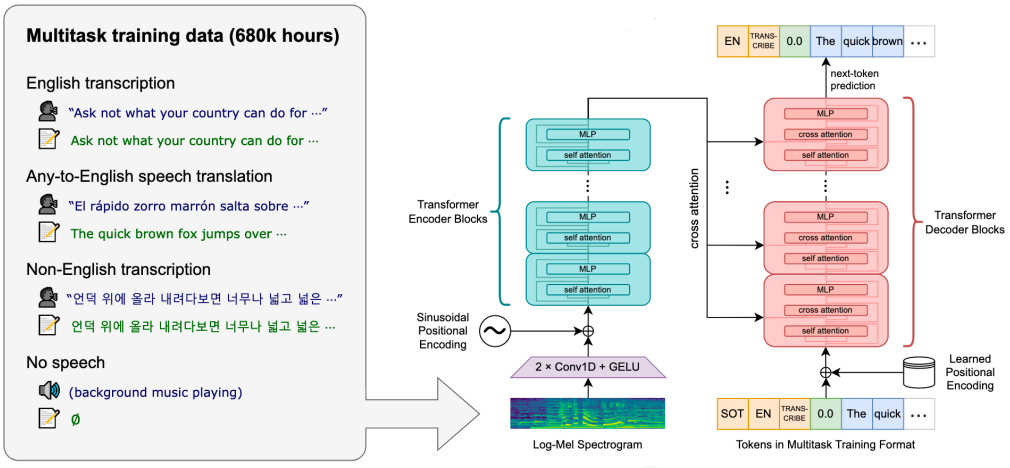

Figure 3.3: Visual representation of Whisper's sequence-to-sequence learning (Radford et al., 2023, p. 3).

Whisper can thus be seen as the culmination of the end-to-end paradigm, as it proved that architectural simplicity combined with massive weakly supervised data can outperform carefully engineered pipelines. In this sense, it represents a huge methodological shift in how robustness and generalization are pursued in contemporary ASR research (Radford et al., 2023).

3.4 ASR and Conversational Data

As ASR system have become increasingly accurate, linguists have begun to question whether they could be effectively integrated into corpus-building pipelines. Manually transcribing spoken data is an extremely time-consuming task, with transcription-time to real-time ratios reaching up to 47 for the most challenging multilogues (Gorisch and Schmidt, 2024). It is also an expensive task: generally, following a fully manual workflow involves hiring or training

professional transcribers. Given these premises, ASR-based transcription appears to be a potentially faster and lower-priced alternative, estimated to be twenty to forty times cheaper than manual transcription(Umair et al., 2022).

However, as already documented during the statistical ASR era (Rabiner and Jung, 2007), one of the main obstacles to robust speech recognition has always been the ability to cope with the variability and interactional complexity of spontaneous speech (Wang et al., 2019). When dealing with conversational data, transcription extends beyond mere speech-to-text conversion. Spontaneous interaction is characterized by overlapping turns, speaker variability, linguistic variation and complex interactional dynamics (Ghyselen et al., 2020; Liesenfeld et al., 2023). Such complexity arises from phenomena such as backchannels, interjections, interruptions and filled pauses, as well as from paralinguistic cues such as intonation, volume, and speech rate (Umair et al., 2022; Zayats et al., 2019). Conversation is typically made of brief turns and short utterances, which constitute a further obstacle for ASR systems. Evidence shows that such elements are disproportionately likely to be omitted from ASR output: Cumbal et al. (2021) illustrate frequent recognition failures for utterances containing four words or fewer, often resulting in empty outputs, while Lopez et al. (2022) found that monosyllabic function words and minimal responses are systematically underrepresented or missed in ASR transcripts. All these elements contribute significantly to transcription challenges, also for human transcribers. Manual transcriptions are indeed not infallible: human error rates range from 3% to 10%, depending on the nature of the input speech and the amount of time dedicated to the transcription task (Ghyselen et al., 2020). Studies show that disagreements among transcribers mainly involve function words and backchannels, while content words are less prone to error (Zayats et al., 2019).

ASR systems, however, largely ignore interactional features. To my knowledge, GailBot is the only known tool explicitly designed to address this gap, combining ASR with modular plugins for detecting non-lexical and paralinguistic elements such as overlaps, silencer, laughter, speech rate variation and other turn-taking related phenomena (Umair et al., 2022). Despite its extensibility, GailBot's performance is constrained by plugin accuracy and design, and it has never been adapted for Italian. Applying it to KIParla would require developing plugins tailored to both language and corpus-specific needs.

More broadly, off-the-shelf ASR systems perform well in controlled and pre-arranged settings but underperform in real-world, spontaneous conversa-

tional contexts (Gaur et al., 2016; Cumbal et al., 2021). Trained primarily on standard, fluent and monologic speech, these systems treat conversational phenomena as noise to be removed (Lin et al., 2025; Dinkar et al., 2023). In classical ASR pipelines, the dominant objective has always been to produce fluent, text-like output, leading to the widespread adoption of, for instance, disfluency detection and removal as a post-processing step. As a result, ASR models have largely focused on recognizing and removing interactional phenomena, treating them similarly to stop words and considering conversational speech as unprofessional rather than as a structured signal in its own right (Yang et al., 2003; Dinkar et al., 2023; Wang et al., 2025b).

Whisper (Radford et al., 2023), for example, illustrates both the strengths and limitations of modern ASR systems. Despite its robustness and strong generalization, applying Whisper to conversational data has been shown to pose several challenges. Work by Yamasaki et al. (2023) reports a tendency to hallucinate words, omit disfluencies and discourse markers, and transcribe background voices when speakers share a physical space. These behaviors inevitably affect the output quality and remove features that are crucial for analyzing conversational dynamics. For linguistic research, such shortcomings underscore the importance of achieving highly accurate transcriptions (Ghyselen et al., 2020; Gaur et al., 2016).

Transcription accuracy is commonly evaluated using the Word Error Rate (WER) (Klakow and Peters, 2002), where a lower WER indicates higher transcription quality. Gaur et al. (2016) suggest that a WER above 30% marks the point where manual transcription becomes more efficient than ASR assisted transcription. However, despite being the standard metric for transcription accuracy, it has been criticized as a means to evaluate conversational data: Liesenfeld et al. (2023) argue that WER oversimplifies performance and ignores key discourse features such as disfluencies or overlaps. Similarly, Gorisch and Schmidt (2024) show that WER tends to disproportionately penalize omissions, especially of hesitation markers and backchannels. These omissions concern elements that are critical for CA, yet they are often absent in ASR output. As a consequence, the literature does not support ASR-assisted transcription workflows, citing no clear gains in either efficiency or quality.

Crucial work by Vitale et al. (2024b,a) demonstrated that end-to-end ASR systems do encode interactional and hesitation-related phenomena within their internal representations. However, the decoding stage plays the critical role in shaping how and if such information will be represented in the

output. Also, different decoding strategies produce and bias such phenomena in distinct ways, despite sharing the same encoder architecture. Vitale et al. (2024b) therefore provides a key view within this framework: interactional phenomena may be encoded but selectively normalized at decoding time. Related findings from language modeling research further suggest that the absence of backchannels and fillers in model output does not necessarily imply a lack of internal representation, but may instead reflect training objectives and data biases (Wang et al., 2025b). This is consistent with studies such as de Zuazo et al. (2025), who show that Whisper's output is highly sensitive to decoding-time configuration choices. The systematic optimization of decoding parameters appears to be a possible path that may lead to substantial variations in transcription behavior without modifying model weights.

3.5 Automatic Optimization of Decoding Parameters

According to Romero et al. (2024), ASR performance is often driven by a small subset of highly influential parameters, whose impact only emerges when systematic analysis techniques are applied. As a result, a proper configuration of hyperparameters is essential for achieving performance close to full fine-tuning (Wang et al., 2025a). Beyond model training, however, system performance also depends critically on decoder configuration, which must balance transcription accuracy with computational efficiency in order to produce accurate transcriptions with the fastest possible speed Chandrashekaran and Lane (2016). All decoders have their own set of hyper-parameters, which govern the accuracy of the transcriptions, the time taken to transcribe the utterance, as well as the computational footprint. Manual tuning can be performed, however when multiple objectives and multiple hyper-parameters are present, human intuition may not be sufficient to obtain the best selection. Also on a practical level, the human expert may not perform an exhaustive search over the search space[10] to find the most optimal hyper-parameter settings for the given task (Chandrashekaran and Lane, 2016).

[10]The search space defines which hyperparameters can be tuned and the range of values over which the optimization algorithm is allowed to search (Akiba et al., 2019).

The automatic optimization of ASR decoding parameters has been investigated well before the advent of neural speech recognition. El Hannani and Hain (2010), for instance, framed decoder tuning as a multi-objective optimization problem, showing that manual parameter selection is insufficient due to strong interdependencies among decoding parameters. By describing optimization as a trade-off between Word Error Rate (WER) and Real-Time-Factor (RTF)[11], respectively representing quality and computational cost, they demonstrated that automated optimization strategies can identify better-performing configurations with modest computational cost.

Six years later, Chandrashekaran and Lane (2016) also addressed decoder parameter tuning as a multi-objective optimization problem, this time in the context of large-vocabulary continuous speech recognition (LVCSR). Their study compared traditional manual optimization with a range of automated search strategies: by systematically exploring the decoder parameter space, these methods are able to account for complex interactions among parameters that are difficult to capture through manual tuning. They found that automated optimization techniques consistently outperformed both manual optimization and random sampling[12], yielding more stable and reproducible configurations and achieving lower WER, while respecting real-time constraints. This advantage derives from the ability of automated methods to explore a broader portion of the search space and to avoid biases introduced by heuristic parameter choices (Chandrashekaran and Lane, 2016).

More recently, automated hyperparameter optimization has been applied to fine-tune large neural ASR models, including Whisper. Both Almahmood et al. (2024) and Wang et al. (2025a) combined Whisper with Optuna (see Section 3.5.1) to optimize hyperparameters in under-represented domains, reporting substantial reductions in WER. Specifically, Almahmood et al. (2024) worked on the Bahraini Arabic dialect: their approach therefore combined dialect-specific training data with systematic hyperparameter search, allowing key training parameters to be selected automatically. Their findings illustrated high improvements in WER, demonstrating that opti-

[11]It is defined as 'the time of an utterance divided by the time it takes the system to process it. A system is considered real-time ready if the RTF is less than one, which means that the system can transcribe a speech faster than it is being spoken' (Arriaga et al., 2025, p. 2).

[12]In random sampling, each hyperparameter is randomly chosen by the user within predefined ranges. Hyperparameters are then evaluated by measuring the performance of the resulting model (Chandrashekaran and Lane, 2016).

mization supports effective generalization on new, unseen data. Wang et al. (2025a), instead, combined Low-Rank Adaptation (LoRA)[13](Hu et al., 2021) with Optuna in a parameter-efficient fine-tuning framework. A dual strategy based on pruning and early stopping[14] enabled effective control over validation loss and WER improvement rates, resulting in enhanced model performance alongside reduced training time.

Related evidence on Optuna's effectiveness on Whisper is provided by Deußer et al. (2024), who employed the ASR model as a part of a multimodal pipeline for automatic dementia detection based on spontaneous speech. Their findings showed that performance for this type of task varies depending on how the transcription and modeling pipeline is configured, with Optuna-based hyperparameter optimization playing a key role in selecting configurations that lead to measurable improvements in clinical classification accuracy.

All studies aforementioned evidence that Whisper is highly sensitive to parameter choices and that systematic, automated optimization strategies are preferable to heuristic training approaches (Wang et al., 2025a).

3.5.1 Optuna

Optuna is an hyperparameter optimization software, designed to address some of the main limitations of earlier optimization tools, particularly in terms of flexibility, efficiency and ease of use. It is introduced as an innovative, next-generation optimization framework built around three core design criteria: a *define-by-run* programming model, efficient sampling and pruning strategies and, finally, a versatile architecture (Akiba et al., 2019).Hyperparameter optimization in Optuna is defined as the process of minimizing

[13]Low-Rank Adaptation (LoRA)is a parameter-efficient fine-tuning technique in which the original weights of a pre-trained model are kept fixed, and only a small number of additional trainable parameters are introduced to adapt the model to a new task. This approach significantly reduces training cost while retaining most of the knowledge of the original model (Hu et al., 2021).

[14]While pruning removes unpromising parameter configurations, early stopping terminates training runs based on the validation metrics Wang et al. (2025a).

or maximizing an objective function[15], which takes a set of hyperparameters as input and returns a scalar evaluation score. Each optimization process is referred to as a *study*, while each evolution of objective function is a *trial*. A study therefore consists of an iterative sequence of trials, in which Optuna repeatedly proposes new hyperparameter configurations, evaluates them through the objective function and records their outcomes. At each iteration, the information collected from previous trials is used to guide subsequent sampling decisions, progressively guiding the search toward more promising regions of the hyperparameter space (Akiba et al., 2019).

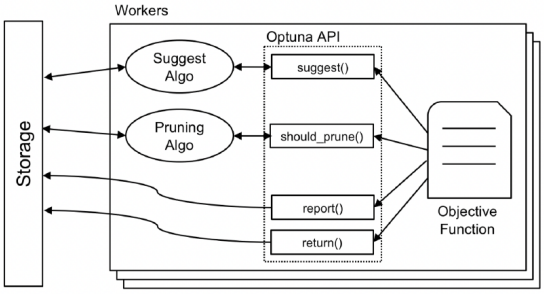

Figure 3.4: Optuna's system design, from (Akiba et al., 2019, p. 5).

Its *define-by-run* programming allows for the search space to be constructed dynamically by the user during the execution of the objective function. Also, sampling and pruning algorithms are respectively customizable and automated: while the first are responsible for proposing new hyperparameter configurations based on the outcomes of previous trials, pruning algorithms perform automated early stopping by terminating trials that are unlikely to produce competitive results. Among the sampling strategies supported by Optuna, a widely used approach is the Tree-structured Parzen Estimator (TPE) (Bergstra et al., 2011). As far as pruning mechanisms are concerned, they allow Optuna to terminate unpromising trials before their

[15]In optimization theory, an objective function quantitatively evaluates each possible solution by mapping it to a numerical value, which represents the criterion according to which solutions are compared. An optimization problem can be formulated either as a minimization problem, which consists in finding the solution that yields the smallest value of the objective function, or as a maximization problem, which aims at identifying the solution that produces the largest value of the objective function (Gunantara, 2018).

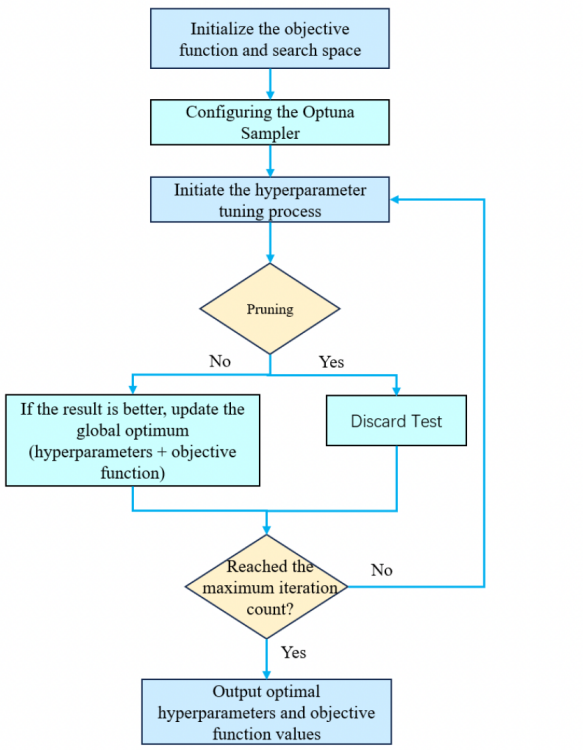

Figure 3.5: Optuna flowchart diagram, from (Wang et al., 2025a, p. 6).

completion, based on intermediate evaluations of the objective function observed during training or evaluation, as illustrated in Figure 3.4. This design is particularly effective in machine learning scenarios, where full evaluations may require substantial computational resources. By discarding poorly performing configurations early, Optuna allows for substantial computational savings (Akiba et al., 2019). The sequential workflow of the optimization process is shown in Figure 3.5, starting from hyperparameter sampling to evaluation, pruning decisions and termination criteria.

Finally, Optuna is designed to be versatile and easy to deploy, supporting both minimal and complex tasks with minimal setup. This combination of flexibility and efficiency makes it particularly suitable for scenarios in which hyperparameters must be systematically tuned or retrained, but manual configuration would be impractical (Akiba et al., 2019).

Chapter 4

The KIParla Corpus

4.1 Corpus Description: an Overview

KIParla[1] is a growing corpus of spoken Italian and a product of a collaborative effort between the Universities of Bologna and Turin. At the moment of writing, the corpus consists of around 223 hours of recording and approximately 2M transcribed tokens. Four modules are currently available for consultation, namely KIP (Goria and Mauri, 2018; Mauri et al., 2019b), ParlaTO (Cerruti and Ballarè, 2020a,b), KIPasti (Mauri et al., 2024b), and ParlaBO (Mauri et al., 2024a). Two modules, instead, are accessible in their DEMO format: Stra-ParlaBO and Stra-ParlaTO. The corpus is freely available through a custom NoSketchEngine[2] service that provides transcriptions aligned with audio files, as well as full transcripts both in orthographical (i.e., following Italian standard orthography) and conversational format. The corpus is also available on Github[3], where it is possible to consult files both in ortographical and Jefferson format, as well as metadata, `.eaf` and `.tsv` format files.

The corpus was created to address the shortcomings of previously existing resources of spoken Italian. The innovations introduced by the KIParla represent three key aspects within the context of corpus-based research: (1) accessing to speaker metadata, particularly concerning age and social group; (2) the possibility of consulting the corpus online; (3) transcriptions aligned

[1]https://kiparla.it/
[2]https://search.corpuskiparla.it/corpus/crystal/#open
[3]https://github.com/KIParla

with audio files. Moreover, the KIParla is structured in an incremental and modular fashion that allows the addition of new sections over time. Such incrementality and modularity also represent a novelty within the scene of spoken data resources and respond to the imperative of capturing the linguistic diversity inherent in spoken language, both in terms of conversation type and speakers. To guarantee that the KIParla is an actual portrait of spoken language, the conversations are always recorded in ecological settings, can involve up to five speakers and often capture non-standard linguistic features. Conversation types in the resource are also quite variable, including near-monologues (recorded from university lectures), semi-structured interactions (e.g., oral exams, office hours and interviews) as well as completely free conversations. This variability significantly affects the number of conversational overlaps between speakers, which inevitably reduces the accuracy of off-the-shelf ASR models, as they typically struggle with speaker diarization. The KIParla already encompasses a diverse range of Italian spoken varieties, featuring participants of various ages, genders, origin, professions and education levels. However, a great deal of work is yet to be done to fully represent spoken Italian in all its nuances, therefore more data collection and transcription is surely envisaged (Ballarè and Mauri, 2021; Mauri et al., 2019a).

4.2 Corpus Composition

The KIParla corpus currently consists of four main modules (KIP, KIPasti, ParlaTO and ParlaBO), which together form the overall resource. The modules can be consulted independently or through a joint query mode that provides access to the entire corpus as a unified dataset. In addition to the full modules, two demo versions are available: Stra-ParlaBO and Stra-ParlaTO, that correspond to preview versions of the forthcoming Stra-Parla modules. When demo data are included, the total size of the resource reaches 2,540,247 tokens. The overall composition of the corpus, including both full modules and demo versions, is summarized in Table 4.1.

Participants range from one to six per recording. Data was collected in 22 Italian provinces, ensuring geographical diversity across the entire corpus. KIParla includes a variety of interactional settings, such as free conversations, dinner table conversations, exams, office hours, lectures and semi-structured interviews, thereby capturing both symmetrical and asymmetrical interac-

Corpus	Type	Tokens
KIParla	*Overall corpus*	**2,069,384**
Modules		
KIP	Module	581,834
KIPasti	Module	404,896
ParlaBO	Module	606,634
ParlaTO	Module	476,047
Demo versions		
Stra-ParlaBO	Demo	378,913
Stra-ParlaTO	Demo	169,431
Total (including demos)		**2,540,247**

Table 4.1: Composition of the KIParla corpus, including full modules and available demo versions.

tional configurations.

Metadata are available at both the conversation and speaker levels. Instead, conversation-level metadata include interaction type, number of participants, participant relationship (symmetric vs. asymmetric), presence or absence of a moderator, year of recording (2017-2024) and data collection place. Speaker-level metadata provide additional sociolinguistic information, such as gender, age range, region of origin, educational background and occupation. Age categories range from 16 to over 85 years, while educational levels span from primary school to PhD. Occupational categories reflect a broad spectrum of professional backgrounds, including students, intellectual professions, technical occupations, manual workers and retirees [4].

4.2.0.1 KIP

The KIP module (Goria and Mauri, 2018; Mauri et al., 2019b) represents the first core component of the KIParla project. Its construction began in 2016 and was published in May 2019. The module comprises approximately 70 hours of recorded interaction, corresponding to 661,175 tokens.

[4] https://kiparla.it/consultazione-congiunta/

Recordings were collected at the Universities of Turin and Bologna and include five communicative settings: university lectures (25:45:12), oral exams (6:20:22), office-hour consultations (6:48:19), semi-structured interviews with students (14:06:15), and free conversations (16:23:33), for a total duration of 69:23:08. The dataset is characterized by a relatively homogeneous sociolinguistic profile and participants are predominantly university students and academic staff[5].

KIP therefore represents an interactionally diverse subcorpus, offering access to both institutional and semi-spontaneous speech contexts within an academically homogeneous population.

4.2.0.2 ParlaTO

The ParlaTO (Cerruti and Ballarè, 2020a,b), module constitutes the second component of the KIParla corpus and includes approximately 50 hours of spoken interaction. It consists of semi-structured interviews collected in Turin between 2018 and 2020, involving over one hundred speakers with diverse geographical origins and socio-economic backgrounds. Interview topics cover a wide range of everyday domains, including education, employment, leisure activities, retirement, personal memories, and urban life.

The recordings are distributed almost evenly across three age groups: young speakers (16–29 years; 16:56:47), adults (30–59 years; 15:39:01), and elderly speakers (60+; 16:15:26), for a total of 48:51:14 hours. Approximately 7 hours and 45 minutes of the recordings involving younger speakers overlap with the KIP module.

Unlike KIP, ParlaTO is characterized by substantial sociolinguistic heterogeneity. Speakers vary in terms of occupation, educational background, age, gender, and region of origin, thus offering a broader representation of contemporary Italian diastratic and diatopic variation. The dataset also includes dialectal insertions: in the transcription, the presence of dialect within a transcription unit is marked by a # symbol placed before the first word of the unit containing at least one dialectal item [6].

[5] https://kiparla.it/kip/
[6] https://kiparla.it/parlato/

4.2.0.3 KIPasti

The KIPasti module (Mauri et al., 2024b) represents the third component of the KIParla corpus and consists of 63 table conversations recorded across 13 Italian regions. The dataset includes over 40 hours of interaction (42:49:19) and 482,892 tokens, involving a total of 145 speakers with diverse geographical origins and socio-social profiles.

KIPasti was designed to approximate the demographic distribution of the Italian population across the macro-areas of the North, Centre, and South (with a maximum deviation of 4% from ISTAT population data). Recording time is distributed as follows: North (19:25:31), Centre (6:50:19), and South and Islands (16:33:29). This structure ensures a geographically balanced representation of informal spoken Italian across the country.

The uniqueness of KIPasti is represented by the presence of highly spontaneous interaction in domestic settings. Dialectal insertions are particularly frequent, occurring in 78% of the conversations [7].

4.2.0.4 ParlaBO

The ParlaBO (Mauri et al., 2024a) module represents the fourth component of the KIParla corpus and comprises over 65 hours of spoken interaction (65:43:25). The dataset includes semi-structured interviews collected in Bologna between 2021 and 2024 and involves more than 150 speakers with diverse geographical origins and socio-economic backgrounds. Interview topics span a broad range of domains, including education, employment, leisure activities, retirement, personal memories, urban life, local traditions, and regional practices. As ParlaTO and KIPasti, this module also contains dialectal varieties.

Recording time is distributed relatively evenly across age groups: young speakers (16–29 years; 18:46:50), adults (30–59 years; 24:21:43), and elderly speakers (60+; 22:34:52). As in ParlaTO, the design ensures balanced age representation while maintaining sociolinguistic diversity.

[7] https://kiparla.it/kipasti/

4.2.0.5 Stra-Parla

As already mentioned, two modules are accessible in their DEMO format (Table 4.1): Stra-ParlaBO and Stra-ParlaTO[8]. In their complete version, they will contain approximately 50 hours each of oral data from semi-structured interviews and free conversation involving speakers with a history of international migration (SIMB) residing in the urban areas of Bologna and Turin. Participants belong to four different linguistic communities and vary with respect to country of origin, first language, age, educational background, length of residence in Italy, and type of occupation.

4.3 Data Collection and Transcription Methodology

Data collection is carried out by researchers, with the collaboration of students and interns of the Universities of Bologna and Turin. The process is preceded by a period of training related to data collection. Conversations are, if possible, recorded via the recorder Zoom H4n Pro or, as an alternative, with a smartphone. Before every recording, participants agreed and signed a consent form that complies with the General Data Protection Regulation (G.D.P.R.) of the European Union. The collected data is transcribed and anonymized before publication. All conversations in KIParla are transcribed manually, to guarantee a high level of accuracy and allow special cases to be treated individually. The software adopted for the transcription is ELAN (Max Planck Institute for Psycholinguistics, 2025), that allows for precise time-alignment. The conversations are segmented into transcription units (TUs) based on prosodic and/or semantic criteria. Content is transcribed in accordance with Italian orthographic norms and a set of conventions to represent prosodic and interactional features (e.g., overlaps, pauses, hesitations...) inspired by Jefferson notation (Jefferson, 2004). The conventions adopted also include symbols to mark overlapping speech, non-verbal behavior e.g., ((*applaude*)), 'claps' ((*ride*)), 'laughs', etc., as shown in Table 4.2.

To ensure that the corpus remains fully queryable with concordance tools while maintaining maximal faithfulness to the data, transcribing the conversations of the KIParla requires many punctual and specific decisions, hence

[8] https://kiparla.it/moduli-in-costruzione/

symbol	meaning
,	weakly ascending intonation
?	ascending intonation
.	falling intonation
:	prolonged sound
(.)	short pause
=	prosodically-linked units
#	code-switching/code-mixing to/with dialect
$	code-switching/code-mixing to/with foreign language
>ciao<	pronunciation (faster)
<ciao>	pronunciation (slower)
CIAO	volume (higher)
°ciao°	volume (lower)
cia-	interrupted speech
[ciao]	overlap between speakers
(ciao)	transcriber's best guess
(ciao)	unintelligible sequence
((ride))	nonverbal behavior

Table 4.2: Symbols for transcription.

the manual approach. In order to get from source audio to fully queryable transcript, a team of linguists with various areas of expertise are involved, including properly trained and regularly monitored student interns. It is estimated that, up to now, around 80 interns have taken part in the transcription process. The workflow proceeds as follows: transcriptions are initially produced by interns and then undergo a full revision by an expert linguist. This two-step process helps mitigate inconsistencies introduced by individual transcribers, who may tend to over- or underrepresent certain phenomena, and ensure internal consistency across the corpus (Ballarè and Mauri, 2021; Mauri et al., 2019a).

However, it is important to underline that such revision inevitably carries the subjectivity of the expert linguist, especially regarding the Jefferson notation. In linguistic research, data are not direct reflections of physical phenomena, but rather representation of linguistic events. Consequently, transcriptions are not an objective reflection of speech, but rather abstract and interpretative acts shaped by methodological choices and theoretical frameworks. In this sense, it is impossible to escape subjectivity, which is not con-

sidered noise to eliminate, but an integral part of the data itself (Lehmann, 2004).

Chapter 5

Methodology

5.1 Chapter Overview

This chapter describes the methodological framework adopted in the study, from dataset selection and manual annotation to ASR transcription, decoding optimization and evaluation procedures. The goal of this chapter is to provide a transparent and reproducible account of the experimental design, ensuring that both linguistic and computational components of the workflow are clearly motivated and documented. The chapter is organized into five main sections. Section 5.3 introduces the dataset. It describes the criteria used for selecting the conversations, the subdivision of the corpus into annotated and exploratory subsets, and the rationale behind the creation of Subset A and Subset B for optimization and control purposes. Section 5.4 presents the manual annotation procedure. It introduces the annotation environment (INCEpTION), the data preparation workflow, and the criteria adopted for identifying interactional phenomena. This section also explains how metadata were integrated into the annotation files and how consistency across annotations was ensured through explicit operational definitions. Section 5.5 describes the ASR transcription pipeline. After introducing the transcription tool and its diarization and segmentation components, the following Section (5.6) illustrates the four initial decoding configurations defined for the study. It then details the automatic optimization procedure implemented with Optuna, including the definition of the search space, the evaluation workflow and the distinction between WER-based and Interaction-aware objective functions. Section 5.7 focuses on the processing of gold standard

annotated transcripts and on the evaluation methodology. It explains the conversion of annotated files into a vertical format, the extraction of descriptive statistics, the word-level alignment procedure and the normalization steps applied before computing evaluation metrics. The definition of the mean match ratio, the substitution rates and overlap analysis calculated and implemented are also described. Finally, Section 5.8 describes the qualitative analysis, which was implemented to support and illustrate quantitative findings. All these methodological steps establish the empirical basis for the results presented in the following chapter.

5.2 Research Questions and Analytical Framework

The study is guided by the following research questions:

- **RQ1**: How does decoding optimization affect overall transcription accuracy in spontaneous conversational speech, as measured by standard Word Error Rate (WER)?
- **RQ2**: How does decoding optimization affect the recognition and preservation of interactional phenomena (backchannels, filled pauses, self-repairs and other-initiated repairs), as measured through event-level match ratios and error distributions?
- **RQ3**: Are certain interactional phenomena structurally more vulnerable to omission or normalization, regardless of optimization strategy?
- **RQ4**: Does conversational overlap significantly affect the recognition of interactional phenomena, and is this effect consistent across decoding strategies?

5.3 Dataset

All the conversations selected for this study derive from the KIParla corpus, described in Chapter 4. For the purpose of the present analysis, the main selection criterion was the presence of two speakers. Audio quality

subset	total duration (hh:mm:ss)
full dataset	15:28:35
annotated subset	8:40:00
exploratory subset	6:48:35

Table 5.1: Overview of the dataset duration and its subdivision into annotated and exploratory subsets.

was not considered as a strict exclusion criterion: while most recordings are of relatively high quality, some are clearly recorded using mobile devices or in public environments where background noise may be present. This choice was made in order to preserve a realistic representation of the corpus and the conditions under which conversational speech naturally occurs. Twenty-six conversations were therefore selected for the experiment: twelve semi-structured interviews from both ParlaBO and ParlaTO, eight student-professor meetings from KIP (six of which are oral exams and two are office hours) and six free conversations both from KIP and KIPasti. The Table containing all 26 conversation IDs is provided in Appendix A.1.

The first 20 minutes of each conversation were selected for annotation, representing the gold standard dataset. Table 5.1 provides an overview of the dataset in terms of its time division. It comprises a total of 15 hours and 28 minutes of audio, of which 8 hours and 40 minutes were manually annotated and used as reference. The remaining portions of the recordings, starting from the twentieth minute onward, were reserved for optimization and exploratory analyses. As shown in Table 5.1, 6 hours and 48 minutes of audio were allocated to this phase.

The exploratory subset was further divided into two subsets with comparable overall duration: Subset A and Subset B. While the first was used for the automatic optimization of ASR decoding parameters, the second was reserved for control analysis. In particular, Subset B was not involved at any stage in the parameter optimization process and was used to verify whether the observed behavior of Subset A was not dataset-specific, therefore exploring how it generalized to unseen data. This choice allows performance to be assessed both on the data used for parameter selections and on unseen material, reducing the risk of overfitting the optimization procedure to a specific subset. Subset A amounts to 3 hours and 25 minutes, while Subset B comprises 3 hours and 22 minutes, resulting in a near-balanced temporal distribution. The subdivision was carried out with the primary goal of

equalizing total duration rather than enforcing a strict balance across interaction types. Sure enough, as summarized in Table 5.2, in this respect the composition of the two subsets is not perfectly symmetrical. This is due to the differences in the original varying length and structure of the conversations. Subset A contains a larger proportion of free conversation, whereas Subset B includes more student-professor meetings. These differences are a consequence of the corpus structure itself and were deemed acceptable, as both subsets remain representative of the interactional diversity of the whole dataset and, more broadly, of the KIParla.

interaction type	**subset A (hh:mm:ss)**	**subset B (hh:mm:ss)**
total duration	3:25:38	3:22:31
semi-structured interviews	2:10:03	1:58:53
oral exams and office hours	0:26:41	0:59:09
free conversations	0:48:54	0:24:29

Table 5.2: Temporal distribution of interaction types across the two exploratory subsets.

As far as the speakers are concerned, the set appears heterogeneous in terms of age range, occupation, gender, birth region and educational degree, in line with the general composition of the KIParla corpus. This variability also extends to linguistic profiles: one semi-structured interview (PBB010) involves an L2 speaker of Italian, while dialectal features are present in five conversations, precisely covering four regions: Campania, Emilia-Romagna, Piedmont and Sicily.

For clarity, the dataset partitioning is summarized in Figure 5.3.

5.4 Annotation Pipeline

The annotation was carried out using the open-source web application and text-annotation environment INCEpTION (Klie et al., 2018), chosen for its extendability and adaptability to individual research requirements. The platform offers a wide range of functionalities, ranging from simple span annotation to knowledge management, active learning and entity linking.

Annotation is organized into layers, which define the types of annotations that can be created and the respective features. Layers can be defined either

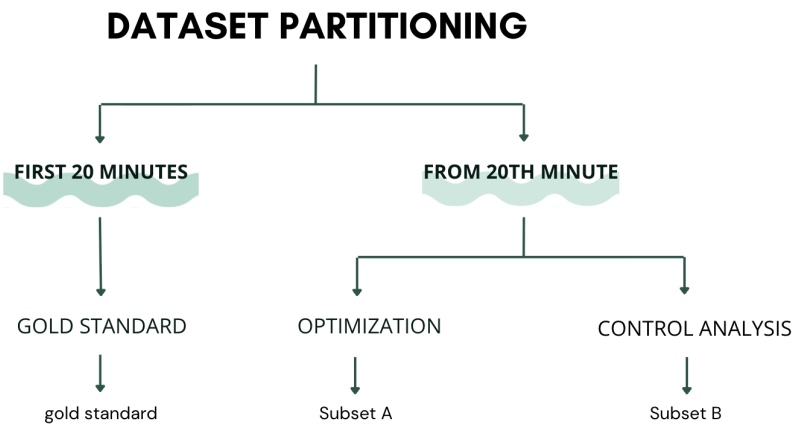

Figure 5.1: Visual summary of the dataset organization.

as spans, relation between spans and chains. A span layer allows to annotate specific segment of text, from one character to another. For example, a filled pause such as *ehm* can be annotated as a span, and a multi-unit backchannel such as *okay okay ho capito* can also be annotated as a span. A relation layer is used to link two span annotations. In the case of repair, for instance, a relation can connect a repair initiation to the corresponding trouble source. Finally, chain layers are used to group together multiple spans that refer to the same object or, as in this case, the same interactional event. In this study, overlapping speech is annotated through a chain layer: therefore, all overlapping segments produced by different speakers within the same overlap episode are linked together as part of a single chain. Each layer, then, is associated with one or multiple features, that specify the properties of the annotated element. A feature may encode, for example, backchannel functional types. Features can take different forms, such as predefined string values, numerical values, booleans, concept references, or references to other annotations.

For this study, the annotation scheme was kept simple. A total of eight layers was defined, of which seven spans (TokenID, Speaker, Jefferson, Intonation, Backchannel, Repair and Filled Pause) and one chain (Overlaps). Only three of them were actively used for annotation and correspond to Backchannel, Repair and Filled Pause categories. The remaining five layers were introduced to preserve useful and relevant metadata in the output,

and also as a reference to provide contextual information during annotation. These metadata layers therefore include TokenID, Speaker, Overlaps, Jefferson and Intonation.

The data preparation workflow consisted of three main steps. First, plain-text transcripts were uploaded to INCEpTION in `.txt` format in order to define the project structure and manually create metadata layers. Second, the transcripts were exported from INCEpTION in `WebAnno TSV (v3.3)` format[1], which is a sentence-oriented format that was created specifically for the application, and is also the most suitable format for the purpose of this annotation. Third, metadata information was added to the WebAnno files using a custom Python script, detailed in Subsection 5.4.1. The resulting enriched files were then re-imported into INCEpTION, where annotation layers were created and manual annotation was carried out.

Figure 5.2 shows the interface with the imported metadata layers only. During annotation, the user can decide to make visible only some of the layers, according to their specific needs.

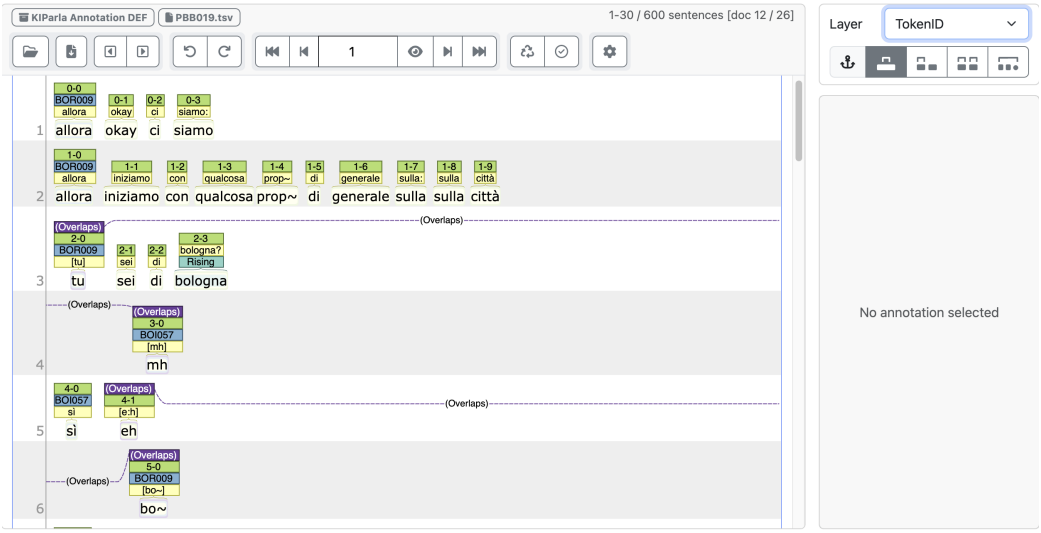

Figure 5.2: Layers with imported metadata

A string feature was linked to each layer. Then, for each interactional phe-

[1]More information can be found in the User Guide in *Appendix D: WebAnno TSV 3.3 File format*: `https://inception-project.github.io/releases/34.2/docs/user-guide.html`

nomenon, a set of tags was used: BC for backchannels, OR for other-initiated repair, SR for self-repair and FP for filled pauses. Figure 5.3 provides a visual representation of backchannel, filled pauses and self-repair annotation.

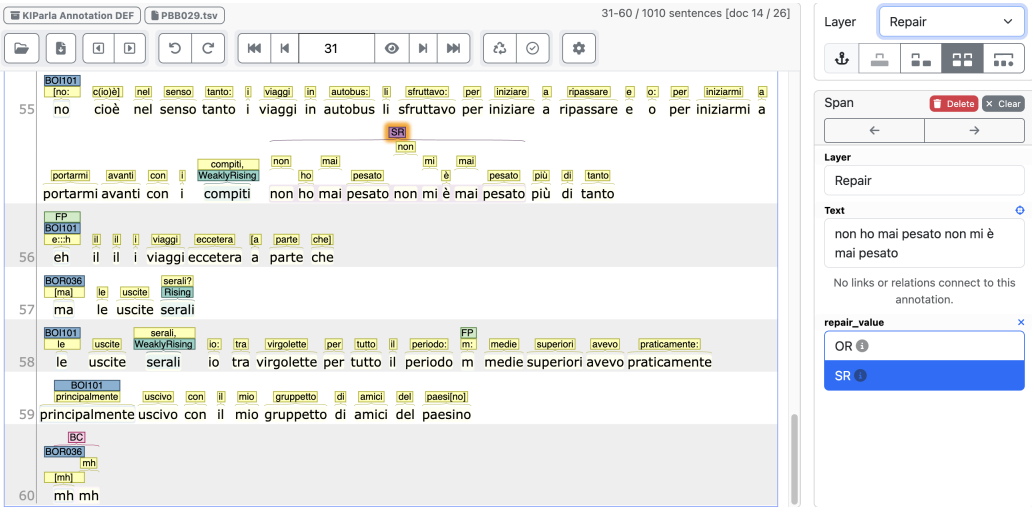

Figure 5.3: Interactional phenomena annotation.

Throughout the annotation process, listening to the conversations was essential to accurately interpret conversational dynamics. For this reason, the audio files were opened in ELAN (Max Planck Institute for Psycholinguistics, 2025) together with their corresponding `.eaf` transcription file and used as a reference.

5.4.1 Files Conversion: From .vert.tsv into WebAnno TSV (v3.3)

A custom Python script was designed to add metadata information from `.vert.tsv` files[2] to the `WebAnno TSV (v3.3)` file. The script reads each token with its associated metadata (token IDs, speakers, Jefferson features, intonation and overlap information) and reconstructs the full text units with the specific WebAnno structure, including the project layers and features.

[2]These files are available on Github for each module. For more information: `https://github.com/KIParla/KIP?tab=readme-ov-file#verticalized-content`.

The script assigns WebAnno-compliant sentence and token identifiers, computes character offsets and handles prosodic and overlap features, splitting tokens when partial overlaps occur. Each transcription unit is then exported with its corresponding `#Text=` line and token-level feature columns, producing a fully structured file ready for import and annotation into INCEpTION. This workflow made it possible to perform manual annotation on transcripts enriched with detailed conversational metadata, while maintaining full compatibility with the INCEpTION annotation environment.

The diagram in Figure 5.4 illustrates all the steps, from transcript upload and layer creation to metadata integration and annotation of interactional phenomena.

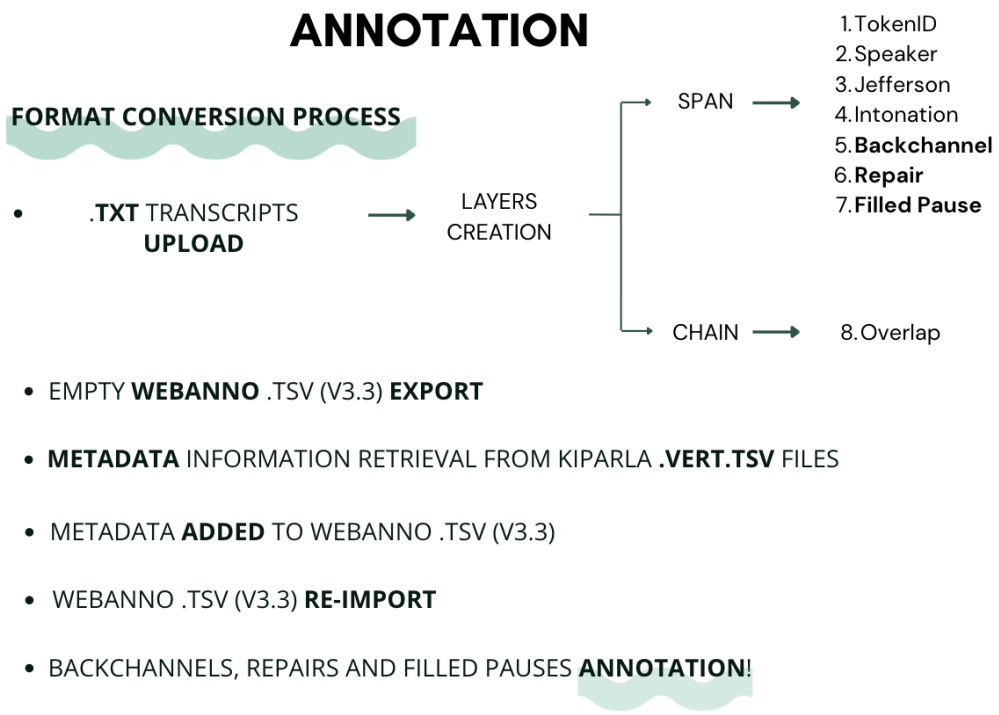

Figure 5.4: Overview of the format conversion process.

5.4.2 Criteria for Annotation

Before starting annotation, a set of criteria for each phenomenon was defined to facilitate their identification, ensure consistency across annotations and streamline the annotation workflow. Table 5.3 provides an overview of the annotated phenomena, their corresponding tags and examples.

Phenomenon	TAG	Subtype / Scope	Example
Backchannels	**BC**	Continuers, assessments, agreement, incipient speakership	*mhmh*, *va bene*, *ho capito*
Filled pauses	**FP**	Vocalic hesitation markers	*eh*, *ehm*, *mh*
Self-repair	**SR**	Same turn, transition space, third position	*vennero anda- ... andarono*
Other-initiated repair	**OR**	Open and restricted initiators	*eh?*, *che cosa?*

Table 5.3: Overview of annotated interactional phenomena with examples.

The following sections describe in detail the annotation criteria adopted for each phenomenon.

While much of the literature focuses primarily on vocalized forms of backchannels (such as *mhmh*) among those that are unlikely to be reproduced by ASR systems, the present study also considers multi-unit and lexicalized backchannels, inspired by the classification proposed by Mereu et al. (2024). Therefore, not ontly continuers were included, but also assessments, agreement and incipient speakership tokens (see Section 2.2). This choice is motivated by the interactional context in which backchannels typically occur: they frequently overlap with the current speaker's utterance, and, as shown in the literature, overlapping speech represent another major challenge to handle for ASR systems. As a consequence, lexical and multi-unit backchannels occurring in overlap may also be prone to omission or misrecognition. Among the variety of annotated backchannels it is therefore possible to find, for example, short lexical feedback such as *ho capito* ('I see' / 'I understand') or *va bene* ('okay' / 'all right'), as well as longer sequences such as *okay va bene sì sì ho capito*. An additional motivation for their inclusion, mainly for shorter sequences, is that ASR systems are also known to suppress very

short turns (Cumbal et al., 2021). As a result, all the described forms were considered to be particularly relevant for the study.

Regarding repair, it was decided to differentiate between sequences of other-initiated repair and self-repair, given their distinct sequential organization and interactional functions (see Section 2.3). For other-initiated repair, the annotation focuses on open initiators, such as *eh?* ('huh?'), *che cosa?* ('what?'), *prego?* ('pardon?'), and restricted requests, which include wh-questions. As in the case of backchannels, while it is unsurprising that non-lexical forms such as *eh?* may be missed, ASR systems may also suppress very short turns, thus increasing the possibility that these types of initiators may not be represented in the output. Instead, restricted offers, which are more elaborate, were not included in the annotation. These sequences typically do not occur in overlap, as recipients typically wait for the current speaker to resolve the trouble through self-repair (Schegloff et al., 1977).

Concerning self-repair, all the cases described in 2.3 were considered: those within the same turn, in the transition space and in third position. However, some specific cases require classification.

Interruptions followed by a reformulation were annotated as self-repair when they involved an explicit correction or reconstruction of the speaker's own ongoing talk. This includes cases such as:

> *perché non partecipò lui attivamente* ***allo squadri- [PAUSE] al [PAUSE] cioè era una figura che si avvantaggiava*** *del eh [PAUSE] partito nazionale fascista [...]*

Simpler cases were also included, such as *vennero anda- [PAUSE] andarono*. Self-repair sequences involving interruptions due to mispronunciation or articulation problems were likewise annotated, as in: *ventiminu- [PAUSE] ventimila volumi*. Other types of interruptions associated with hesitation phenomena or caused by overlap with another speaker were not considered, as they do not constitute self-repair operations.

Reformulations were annotated as self-repair when they resulted in repetition or partial recycling of previously produced material, reflecting an attempt to reconstruct the utterance. Examples include:

> (1) *di* ***gran parte*** *[PAUSE] diciamo della perdita di* ***gran*** *[PAUSE]* ***parte*** *[...]*
>
> (2) ***l'unico*** *censi- [PAUSE] diciamo* ***l'unico*** *[PAUSE] calcolo approssimativo [...]*

In many cases, these sequences follow a recurrent pattern in which the repairable followed by a pause, a discourse marker (e.g., *cioè, diciamo* and/or a filled pause, and culminating in a reformulation. Reformulations that functioned instead as exemplifications or expansions without repetition were excluded. This distinction was motivated by the focus of the present study: while repetitions are frequently subject to omission in ASR output, especially non-final repetitions (Goldwater et al., 2008), reformulations relying on entirely different lexical material are less likely to be suppressed and therefore fall outside the scope of the analysis.

As far as filled pauses are concerned, in line with Spreafico (2012), they were distinguished from other instances of hesitation such as false starts and final-word prolongations. It should be highlighted, however, that in some cases the distinction between a filled pause and other short vocalic items, such as the third-person singular indicative of the verb *essere* ('to be') or the conjuction *e* ('and'), is not categorical, but may depend on the perception of each transcriber. These cases often involve subjective judgments related to prosody, duration and interactional context. For this reason, annotation was guided by careful listening of the audio. All instances perceived as filled pauses were retained, including those that were not consistently represented in the transcripts. When a filled pause occurred as a part of a self-repair sequence, it was annotated both as a filled pause and as part of the self-repair. Moreover, cases of co-occurring *eh* and *mh* were grouped under a single filled pause category, as they function as a single hesitation marker *ehm*.

More generally, it must be taken into consideration the fact that the KIParla corpus is a modular and incremental resource, and, as such, it was not always possible to control and uniform spelling variation with regard to these phenomena. Also, the project has grown over the years and not all transcription conventions were solidly defined from the start.

5.5 ASR Transcription

The choice on the ASR tool for this study was guided by the need to support the transcription and analysis of interactional phenomena in conversational speech. For this reason, diarization was considered a central component of the transcription process: accurate speaker attribution, appropriate segmentation strategies and the preservation of short turns are essential requirements

in this context.

5.5.1 The Tool

The transcription pipeline adopted in this study is based on a custom tool currently under development at the Department of Translation and Interpreting by Gabriele Carioli[3]. By splitting the input into shorter blocks and processing them sequentially, its primary goal is to address the limitations of both speaker diarization systems (PyAnnote[4]) and ASR models (Whisper) when applied to very long audio files. Although the tool was designed specifically for long recordings (specifically, podcasts), it can be adapted to short conversational audio as well. The audio files used in the present study are relatively short: in this case, the pipeline operates on a single block. The modular structure of the tool allows control over diarization, segmentation, filtering and transcription parameters, and can produce time-aligned and speaker-labeled output, which is particularly relevant in the case of KIParla.

The first stage of the pipeline performs speaker diarization with PyAnnote: the system estimates the number of speakers given a predefined range by the user and then segments the audio accordingly. An optional *exclusive mode* can be enabled, which assigns overlapping speech to the predominant speaker rather than attributing it to multiple speakers. This option may be less suitable for short conversations where overlaps are interactionally meaningful.

After diarization, the pipeline allows for manual inspection and normalization of speaker labels through a speaker map. This step ensures consistent speaker identities across segments and prevents the inclusion of 'ghost' speakers, which may arise from overlapping segments or background speech.

The filtering phase then aggregates consecutive segments produced by the same speaker if they are separated by short pauses and optionally removes isolated segments below a minimum duration threshold. In this case, overly aggressive filtering risks excluding short turns, that may last below half a second: consequently, a low minimum duration threshold was adopted to preserve short utterances that would otherwise be discarded.

Transcription is performed by Whisper. The pipeline allows the user to choose several decoding parameters to optimize transcription: in this study,

[3]https://github.com/bilo1967/DIT.DaT
[4]https://www.pyannote.ai/

an automated hyperparameter tuning using Optuna is performed. This approach allows for an exploration of the decoding parameter space without retraining the model, identifying configurations that better represent conversational speech. Parameters include, for example, the decoding temperature, beam size, number of alternative hypotheses considered (best_of), threshold for detecting non-speech segments (no_speech_threshold) and constraints aimed at reducing hallucinated output (compression_ratio_threshold).

Finally, the tool generates speaker-aligned subtitle files that can be manually reviewed and corrected using standard subtitle-editing software. These corrected subtitles are then converted into textual formats.

Alternative solutions such as WhisperX (Bain et al., 2023) were also considered. However, WhisperX relies on an older version of the PyAnnote diarization framework and control over diarization and segmentation is limited. Based on preliminary tests and discussions with the tool developer, WhisperX resulted in less reliable speaker segmentation on the present conversational data. Given the relevance of turn structure and speaker attribution for the phenomena under investigation, a more modular and configurable pipeline was preferred. Also, the workflow adopted allows ASR output to be systematically controlled and revised.

5.5.2 Pre-Processing

ASR-generated transcripts were pre-processed in order to make them comparable with the gold standard. In particular, this procedure aimed to reduce superficial mismatches unrelated to transcription accuracy. All text was lowercased, numerical digits were converted into their corresponding word forms and punctuation was removed, with the exception of apostrophes. Extra whitespace was also normalized. Speaker identifiers were preserved and kept separate from the transcribed content in order to maintain alignment with the reference format.

5.6 Decoding Configurations

Four configurations were defined and applied to the same set of audio files to perform segmentation and diarization. As described in the dataset section (Section 5.3), they were evaluated on a fixed validation subset, Subset A.

All configurations share a common processing pipeline and differ only with respect to a limited set of parameters, allowing for controlled comparison. Table 5.4 shows the four configurations, denominated A, B, C and D.

config.	**exclusive mode**	**min. pause (s)**	**min. duration (s)**	**segmentation sensitivity**
A	yes	2.0	0.25	low
B	no	2.0	0.25	medium
C	no	1.0	0.25	high
D	no	1.0	0.20	very high

Table 5.4: Overview of the four diarization and segmentation configurations used in the optimization process.

Across all configurations, speaker diarization was performed assuming a fixed number of two speakers per interaction, and automatic speaker mapping was enabled. Given that the audio files are relatively short, block-based processing was retained for comparability with the pipeline; however, block segmentation did not introduce any additional partitioning and each recording was processed as a single block. The configurations primarily vary along three dimensions: handling of overlapping speech, minimum pause duration for merging consecutive segments and minimum segment duration threshold.

Configuration A enables exclusive mode during diarization, forcing overlapping speech segments to be assigned only to the dominant speaker. This setting reduces the number of overlapping segments and produces cleaner speaker turns, at the cost of potentially suppressing brief utterances produced in overlap. Segmentation parameters were set conservatively, with a minimum segment duration of 0.25 seconds and a minimum pause of 2 seconds to merge adjacent turns. This configuration represents the most restrictive setup and approximates a pipeline optimized for fluent, mostly non-overlapping speech.

Configuration B is identical to Configuration A except for the handling of overlaps: exclusive mode is disabled, allowing overlapping speech to be attributed to multiple speakers. This choice tries to preserve overlaps while maintaining the same conservative segmentation threshold (minimum duration 0.25 seconds and minimum pause of 2 seconds).

Configuration C further relaxes the segmentation constraints by reducing the minimum pause threshold to 1 second, while keeping the minimum

segment duration at 0.25 seconds. This setting allows for a finer-grained segmentation of speaker turns, increasing sensitivity to rapid turn-taking and brief interactional events.

Finally, Configuration D represents the least restrictive setup. In addition to disabling exclusive mode and using a 1-second pause threshold, the minimum segment duration is further reduced to 0.20 seconds. This configuration is designed to retain very short speech segments, including minimal responses and hesitation markers, at the expense of increased fragmentation. Therefore, it provides an upper bound on segmentation sensitivity and allows to test how aggressively preserving short turns affects transcription quality and error patterns.

5.6.1 Automatic Optimization of Decoding Parameters

In order to explore the impact of decoding parameters on ASR performance, an automatic optimization pipeline was implemented using Optuna Akiba et al. (2019). Decoding parameter selection was framed as a minimization problem, as the objective function consisted of error-based metrics (i.e., Word Error Rate and a composite loss), which are conventionally optimized by reducing their value. Both metrics were computed against manual reference transcriptions.

Each optimization trial corresponds to a full execution of the transcription pipeline, under a specific combination of decoding parameters. The search space includes a set of inference-time parameters which are known to influence Whisper's decoding behavior. In addition, the optimization process allowed for the selection among different Whisper model sizes. Parameter values were automatically sampled at each trial using the Optuna Tree-structured Parzen Estimator (TPE) (Bergstra et al., 2011) sampler. The parameters that were explored, together with the value ranges, are listed below[5].

- **Temperature** $[0.0, 1.0]$: controls the randomness of the decoding process by modifying the probability distribution over the model's predictions. Lower values lead to more deterministic and confident output,

[5]https://whisper-api.com/docs/transcription-options/#setting-the-model-size

while higher values increase randomness and allow more diverse hypotheses.

- **Beam size** $[1, 10]$: specifies the number of alternative hypotheses retained during decoding. Larger values increase search breadth at the cost of higher computational time and memory usage.
- **Best of** $[1, 10]$: number of candidate hypotheses generated internally and considered for the final transcription. Together with beam search, it balances exploration and output quality.
- **No-speech threshold** $[0.2, 0.8]$: controls how sensitive the model is to detecting silence or non-speech segments. Higher values make the model more likely to classify low-confidence audio as non-speech and discard it from transcription.
- **Compression ratio threshold** $[1.5, 2.8]$: helps detect and suppress hallucinated or repetitive output by comparing the length of the generated text with what would be expected given the audio content. Lower values apply stricter filtering.
- **Condition on previous text** (true, false): determines whether decoding is conditioned on previously generated text. Enabling this option can improve global coherence across segments, but may also propagate errors from earlier output.
- **Patience** $[0.2, 0.8]$: determines how long the decoding process continues searching for better transcription hypotheses. Higher values allow a more thorough search at the cost of increased decoding time.
- **Length penalty** $[0.2, 0.8]$: controls the preference for shorter or longer transcription hypotheses during decoding. Lower values encourage shorter outputs, while higher values favor longer ones.
- **Model size** (small, medium, large-v3, turbo): determines the balance between transcription accuracy and computational cost. Larger models are generally more accurate but slower, while smaller models are faster but potentially less accurate, depending on audio quality.

For each trial, each configuration described in Section 5.6 was applied to Subset A. Specifically, the optimization was carried out on 50 trials. In

each trial, Optuna automatically sampled a new combination of decoding parameters according to the predefined ranges described above. Each sampled configuration defines a complete decoding setup.

The complete transcription workflow was therefore executed, and subtitles and plain-text transcripts were generated. Each transcript was subsequently normalized through the pre-processing pipeline described in Section 5.5.2 and aligned at the word level with the corresponding gold-standard transcripts. Once all conversations in the subset were processed, their individual scores were averaged to obtain a single performance value for the trial.

To reduce computational cost, evaluation was performed incrementally. Instead of processing all conversations at once, the validation subset was divided into small chunks (seven conversations at a time). After each chunk was evaluated, the current mean score was reported to the optimization engine. A median-based pruning strategy was applied: if a trial's intermediate performance was worse than the median performance of previously completed trials, that trial was terminated early. This prevented unnecessary computation on clearly suboptimal parameter configurations.

At the end of the optimization process, Optuna selected the parameter configuration that achieved the lowest overall score among the completed trials. The full set of trial results and the best parameter configuration were stored in two `.json` files for reproducibility (respectively, `optuna_trials.json` and `parameters.json`).

This approach treats Whisper as a black box, and ensures that parameter selection is based on empirical transcription performance, while maintaining computational efficiency through early stopping of weak configurations. As described at the beginning of this Subsection, the optimization procedure was implemented in two variants, differing in the definition of the objective function.

5.6.2 WER-based Optimization

In the WER-based optimization setting, decoding parameters were optimized using the Word Error Rate (WER) (Klakow and Peters, 2002) as the objective function. WER was computed by aligning ASR output with the corresponding manually transcribed gold standard transcriptions at the word level and measuring the proportion of substitutions, deletions and insertions.

Specifically, WER is defined as:

$$\text{WER} = \frac{S + D + I}{N} = \frac{S + D + I}{S + D + C}$$

where:

- S = substitutions
- D = deletions
- I = insertions
- C = correct words
- N = total number of words in the reference (i.e., Gold) transcript

Insertions are defined as tokens present in the ASR output but missing in the gold standard; deletions correspond to tokens missing from the ASR output; finally, substitutions occur when both tokens are present but differ. For each aligned word pair, errors were counted accordingly, while the denominator N accumulated the number of reference tokens, excluding insertions. The WER for each transcription was computed by summing all errors (insertions, deletions, and substitutions) and dividing by the total number of reference tokens. This metric provides a normalized and interpretable measure of transcription performance, with lower values indicating higher similarity to the gold standard (Klakow and Peters, 2002).

WER was selected as the primary optimization metric because it represents the standard measure for evaluating transcription accuracy for ASR systems. Although the literature has highlighted limitations of this metric in the evaluation of conversational speech, particularly with respect to interactional phenomena, it remains the dominant benchmark for error analysis and parameter tuning (see Section 3.4. Given the absence of widely adopted evaluation metrics specifically designed to capture interactional complexity in spontaneous speech, WER is used in this study as a baseline objective function.

5.6.3 Interaction-aware Optimization

The limitations of the Word Error Rate in capturing interactional phenomena, together with the absence of established interaction-centric evaluation

metrics, motivate the exploration of an alternative, event-sensitive objective function. This alternative optimization pipeline, that will be referred to as the Interaction-aware optimization strategy, implements an objective function defined as a weighted combination of the standard Word Error Rate (WER) and two event-specific components that aim to capture the tendency of the ASR system to suppress or correctly produce interactional events. The resulting loss function is formalized as follows:

$$\mathcal{L} = \text{WER} + \lambda \cdot \text{Suppression}_{\text{event}} - \mu \cdot \text{Production}_{\text{event}}$$

Suppression captures how frequently interactional events are lost by Whisper. It is defined as the proportion of event tokens that are either deleted or substituted:

$$\text{Suppression}_{\text{event}} = \frac{D_{\text{event}} + S_{\text{event}}}{N_{\text{event}}}$$

Production, instead, measures the proportion of correctly transcribed event tokens, and is defined as:

$$\text{Production}_{\text{event}} = \frac{C_{\text{event}}}{N_{\text{event}}}$$

The coefficients in the loss function λ and μ regulate the relative influence of interactional phenomena. $\lambda = 0.3$ was used to penalize the suppression of interactional events, while $\mu = 0.1$ introduced a mild reward for their correct production. Both of these values were intentionally kept low: in this way, penalization was not excessive and reward did not risk to produce overly-hallucinated output.

Interactional events were identified following the a form-based procedure, using the following predefined set of tokens:

- *eh*
- *ehm*
- *ah*
- *oh*
- *mh*
- *mhmh*
- *mm*
- *okay*
- *sì*

- *bene*
- *esatto*
- *cioè*
- *diciamo*
- *insomma*

Word-level alignments between hypothesis and reference were also used for their identification.

This Interaction-aware optimization should be interpreted as an exploratory extension of the WER-based approach. It is important to underline that its aim is not maximizing the production of interactional events. Instead, it seeks to reduce their systematic suppression when they are present in the reference transcription, while preserving overall transcription accuracy.

5.6.4 Quantitative Error Analysis

Following both optimizations, a quantitative error analysis was conducted in order to examine the distribution of error types produced by the optimized decoding configurations. This analysis was based on word-level alignments between the ASR output obtained from the best-performing optimization trial and the gold standard transcripts, generated for each configuration.

An R (R Core Team, 2021) script was used to calculate summary statistics of both WER and Interaction-aware optimization values obtained across Optuna trials, for each configuration. Specifically, mean, variance and standard deviation were calculated, while the WER distribution was visualized with boxplots.

For each aligned word pair, an edit operation was assigned according to standard WER categories: deletion (DEL), insertions (INS), substitution (SUB) or match (OK). In order to investigate the behavior of the ASR system with respect to interactional phenomena, aligned tokens were further classified into two categories: *event* and *other*. Event tokens were identified using through the same form-based definition listed in Section 5.6.3, thus without specifying their functional distinction.

Error distributions were aggregated across all conversations for each configuration, yielding normalized proportions of deletions, insertions, substitutions, and correct matches for event tokens.

It is important to underline that this analysis is exploratory in nature. The identification of interactional phenomena relies on a predefined set of

form-based tokens and, therefore, does not aim to exhaustively capture all conversational events. For example, multi-unit phenomena may not be explicitly considered within this framework. As a result, the analysis does not provide a complete account of interactional behavior in ASR output, but rather an approximation that allows suppression or substitution to be quantitatively examined. This quantitative error analysis provides the empirical basis for the subsequent exploration of an Interaction-aware optimization strategy, aimed at mitigating the systematic suppression of interactional phenomena observed under WER-based optimization.

5.6.5 Control Analysis on Subset B

To verify that the observed behavior was not specific to the optimization subset, a complementary analysis was conducted on an additional set of conversational data (Subset B) not used during parameter tuning.

For each optimization strategy, only the best-performing configuration (i.e., configuration A, see Tables 6.1 and 6.3) was retained and applied unchanged to Subset B. This allowed to assess whether the differences observed between the two optimization strategies generalize beyond the data used for parameter selection.

To summarize, Figure 5.5 illustrates the processing pipeline applied to Subset A, including segmentation configurations, automatic parameter optimization under two objective functions (WER-based and Interaction-aware), quantitative error analysis and final control evaluation on Subset B.

5.7 Gold Annotated Transcripts

5.7.1 Conversion of INCEpTION Annotations to Vertical Format

Before extracting quantitative data from annotations, transcripts were converted from the WebAnno TSV V3.3 into a vertical format (`.vert.tsv`) compatible with the `.vert` files available on GitHub for each KIParla module. This step was therefore carried out to preserve the original vertical structure while enriching it with annotations. Maintaining the same format

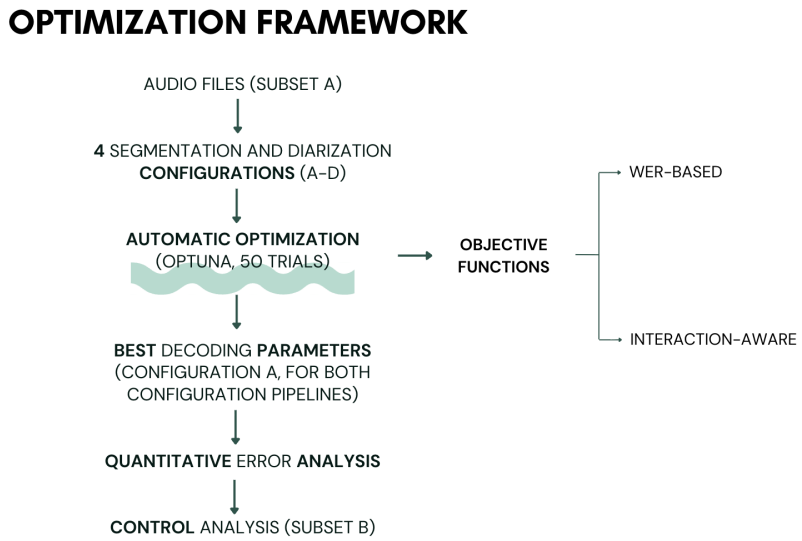

Figure 5.5: Schematic representation of the decoding and optimization framework.

also facilitates the extraction of interactional statistics and the readability of the annotated data at the token level.

For each conversation, the corresponding and pre-existing vertical file (`.vert.tsv`) was retrieved. Vertical files represent each token as a structured row in a column-based format, making the alignment between textual content and annotation layers explicit. Interactional annotations were therefore exported from the INCEpTION outputs to the vertical files. Specifically, they were updated by introducing an additional *event* column.

The resulting enriched vertical files provide a unified, structurally transparent representation of the corpus, in which tokenization, speaker structure and interactional annotations are integrated in a single format.

5.7.2 Statistics

Interactional events were extracted from the manually annotated dataset to obtain a descriptive analysis of each phenomenon. They were identified through a regular-expression-based matching procedure that captured both single-token annotations (e.g. BC) and span-based annotations (e.g., BC[2]). For span annotations, the script ensured that multi-token spans were counted

as a single event. Two distinct normalization procedures were followed, each corresponding to a different analytical level.

At the speaker level, event frequencies were normalized per minute of individual speech, that was extracted from the corresponding ELAN (`.eaf`) files by summing the durations of each speaker's transcription units within the same 20-minute window. This allowed the calculation of normalized event frequencies per minute of speech:

$$\text{Events per minute of speech} = \frac{\text{Raw event count}}{\text{Speech duration (minutes)}}$$

This measure captures the relative density of interactional phenomena in relation to each speaker's actual speaking time and allows for meaningful comparisons across participants with different speech contributions: very often speaker differ substantially in the amount of time they hold the floor.

At the interaction level (i.e., free conversations, exams/office hours and semi-structured interviews), event counts were instead normalized per minute of conversation. This measure reflects the overall density of interactional phenomena within the conversational context, independently of how speaking time is distributed among participants. This analysis provides a macro-level perspective on how interaction types differ in their global interactional dynamics.

These two normalization procedures therefore serve complementary purposes: the former captures individual behavior relative to speech production, while the latter characterizes the structural density of interactional phenomena within entire conversations. They are not directly comparable, however they address different analytical questions.

5.7.3 Word-level Alignment

To compare ASR outputs with the manually annotated data, a word-level alignment procedure was implemented. Alignment was performed separately for the two decoding strategies (WER-based and Interaction-aware). Non-verbal annotations and undefined speaker labels were excluded from the alignment, only lexical tokens were retained.

ASR outputs were loaded from plain text files containing speaker-labeled utterances. Prior to alignment, a minimal normalization step was applied to

both sequences: the *ok* from the ASR output was systematically normalized to *okay* in order to ensure consistency with the KIParla conventions.

5.7.4 Normalization

Before computing the evaluation metrics, two procedures were carried out in order to identify potential sources of systematic mismatch: (1) an exploratory extraction of raw substitution patterns, and (2) a manual inspection of alignment outputs. This preliminary inspection served to identify systematic mismatches that did not reflect genuine recognition errors, but rather orthographic variation or tokenization inconsistencies between the gold standard transcriptions and Whisper outputs. Both procedures were therefore followed to guide subsequent normalization choices, ensuring that evaluation decisions were empirically motivated.

All token-level raw substitutions were extracted from the aligned files without applying any normalization criteria. Table 5.5 reports the ten most frequent raw substitution patterns for both decoding strategies, showing that several high-frequency substitutions involve non-lexical vocalizations such as *eh*, *ehm*, and nasal backchannels (*m, mh, mhmh*). These forms display substantial orthographic variability across gold standard transcriptions and ASR output.

WER-based (raw)			**Interaction-aware (raw)**		
Gold	**Whisper**	**Freq.**	**Gold**	**Whisper**	**Freq.**
eh	e	50	eh	e	49
mh	grazie	26	mh	grazie	28
eh	è	17	eh	è	17
ehm	e	16	mh	e	13
mh	e	15	ehm	e	12
mh	non	13	mh	non	11
mh	è	12	mh	che	10
sì	grazie	11	sì	grazie	10
mh	che	10	mh	è	9
eh	grazie	9	okay	e	9

Table 5.5: Top 10 raw substitution patterns across optimization strategies.

In parallel, a manual reading of the alignments revealed additional struc-

tural inconsistencies unrelated to lexical recognition performance. First, orthographic inconsistencies related to apostrophes were resolved. In the gold standard transcriptions, forms separated by apostrophe were tokenized into two units (e.g., *all'* + *interno*), whereas Whisper produced a single token (*all'interno*). To prevent mismatches derived from tokenization differences, consecutive tokens were merged when the first ended with an apostrophe. Second, truncated tokens were examined. Truncations, marked with a tilde in the gold reference transcription, are characteristic of repair phenomena but cannot be explicitly encoded by Whisper. Therefore, gold tokens ending with the truncation marker were considered correctly matched if the corresponding Whisper token began with the same orthographic sequence, preventing a structural penalization of repair sequences due to architectural limitations of the ASR system. Third, [PAUSE] labels were excluded from substitution counts, since Whisper does not explicitly model pause markers. Including them would have introduced a systematic penalty unrelated to lexical recognition performance. At the same time, these adjustments may lead to a partial overestimation of the system's ability to handle certain interactional phenomena. By reducing penalties associated with truncations and pauses, the evaluation becomes less sensitive to aspects of the signal that are relevant from a conversation-analytic perspective. These decisions therefore reflect a methodological trade-off: they aim to avoid structural bias in the metric, while acknowledging that they introduce a different form of evaluative bias. Finally, manual inspection revealed substantial orthographic variability in the representation of nasal backchannels. It was observed that in the gold standard transcriptions these forms display orthographic variation (e.g., *m, mh, mm, mhmh*), while Whisper produces alternative but functionally equivalent variants such as *mmm* or *mmh*. Given their non-lexical and discourse-related function, tokens composed exclusively of the characters *m* and *h* were treated as equivalent across Gold and ASR output. Guided by the raw substitution analysis, similar criterion was applied to filled pauses. In cases where the gold standard transcription contained the form *eh* and Whisper produced *e*, tokens were treated as equivalent when annotated as filled pauses (FP). This decision was motivated by their phonetic proximity, as hesitation markers are often realized with reduced vocalic quality that may approximate the conjunction *e*. Moreover, the majority of *eh* $\rightarrow$ *e* substitutions observed in the data (40 out of 50 cases) involved filled pauses, whereas only a minority concerned backchannels. Thus, such normalization was not applied to the backchannel uses of *eh*, which are typically produced

with distinct prosodic properties: equating them with the lexical conjunction would have risked obscuring functional differences. The adjustment was therefore restricted to filled pauses, where the substitution pattern appeared systematic and phonetically motivated.

Raw substitution counts also revealed recurrent cases in which the token *eh* was transcribed by Whisper as *è*. Given their phonetic proximity, normalization was considered. However, manual inspection did not reveal a consistent pattern restricted to a single, specific use. While *eh* functions as a backchannel or filled pause, *è* represents a lexical verb form. Normalizing these cases would therefore have introduced an interpretative assumption rather than correcting a purely orthographic variation. For this reason, such substitutions were retained in the analysis.

5.7.5 Word Error Rate

The overall transcription accuracy on the manually annotated subset was assessed by computing the Word Error Rate (WER) for each conversation and for each decoding strategy, resulting in paired WER values for every annotated recording.

To facilitate comparison, WER values were visualized at two levels. First, a conversation-level plot was generated, displaying the WER obtained under each optimization strategy for every annotated conversation. For reasons of readability, conversations were grouped by interaction type (free conversation, exams/office hours, and semi-structured interviews); this subdivision serves a purely descriptive purpose and does not imply separate statistical analyses at the interaction-type level. Rather, it allows clearer inspection of performance shifts within comparable conversational settings while preserving the paired structure of the data. Second, aggregate boxplots were produced to summarize the overall distribution of WER values across conversations for each optimization strategy, highlighting differences in central tendency and dispersion.

To support graphs, descriptive statistics (mean, variance and standard deviation) were calculated across conversations to provide a quantitative summary of the observed distributions.

5.7.6 Mean Match Ratio

Following the normalization procedure, interactional phenomena were quantified. The analysis was conducted separately for the two decoding strategies, using the aligned `.align.tsv` files as input. Metadata regarding interaction types were imported from `conversations.csv` (see Appendix A, Table A.1) to enable interaction-level aggregation of the results. Each annotated sequence was treated as a single event, including multi-token phenomena that were identified through shared indices (e.g., SR[59]). For each annotated event e, a match ratio was defined as:

$$\text{MatchRatio}_e = \frac{M_e}{T_e}$$

where M_e represents the number of correctly matched tokens in event e, and T_e corresponds to the number of evaluable tokens in the same event. A token was considered correctly matched if it was identical in the gold standard transcription and in the Whisper output, except for normalized cases outlined in Subsection 5.7.4.

For each phenomenon type $p \in \{BC, FP, SR, OR\}$, the overall score was calculated as the simple mean of the event-level match ratios:

$$\text{MeanMatchRatio}_p = \frac{1}{E_p} \sum_{e=1}^{E_p} \text{MatchRatio}_e$$

where E_p denotes the total number of events of type p. A simple mean was deliberately adopted, so that each event contributed equally to the final score, independently of its token length. This prevents longer repair sequences from disproportionately influence the evaluation.

Results were generated at two levels of analysis: at the conversation level and at the interaction level.

5.7.6.1 Conversation-level

At the conversation level, for each conversation and each optimization condition, the script computed the number of events per phenomenon and the corresponding mean match ratio. To better assess whether differences between the two optimization strategies were consistent across individual conversations, paired difference scores were calculated as:

$$\Delta = \text{MeanMatchRatio}_{\text{Interaction-aware}} - \text{MeanMatchRatio}_{\text{WER-based}}$$

Positive values indicate higher performance under the Interaction-aware configuration, whereas negative values indicate higher performance under the WER-based configuration. Boxplots were generated to visualize the distribution of these paired differences, allowing inspection of central tendency (median), dispersion (interquartile range) and potential outliers across conversations.

Because each conversation yields two related measurements (one per optimization strategy), the observations are paired rather than independent. Statistical comparison therefore requires a paired test assessing whether one optimization strategy systematically yields higher mean match ratios than the other.

phenomenon	W	p**-value**
BC	0.917	0.038
FP	0.859	0.002
SR	0.829	$< .001$
OR	0.254	$< .001$

Table 5.6: Shapiro–Wilk normality tests on the paired differences (Interaction-aware – WER-based) for each interactional phenomenon.

The Shapiro–Wilk test (Table 5.6) revealed that the distribution of paired differences deviated significantly from normality for all phenomena ($p < 0.05$). Consequently, non-parametric Wilcoxon signed-rank tests were employed.

5.7.6.2 Interaction-level

At the interaction-level, events from conversations belonging to the same interaction type were grouped together, separately for each optimization condition, and the mean match ratio was computed over this combined set of events. Due to the limited number of conversations per interaction type, additional normality testing within each interactional subset was deemed unreliable. Consequently, inferential analyses were conducted at the overall interaction type level.

5.7.7 Substitution Rates

In addition to the mean match ratio, a complementary error analysis was conducted in order to examine the distribution of omission and substitution patterns. After applying the normalization procedures described in Subsection 5.7.4, token-level errors were extracted from the aligned files and categorized as omissions or substitutions.

Error counts were aggregated separately for the two optimization strategies. All calculations were performed in R, where total error counts, omission rates, and substitution frequencies were computed. Substitution frequencies were subsequently ranked in descending order in a top ten list to identify recurrent patterns. This analysis allows for a more fine-grained understanding of Whisper's error profile, distinguishing between complete omission of interactional phenomena and lexical misrecognition.

5.7.8 Overlap Analysis

To investigate whether conversational overlap affected the recognition of interactional phenomena, a token-level representation was extracted from the alignment files, including both correctly matched and unmatched annotated tokens. Unlike previous analyses restricted to successfully recognized cases, this representation retained all annotated tokens, allowing the estimation of recognition probabilities under different interactional conditions. For each annotated token, two binary variables were recorded:

- *match*, coded as 1 if the token was correctly recognized under the established matching criteria and 0 otherwise;
- *overlap*, coded as 1 if the token occurred in overlapping speech and 0 if produced in non-overlapping speech.

Overlap information was extracted from the `.vert.tsv` files based on the presence of overlap markers in the corresponding metadata column.

The association between overlap and recognition outcome was assessed through contingency tables (crosstabs) and Pearson's chi-square tests of independence implemented in R. Crosstabs display the joint distribution of two categorical variables by reporting the observed frequencies, namely, the actual counts of tokens falling into each combination of categories.

The chi-square test evaluates whether the observed distribution differs from the expected frequencies that would arise under the assumption of independence between the two variables. Expected frequencies are calculated from the marginal totals of the table and represent the counts that would be expected if overlap had no influence on recognition outcome. A statistically significant result therefore indicates that the distribution of matches and mismatches differs between overlap and non-overlap contexts.

Tests were conducted at three levels: (1) globally across all annotated tokens; (2) separately for each interactional phenomenon (BC, FP, SR, OR); and (3) separately for each optimization configuration (WER-based and Interaction-aware).

Assumptions concerning the chi-square test were verified by inspecting expected cell frequencies. In the case of other-initiated repairs (OR), two cells fell below 5: therefore, Fisher's exact test was applied, as it provides a more reliable estimate in situations involving small sample sizes.

Given that multiple follow-up tests were conducted on the same dataset (four phenomenon-specific tests and two optimization-specific tests), p-values were adjusted using the Holm correction procedure across all six comparisons (see Section 6.3.5).

5.8 Qualitative Error Analysis

The following analysis aims to illustrate and complement the quantitative results by examining selected cases in greater detail. All examples were identified from the word-level alignments, either automatically or manually.

The qualitative analysis on filled focused and backchannels exclusively focused on correctly matched cases, in order to examine the context of some of the instances in which Whisper system successfully preserved these phenomena. Examples were reported with the screenshot from ELAN, to better illustrate the placement and the sequential context of each event. In these two cases, differences between optimization strategies were not discussed.

As far as self-repairs are concerned, since they are structurally more complex than backchannels and filled pauses, they were evaluated at the level of event integrity. Therefore, a more structured analysis pipeline was required. With a Python script, each annotated self-repair was classified according to the extent to which its internal structure was preserved in the ASR output. Three categories were defined: *Full* (all components of the repair sequence

preserved), *Partial* (only part of the repair trajectory retained), and *None* (Complete omission of the repair sequence). Based on the aggregated results, two pie charts (one for each optimization pipeline) were subsequently created on Excel to illustrate self-repair integrity across the two decoding strategies. Examples were then retrieved from the word-level alignments and commented.

Since annotated other-initiated repairs are limited in number, all instances were automatically extracted from the word-level alignments and then manually checked. Specifically, a dedicated Python script was implemented to scan all alignment files and identify tokens labeled as OR or OR[n]. For each occurrence, the script retrieved the corresponding gold standard token, the ASR output token, and the associated speaker information, ensuring that all annotated other-repair events were systematically collected across conversations and optimization settings. The extracted instances were then inspected manually. Given the relatively low frequency of other-repair phenomena in the corpus, the qualitative analysis focused on both match and mismatch examples, provided as screenshots in the ELAN transcription environment.

The entire evaluation procedures applied to the manually annotated subset are summarized in Figure 5.6.

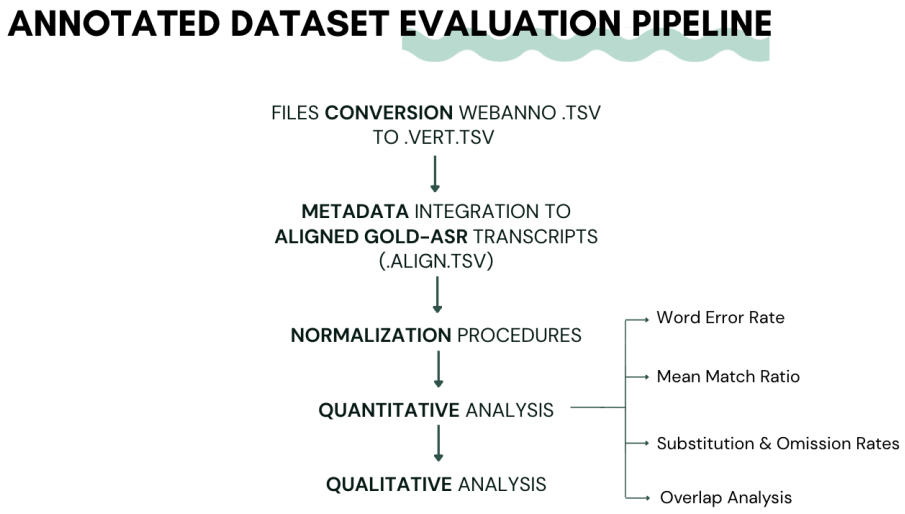

Figure 5.6: Evaluation pipeline of the gold annotated dataset.

Chapter 6

Results and Discussion

6.1 Chapter Overview

This chapter presents the results of the decoding optimization procedures and evaluates their impact on the transcription of interactional phenomena. The discussion integrates quantitative and qualitative analyses in order to assess both overall transcription accuracy and analyzed the contexts in which conversational phenomena are recognized or omitted.

The chapter is organized into four main parts. First, Section 6.2 reports the results of the decoding optimizations. WER-based and Interaction-aware optimization strategies are presented separately. For each configuration, the best-performing parameter set identified through Optuna is described, followed by an evaluation of global performance in terms of WER statistics (mean, variance, dispersion). A complementary evaluation on Subset B, which was not used during parameter tuning, is also included to test generalizability. Second, Section 6.3 presents the quantitative analysis of annotated interactional phenomena. After describing their distribution across interaction types, the section examines recognition performance using the mean match ratio at both the conversation level and the interaction level. Statistical testing is employed to determine whether differences between optimization strategies are significant. Additional analyses focus on omission rates, substitution patterns, and the effect of overlap on recognition accuracy. Third, Section 6.4 provides a qualitative error analysis. Selected examples drawn from the word-level alignments illustrate typical recognition patterns for filled pauses, backchannels, self-repairs, and other-initiated repairs. This

section aims to contextualize the quantitative findings by examining how ASR output reshapes the sequential and structural organization of interaction.

6.2 Decoding Optimization Results

6.2.1 WER Optimization

A WER-based optimization was performed independently for each of the configurations listed in Table 5.4 using Optuna, with the objective of minimizing WER and enhancing interactional phenomena representation through decoding parameter search. The complete list of optimization trials is reported in Appendix B. Table 6.1 reports the decoding parameters corresponding to the best-performing trial identified for each configuration. The WER value reported in the first row corresponds to the objective value achieved by the selected trial during optimization. It therefore represents the minimum WER obtained within the explored search space.

parameter	**A**	**B**	**C**	**D**
wer	0.357	0.391	0.368	0.364
temperature	0.143	0.636	0.555	0.375
beam size	3	10	4	4
best-of	6	2	5	9
no speech threshold	0.203	0.415	0.600	0.249
compression ratio threshold	2.121	2.501	2.240	2.613
model	large-v3	large-v3	large-v3	large-v3
condition on previous text	true	true	true	false
patience	0.592	0.626	0.344	0.217
length penalty	0.479	0.616	0.700	0.241

Table 6.1: Best-performing trial selected through Optuna optimization, for each configuration.

All optimal configurations rely on the large-v3 model, but differ substantially with respect to decoding hyperparameters. Configuration A is characterized by a relatively low temperature (0.143) and a small beam size (3), combined with a moderate best-of value (6). The no speech threshold is comparatively low (0.203), while compression ratio threshold and patience remain

moderate. Conditioning on previous text is enabled. Overall, this configuration favors relatively constrained decoding with limited search breadth and low stochasticity. Configuration B differs markedly from A, with the highest beam size (10) and a higher temperature (0.636), suggesting a broader and more exploratory decoding strategy. The best-of value is lower (2), but patience and length penalty are among the highest across configurations. Again, conditioning on previous text remains enabled. This combination reflects a more extensive search procedure, potentially aimed at minimizing substitution and insertion errors through increased hypothesis exploration. Configuration C adopts intermediate values for most parameters. The temperature (0.555) remains relatively high, while beam size (4) and best-of (5) indicate moderate search breadth. Notably, it exhibits the highest no speech threshold (0.600) and length penalty (0.700), suggesting a stronger bias toward longer sequences and more conservative silence filtering. Conditioning on previous text is also enabled in this setting. Configuration D, in contrast, is the only configuration in which conditioning on previous text is disabled. It combines a moderate temperature (0.375) with a beam size of 4 and the highest best-of value (9). Patience and length penalty are lower than in the other configurations. This setup reflects a decoding strategy that relies less on contextual continuation and more on hypothesis diversification through best-of sampling.

In terms of optimization outcome, Configuration A achieves the lowest objective value (WER = 0.357), followed by D (0.364) and C (0.368), while B shows the highest optimized WER (0.391). These values anticipate the general ranking later observed in the aggregated statistics, although the differences remain relatively contained.

The absence of a consistent pattern across configurations suggests that WER-based optimization does not converge toward a single robust decoding strategy but instead yields configuration-specific optima.

configuration	**mean**	**variance**	**SD**	**best**	**worst**
A	0.357	0.0090	0.095	0.206	0.535
B	0.399	0.0090	0.095	0.225	0.535
C	0.369	0.0094	0.097	0.203	0.563
D	0.365	0.0095	0.097	0.215	0.568

Table 6.2: Summary of WER statistics across configurations.

Table 6.2 reports the summary statistics of WER values across configu-

rations, while Figure 6.1 visualizes the distribution of WER values for each configuration. Overall, the four configurations exhibit comprable dispersion patterns, with variance values ranging between 0.0090 and 0.0095 and standard deviations around 0.095–0.097. The boxplots further confirm that the interquartile ranges are broadly aligned across configurations, indicating a relatively similar level of variability in performance.

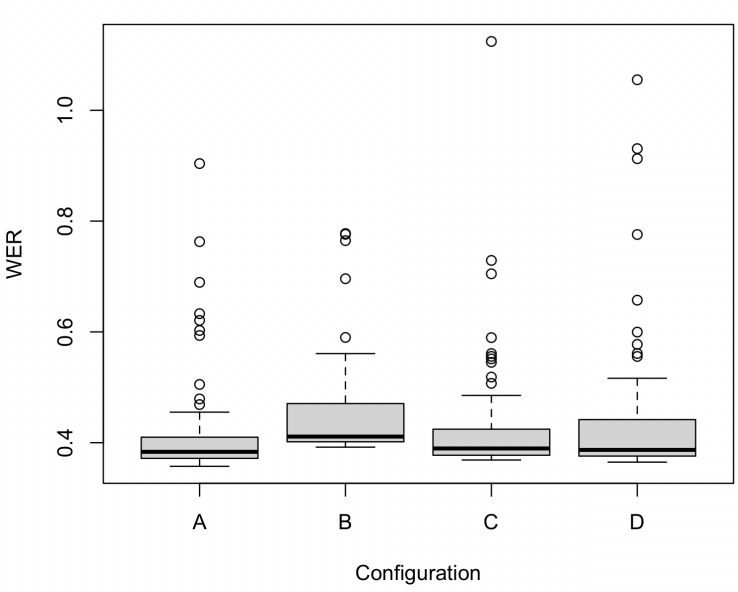

Figure 6.1: Distribution of WER across configurations.

In terms of central tendency, Configuration A achieves the lowest mean WER (0.357), followed closely by D (0.365) and C (0.369), while configuration B shows the highest (0.399). This pattern is also visible in Figure 6.1, where the median for Configuration B appears slightly higher than those of the other configurations. Although the differences are not dramatic, configuration A appears to provide the most stable and overall favorable performance among the four, combining the lowest central tendency with a relatively compact interquartile range.

With respect to extreme values, the best-case WER values are relatively close across configurations, with C achieving the lowest minimum (0.203), closely followed by A (0.206). However, the worst-case performance varies more noticeably: configurations C and D reach higher maximum WER values

(0.563 and 0.568, respectively), which is reflected in the more pronounces upper outliers visible in Figure 6.1. In contrast, both configurations A and B show a slightly lower worst-case value (0.535), indicating a more contained performance spread.

This data suggest that while global performance differences remain moderate, configuration A combines the lowest mean WER with comparatively controlled variability, making it the most balanced option among the evaluated configurations.

6.2.2 Interaction-aware Optimization

An Interaction-aware optimization was performed independently for each of the configurations listed in Table 5.4 using Optuna, with the objective of optimizing event-level preservation of interactional phenomena through decoding parameter search. The complete list of optimization trials is reported in Appendix C. Table 6.3 reports the decoding parameters corresponding to the best-performing trial identified for each configuration.

parameter	**A**	**B**	**C**	**D**
temperature	0.053	0.266	0.240	0.158
beam size	7	8	8	1
best-of	2	1	2	3
no speech threshold	0.539	0.354	0.617	0.465
compression ratio threshold	1.734	1.630	2.227	2.492
model	large-v3	large-v3	large-v3	large-v3
condition on previous text	false	true	false	true
patience	0.506	0.308	0.705	0.316
length penalty	0.241	0.295	0.614	0.397

Table 6.3: Best-performing trials selected through Interaction-aware optimization, for configurations A, B, C and D.

The optimized decoding parameters reveal both shared tendencies and configuration-specific adjustments. As in the WER-based setting, all configurations converge on the use of the large-v3 Whisper model. However, the decoding hyperparameters differ substantially. Configuration A is characterized by an extremely low temperature (0.053) combined with a relatively large beam size (7). This suggests a decoding strategy in which variability is

minimized while maintaining moderate hypothesis exploration during beam search. In this configuration, conditioning on previous text is disabled. Configuration B adopts a higher temperature (0.266) while maintaining a comparable beam size (8), but sets best-of to 1. Conditioning on previous text is enabled. Compared to A, this configuration allows greater decoding variability while still relying on a broad search over candidate hypotheses. The lower patience (0.308) and length penalty (0.295) values indicate a less constrained sequence expansion strategy relative to other configurations. Configuration C also employs a relatively high temperature (0.240) and beam size (8), but differs from B in several key parameters. It sets best-of to 2, exhibits the highest patience value (0.705), and the highest length penalty (0.614), while disabling conditioning on previous text. This combination reflects a decoding strategy that permits variability but strongly regulates sequence continuation, potentially encouraging longer or more structured outputs under the Interaction-aware objective. Configuration D is characterized by the lowest beam size (1), effectively reducing beam search to greedy decoding. It combines a moderate temperature (0.158) with the highest best-of value (3) among configurations and enables conditioning on previous text. Patience (0.316) and length penalty (0.397) remain intermediate. This profile suggests a more conservative search strategy, limiting hypothesis branching while relying on contextual continuation and limited sampling diversity. Across configurations, no clear trend emerges for no-speech threshold and compression ratio threshold. Although their values vary, this variation does not reveal a consistent decoding strategy. Overall, the absence of convergence toward a single decoding profile indicates that Interaction-aware optimization yields configuration-specific solutions rather than a uniform decoding strategy.

configuration	**mean**	**variance**	**SD**	**best**	**worst**
A	0.360	0.0085	0.092	0.207	0.543
B	0.393	0.0082	0.090	0.229	0.529
C	0.365	0.0077	0.088	0.199	0.521
D	0.366	0.0085	0.092	0.199	0.515

Table 6.4: Standard WER statistics obtained by the best Interaction-aware configuration, evaluated independently from the optimization objective

.

Table 6.4 illustrates the summary statistics of WER values obtained with the Interaction-aware configurations, while Figure 6.2 illustrates their distri-

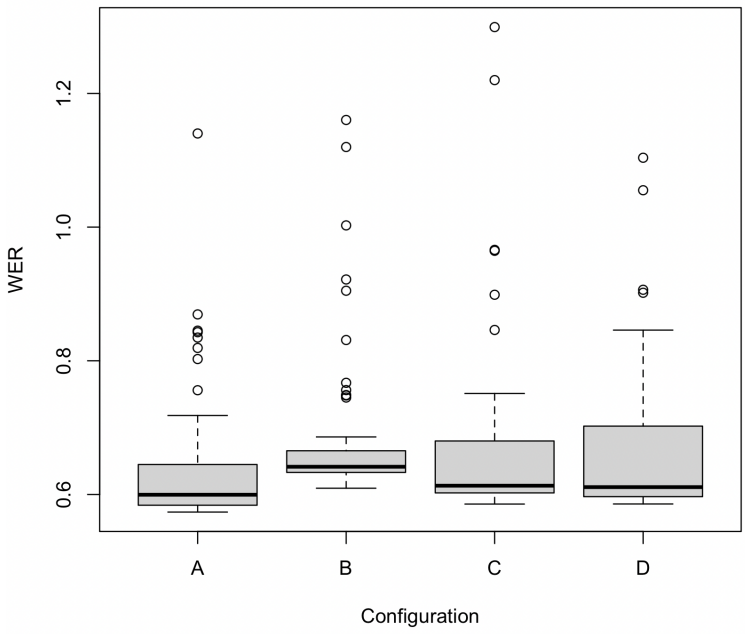

Figure 6.2: Distribution of WER across configurations.

bution across configurations. As in the WER-based setup, variability remains broadly comparable across configurations, with variance values ranging between 0.0077 and 0.0085 and standard deviations between 0.088 and 0.092. The boxplots confirm that the interquartile ranges are similar in width, indicating a comparable level of dispersion in transcription performance. Overall, dispersion appears slightly lower than in the WER-based case, suggesting a marginally more contained performance spread across conversations.

In terms of central tendency, Configurations A (0.360), C (0.365) and D (0.366) show very similar mean WER values, while B (0.393) again yields the highest mean WER. This pattern is visible in Figure 6.2, where the median for Configuration B appears slightly elevated relative to the others, and mirrors what was observed in the WER-based configurations, where B also exhibited comparatively weaker performance.

The best-case WER values are achieved by C and D (0.199), indicating that, under favorable conversational conditions, Interaction-aware optimization can reach lower error rates than configuration A. At the same time, worst-case values are slightly more contained for C (0.521) and D (0.515) compared to A (0.543), suggesting a somewhat improved robustness to more

challenging conversations.

Interaction-aware configurations therefore do not lead to substantial degradations in standard WER when compared to their WER-based counterparts. They maintain comparable average performance while exhibiting slightly reduced variance in some configurations, thus showing that incorporating event-sensitive components into the optimization framework does not destabilize global transcription accuracy. Again, Configuration A achieves the lowest WER, indicating that it remains the most favorable configuration in terms of average transcription accuracy even within the Interaction-aware optimization framework.

configuration	**loss mean**
A	0.563
B	0.592
C	0.573
D	0.574

Table 6.5: Mean composite loss values across Interaction-aware configurations.

The reported values in Table 6.5 reports the average of the composite loss function defined in Section 5.6.3, which integrates standard WER with event-specific suppression and production terms. As expected, configuration A achieves the lowest overall loss (0.563), followed by C (0.573) and D (0.574), while B again yields the highest value (0.592).

6.2.3 Quantitative Error Analysis

The comparison between WER-based and Interaction-aware optimization reveals that the introduction of an event-aware objective function does not lead to a systematic improvement in the representation of interactional events, at least at a form-level. Across all configurations, deletions remain the dominant error category, accounting for approximately half of all event-related cases under both optimization settings. This may indicate that the suppression of short interactional tokens is structural and cannot be mitigated by introducing a weak event-oriented signal in the objective function.

However, Interaction-aware optimization seems to modify the distribution of error types affecting such phenomena. This emerges especially when

config.	type	DEL (%)	SUB (%)	OK (%)	INS (%)
A	WER-based	53.39	19.78	21.11	5.73
A	Interaction-aware	54.67	20.44	21.39	3.50
B	WER-based	50.92	23.12	20.57	5.39
B	Interaction-aware	53.04	20.45	20.91	5.61
C	WER-based	51.72	25.86	18.06	4.36
C	Interaction-aware	50.89	25.13	19.67	4.32
D	WER-based	51.17	24.91	19.06	4.86
D	Interaction-aware	51.18	24.83	20.12	3.87

Table 6.6: Distribution of alignment operations for event tokens under WER-based and Interaction-aware optimization.

examining substitution and match rates. In Configuration B, Interaction-aware optimization is associated with a reduction in substitution errors (from 23.12% to 20.45%), which may suggest the tendency to normalize or replace interactional events with more canonical lexical forms. In Configuration C, substitution rates remain largely stable (from 25.86% to 25.13%). Increases in correct matches are instead observed across all configurations, although small. Configuration D shows a largely stable pattern across optimization strategies. Deletion and substitution rates remain virtually unchanged, while a modest increase in correct matches (from 19.06% to 20.12%) is accompanied by a slight reduction in insertion errors. Insertion errors remain consistently low under both optimization strategies: nevertheless, while they remain slightly reduced in A, C and D, a moderate increase is observed in B. This slight trend may suggest that rewarding correct event production can marginally increase the likelihood of producing interactional tokens that are not present in the reference. Such increases remain limited in magnitude and do not dominate the overall error profile, indicating that the Interaction-aware optimization pipeline does not trigger widespread hallucination of interactional events, favoring the introduction of a trade-off between suppression and insertion.

6.2.4 Subset B

Table 6.7 reports the aggregated results obtained on Subset B, which was not involved in the optimization process. This complementary analysis al-

align_type	WER (%)	DEL (%)	SUB (%)	OK (%)	INS (%)
WER-based	34.23	57.92	18.75	20.41	2.90
Interaction-aware	33.24	57.73	19.76	19.75	2.75

Table 6.7: Comparison summary of WER, deletions, substitutions, matches and insertions for the two types

lowed to verify whether the patterns observed in the tuning subset generalize to unseen data. In terms of global transcription accuracy, Interaction-aware optimization does not degrade performance. On the contrary, it achieves a slightly lower mean in WER (33.24%) compared to the WER-based pipeline (34.23%). Although the difference remains moderate, this result indicates that introducing event-sensitive components into the objective function does not negatively impact overall recognition accuracy on unseen data. Regarding the distribution of alignment operations for event tokens, deletions remain the dominant error category under both optimization strategies (57.92% vs. 57.73%), confirming the structural tendency of the ASR system to suppress interactional tokens. This mirrors the behavior observed in Subset A and suggests that event suppression is not merely an artifact of the optimization subset. However, some distributional shifts emerge. Interaction-aware optimization shows a slight increase in substitution rates (from 18.75% to 19.76%) and a marginal reduction in insertion errors (from 2.90% to 2.75%). The proportion of correctly matched events remains broadly comparable across the two systems (20.41% vs. 19.75%).

The results regarding Subset B therefore confirm the general pattern observed in Subset A: Interaction-aware optimization does not radically modify the overall error profile, nor does it introduce instability in global WER. Instead, its impact appears subtle and controlled, suggesting that the composite objective slightly reshapes the internal distribution of event-related errors without disrupting overall transcription quality.

6.2.5 Summary

The results indicate that modifying the decoding objective to incorporate interactional awareness does not fundamentally alter global transcription accuracy. Across configurations and subsets, WER remains broadly comparable between the WER-based and Interaction-aware pipelines. The introduction

of event-sensitive components does not produce substantial gains in overall error reduction, nor does it destabilize performance. Instead, its impact appears subtle: small shifts in the distribution of event-related errors can be observed, particularly in substitution and insertion patterns, however deletions remain the dominant issue across all settings. In practical terms, this means that while the Interaction-aware strategy does not harm transcription quality, it also does not radically improve the system's ability to preserve interactional phenomena. The structural tendency of the ASR system to suppress short conversational events appears to persist regardless of the optimization objective.

6.3 Gold Annotated Transcripts

6.3.1 Statistics

Table 6.8 shows the mean frequency per minute of interactional phenomena across different interaction types. Distributional differences emerge across the three conversational settings, reflecting varying degrees of spontaneity, interactional alignment and formality constraints.

type	**nr.**	**BC/min**	**FP/min**	**SR/min**	**OR/min**
free conversation	6	4.42	1.46	0.40	0.12
exams-office hours	8	4.01	4.76	0.83	0.11
semi-structured interview	12	5.70	3.25	0.26	0.00

Table 6.8: Frequency per minute of interactional phenomena across interaction types.

Backchannels (BC) occur frequently across all interaction types, with the highest rate observed in semi-structured interviews (5.70 events/min), followed by free conversation (4.42 events/min) and exams-office hours (4.01 events/min). This high occurrence suggests that backchanneling serves as a general interactional resource, consistently employed in all conversational settings.

Filled pauses (FP), instead, show marked variation across interaction types. They are relatively infrequent in free conversation (1.46 events/min), but substantially more frequent in academic contexts (4.76 events/min), with

intermediate values in semi-structured interviews (3.25 events/min). This distribution may be connected with increased cognitive load and planning pressure of students in oral exams and meetings with professors, where they are required to produce structured responses.

Self-repairs (SR) also display higher rates in exams and office hours (0.83 events/min) compared to free conversation (0.40 events/min) and semi-structured interviews (0.26 events/min). This suggests a greater tendency toward self-monitoring and correction in asymmetrical contexts, in which speakers have to be evaluated for their knowledge and their discussing abilities.

Other-repair phenomena (OR) show their highest frequency in free conversation (0.12 events/min), followed by exam/reception interactions (0.11 events/min) and being absent in semi-structured interviews. Although overall infrequent, the higher occurrence of OR in free conversation can be interpreted in relation to the looser sequential organization characteristic of this interactional setting. In contexts where topic shifts are frequent and turns are less tightly constrained, speakers may initiate repairs to re-establish coherence with previous turns, as argued by (Drew, 1997). The distribution of interactional phenomena at the speaker level can be visualized in Figures 6.3, 6.4 and 6.5, which display event frequencies normalized per minute across the three interaction types.

Regarding the academic context (Figure 6.3), in line with the aggregate data (4.01 BC/min), backchannels are produced by both professor and students, and display moderate dispersion. The distribution does not appear strongly polarized by role, indicating that both roles contribute to the maintenance of interactional alignment. Filled pauses, which showed the highest overall rate in this context (4.76 FP/min), display variability: higher and more dispersed values are observed among students, whereas professors tend to cluster at lower rates. This speaker-level asymmetry aligns with the interpretation commented above: in evaluative context, students may experience greater planning and cognitive load, which may result in more frequent hesitation phenomena. A similar pattern emerges for self-repairs, as they are more frequent in this type of interaction. The graph confirms that this increase is primarily driven by student speakers, whose distribution includes higher per-minute rates compared to professors. Again, the pattern is consistent with the increased need for precision and reformulation in asymmetrical and evaluation-oriented interactions. Other-initiated repair phenomena remain relatively infrequent, which is consisted with their low aggregate rate

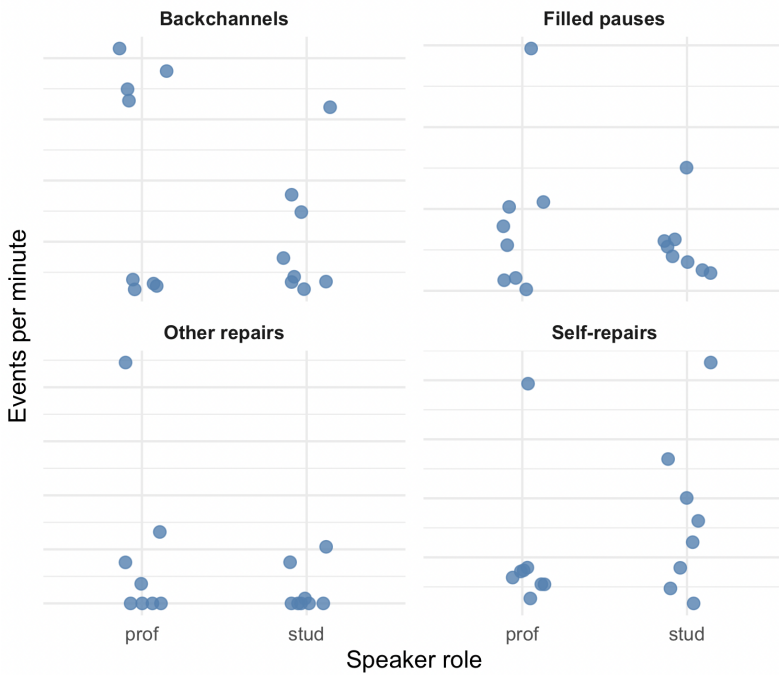

Figure 6.3: Speaker-level distribution of interactional phenomena, exams and office hours context

(0.11 OR/min). Their distribution appears limited and does not exhibit strong role-based polarization.

As far as semi-structured interviews are concerned, Figure 6.4 shows how interactional events are distributed across interviewer and interviewee roles. Backchannels are frequent in both roles and display relatively balanced distributions, with no strong asymmetry between interviewer and interviewee. This aligns with the aggregate data indicating that backchanneling is particularly prominent in this context for both roles (5.70 BC/min). The relatively even distribution suggests that semi-structured interviews foster continuous interactional alignment, with both participants actively contributing to feedback, agreement, assessments and incipient speakership. Filled pauses show moderate dispersion across speakers, however without the pronounced role asymmetry observed in exams and office hours. Although some variability is visible, the overall distribution appears more compact and less polarized. This is consistent with the data reported in Table 6.8, namely, 3.25 events/min. Self-repairs, which display the lowest aggregate rate across all

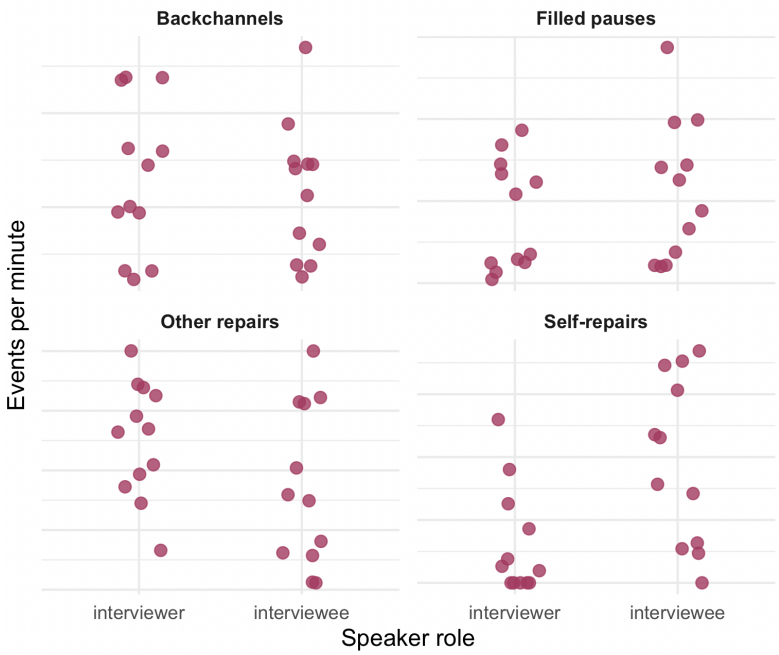

Figure 6.4: Speaker-level distribution of interactional phenomena, semi-structured interviews.

interaction types (0.26 SR/min), remain relatively infrequent at the speaker level as well. While some interviewees exhibit higher values, the overall distribution does not show the marked clustering observed in exams. Other-initiated repairs are present in both roles, although they do not display a strongly polarized distribution between interviewer and interviewee.

Regarding free conversation, unlike exams and semi-structured interviews this setting does not involve a role differentiation; consequently, the Figure 6.5 highlight inter-speaker variability rather than role-based differentiation. Backchannels show a relatively wide dispersion across speakers: several speakers display higher per-minute values, while others cluster at more moderate levels, suggesting that backchanneling in free conversation functions as an idiosyncratic behavior, as Blomsma et al. (2024) previously found. Filled pauses, which are substantially less frequent here than in exams (1.46 vs. 4.76 FP/min), remain comparatively low and more compactly distributed. Again, with the absence of evaluative pressure speakers experience less planning strain and therefore produce fewer hesitation phenomena. Self-repairs show moderate variability, however no clear trend emerges. Other-initiated

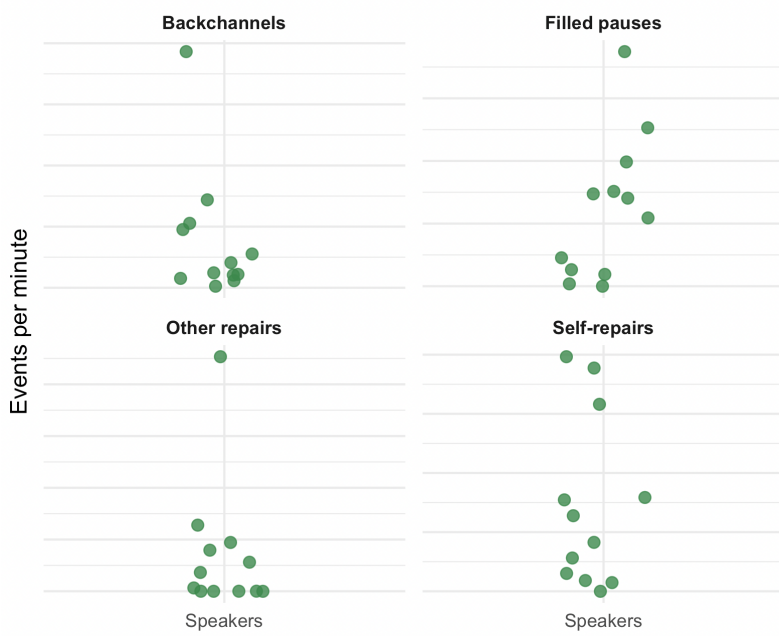

Figure 6.5: Speaker-level distribution of interactional phenomena, free conversations.

repairs, while overall infrequent (0.12 OR/min), are visibly present across multiple speakers. Unlike the other two types of conversation, this context allows greater flexibility and spontaneity both in topic management and sequential organization. As already commented for the aggregate values, this is consistent with Drew (1997): repair initiation in less constrained contexts may arise from the need to re-establish coherence across loosely connected turns.

6.3.2 Word Error Rate

Figure 6.6 presents the Word Error Rate (WER) obtained for each annotated conversation under the two decoding strategies: WER-based and Interaction-aware optimization. Conversations are grouped by interaction type, each represented by paired values, allowing a direct comparison of performance shifts within the same conversational context and recording.

Across all interaction types, the two optimization strategies display highly similar performance profiles. Within each panel, the curves closely mirror

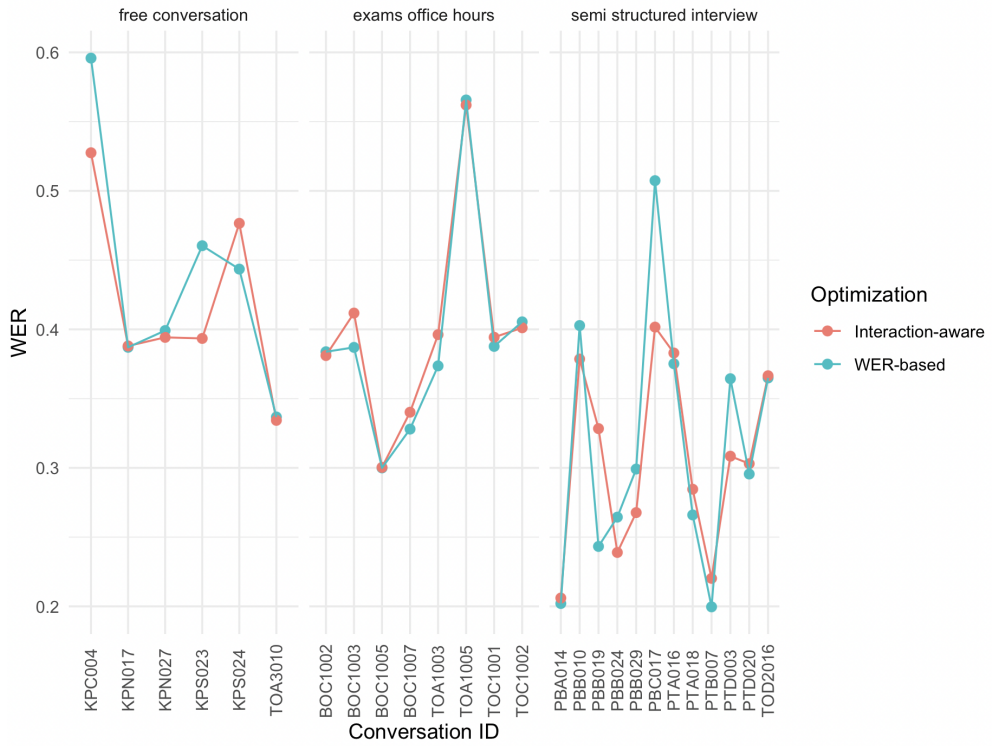

Figure 6.6: Comparison of conversation-level Word Error Rate (WER) between WER-based and Interaction-aware optimization strategies, grouped by interaction type. Each pair of points corresponds to the same conversation under the two decoding conditions.

each other, and the relative ranking of conversations in terms of transcription difficulty remains largely stable. Conversations that yield higher WER values under the WER-based configuration also tend to exhibit higher values under the Interaction-aware configuration (e.g., KPC004, TOA1005). This stability suggests that conversation-specific factors, such as acoustic quality, overlap density, or speaker variability, may represent the primary drivers of global error rates. When examined by interaction type, no systematic pattern emerges. In free conversations, minor fluctuations can be observed, yet the two curves remain closely aligned. A similar stability characterizes the exams/office hours subset, where local deviations occur but do not alter the overall distribution. In semi-structured interviews, although some conversations show slightly lower WER under the Interaction-aware configuration,

the magnitude of these differences remains limited and inconsistent across recordings. These shifts do not follow a clear interaction-type pattern and appear to be limited in magnitude: the two strategies appear to converge toward comparable levels of transcription accuracy, reinforcing the conclusion that decoding optimization exerts only incremental effects on overall error rates.

optimization	**mean WER**	**variance**	**SD**
WER-based	0.367	0.00958	0.0979
Interaction-aware	0.361	0.00716	0.0846

Table 6.9: Descriptive statistics summary of conversation-level WER on the annotated subset.

Table 6.9 reports the descriptive statistics of aggregate, conversation-level WER values under the two optimization strategies. On average, the Interaction-aware configuration yields a slightly lower mean WER (0.361) compared to the WER-based configuration (0.367). Although the absolute difference is small (approximately 0.006), it favors the Interaction-aware setting. In addition to the marginal reduction in mean WER, the Interaction-aware optimization also exhibits lower variance (0.00716 vs. 0.00958) and a smaller standard deviation (0.0846 vs. 0.0979). These tendencies are visually confirmed by Figure 6.7, which displays the distribution of WER values under the two conditions.

The medians are very close, with a slight shift toward lower values for the Interaction-aware configuration. The interquartile range appears marginally narrower, reflecting the reduced dispersion reported in Table 6.9. Importantly, the overall distributional structure remains comparable across the two strategies, and outliers are present in both cases. These results indicate that incorporating interaction-sensitive components into the optimization objective does not substantially alter global transcription accuracy. Rather, its impact appears incremental: it slightly lowers average WER and modestly stabilizes performance across conversations, without introducing systematic degradation. This aggregate pattern is consistent with the interaction-type graphs in Figure 6.6, where differences between strategies remain limited within each conversational setting.

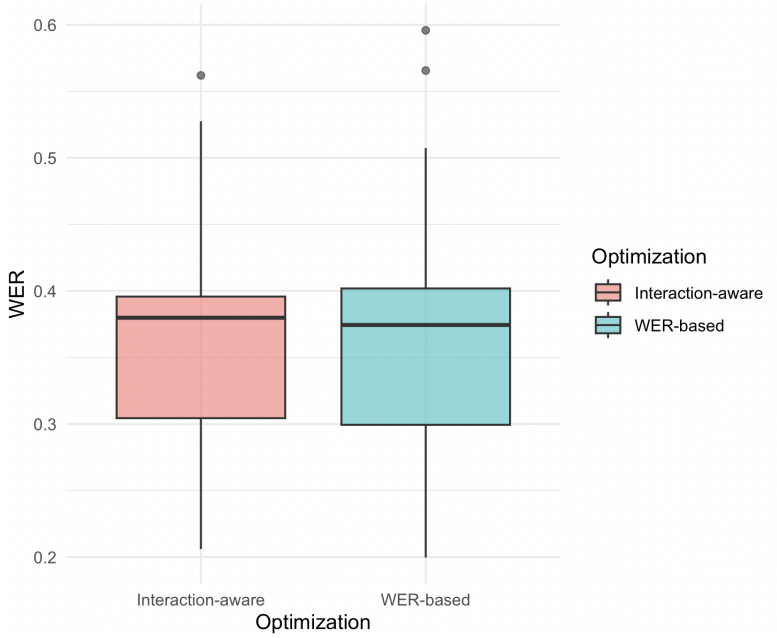

Figure 6.7: Visual distribution of conversation-level Word Error Rate (WER) under WER-based and Interaction-aware optimization, computed on the annotated subset.

6.3.3 Mean Match Ratio

6.3.3.1 Conversation-level

As far as the results at the conversation level are concerned, Figure 6.8 provides the distribution of the differences in mean match ratio between Interaction-aware and WER-based configurations (Δ = Event - WER). Across most phenomena, the distributions are centered around zero, indicating the absence of a systematic advantage for either optimization strategy.

For backchannels (BC), the median difference is effectively null and the interquartile range remains narrow, suggesting that the two configurations yield highly comparable performance across conversations. A similar pattern emerges for filled pauses (FP), where differences are generally small and symmetrically distributed around zero. Given the overall low match ratios observed for FP, the optimization strategy appears to exert only a marginal influence on their recognition. Other-initiated repairs (OR) show

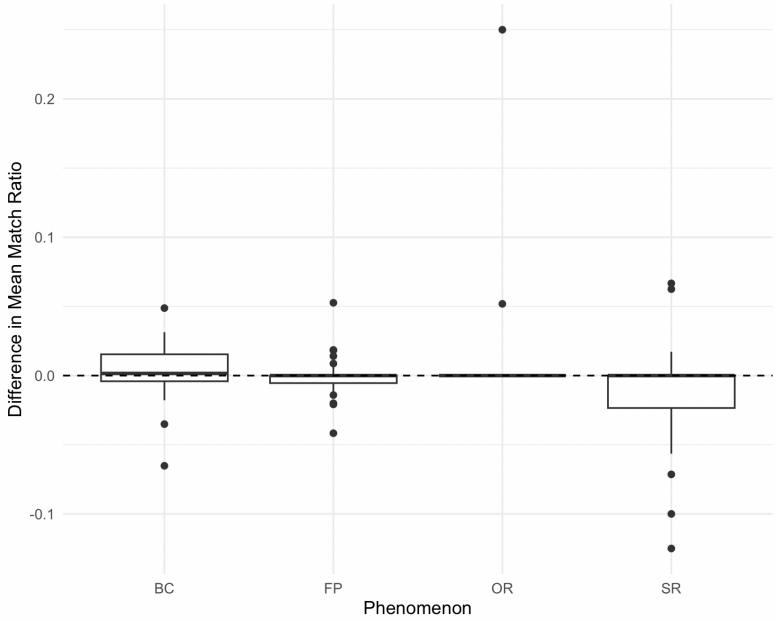

Figure 6.8: Distribution of differences in mean match ratio, for each phenomenon.

minimal variation between configurations, with most differences clustered at zero. The few visible outliers are likely attributable to the small number of OR events per conversation, which makes the metric more sensitive to minor fluctuations. Self-repairs (SR) exhibit comparatively greater variability. Although the median difference is slightly negative, indicating a modest tendency for the Interaction-aware configuration to perform marginally worse than the WER-based configuration in some conversations, the distribution remains relatively balanced, with both positive and negative deviations. No consistent or large-scale shift favoring one optimization strategy emerges: the observed variability appears to be driven more by conversation-specific characteristics than by the choice of optimization strategy itself.

Although the boxplots suggested some variability across conversations, the Wilcoxon signed-rank tests revealed no statistically significant differences between the two configurations for any of the investigated phenomena. As shown in Table 6.10, all phenomena $p > 0.05$. This indicates that the median difference in mean match ratios between the two optimization strategies does not significantly deviate from zero: despite small conversation-level fluctu-

ations, neither configuration demonstrates a systematic advantage over the other in the recognition of annotated interactional phenomena. Given the relatively low number of OR events per conversation, the Wilcoxon test for this phenomenon should be interpreted with caution, as limited sample size may reduce statistical power.

Table 6.10: Wilcoxon signed-rank test (paired)

phenomenon	**p-value**
BC	0.207
FP	0.780
SR	0.197
OR	0.371

6.3.3.2 Interaction-level

Results of the mean match ratio calculation at the interactional level (Table 6.11 reveal clear differences across interactional phenomena, whereas differences between optimization strategies remain generally limited.

Across all interaction types, self-repairs (SR) display the highest mean match ratios. In semi-structured interviews, SR reach 56.11% under the WER-based configuration and 53.33% under the Interaction-aware configuration; in free conversation, the values are 51.80% and 52.49% respectively; and in exams/office hours interactions they range between 45.12% and 46.64%. This suggests that repair sequences, despite their structural complexity, tend to be relatively well preserved. In contrast, filled pauses (FP) consistently show the lowest match ratios across all interaction types. Values remain below 5% in semi-structured interviews and free conversations, and drop to approximately 1–1.5 % in institutional settings. The fact that filled pauses exhibit extremely low match ratios in all interaction types indicates that hesitation markers are particularly prone to be missed by Whisper. The Interaction-aware optimization does not substantially alter this pattern. Backchannels (BC) occupy an intermediate position, with match ratios ranging from approximately 17% to 23%. Semi-structured interviews exhibit slightly higher values under the Interaction-aware configuration (23.01% vs.

21.71%), whereas differences in free conversation and institutional interactions are minimal. Overall, the optimization strategy appears to have only a marginal effect on backchannel recognition. Finally, other-initiated repairs (OR) show relatively high match ratios in free conversation and institutional interactions (approximately 46–58%), though the number of events is limited. Given the small sample size, these values should be interpreted with caution.

interaction type	**phen.**	n	**WER (%)**	**Event (%)**
free conversation	BC	524	16.92	16.89
	FP	174	3.45	4.02
	SR	48	51.80	52.49
	OR	13	46.15	53.85
exams/office hours	BC	636	22.10	21.53
	FP	761	1.38	1.45
	SR	133	46.64	45.12
	OR	18	55.56	58.15
semi-structured interview	BC	1354	21.71	23.01
	FP	776	4.23	4.17
	SR	62	56.11	53.33
	OR	0	–	–

Table 6.11: Mean match ratio (in percentage) by interaction type and optimization strategy.

When comparing optimization strategies directly, differences are generally modest. In some cases, the Interaction-aware configuration yields slightly higher match ratios (e.g., BC in semi-structured interviews and OR in institutional interactions), while in others the WER-based configuration performs marginally better (e.g., SR in semi-structured interviews). Generally, self-repairs appear to be comparatively robust, filled pauses remain highly challenging, and backchannels show moderate preservation across settings. However, no systematic advantage of one strategy over the other emerges across all interaction types and phenomena: this is consistent with the conversation-level statistical analysis.

6.3.4 Substitution Rates

Table 6.12 reports the global omission rates after normalization. Across both decoding strategies, omissions account for more that 82% of total errors. This indicates that the predominant source of error is the complete absence of interactional phenomena. The two optimization strategies show virtually identical omission rates, suggesting that optimization does not substantially affect the likelihood of phenomena being omitted.

optimization	total errors	omissions	omission rate (%)
WER-based	5283	4340	82.2
Interaction-aware	5266	4336	82.3

Table 6.12: Global omission rate after normalization.

Table 6.13 illustrates the most frequent substitution patterns after normalization. Once orthographic variation and consequent normalization have been accounted for (see Section 5.7.4), substitution frequencies remain relatively stable across configurations.

WER-based			Interaction-aware		
gold	**whisper**	**frequency**	**gold**	**whisper**	**frequency**
mh	grazie	26	mh	grazie	28
eh	è	17	eh	è	17
ehm	e	16	mh	e	13
mh	e	15	ehm	e	12
mh	non	13	mh	non	11
mh	è	12	mh	che	10
sì	grazie	11	sì	grazie	10
eh	e	10	mh	è	9
mh	che	10	okay	e	9
eh	grazie	9	eh	che	8

Table 6.13: Top 10 substitution patterns after normalization.

Several patterns are noteworthy. First, nasal backchannels such as *mh* are frequently rendered as lexical items including *grazie*, *non*, *che*, or *è*. This suggests that short non-lexical vocalizations may be interpreted by the ASR system as plausible lexical continuations within the surrounding syntactic

context. In particular, the recurrent mapping of *mh* to *grazie* may indicate a tendency toward lexical normalization of brief feedback signals into more canonical verbal responses.

Second, substitutions of the type *eh* → *è* are among the most frequent substitutions. As already discussed in Section 5.7.4, although the two forms are phonetically similar, no consistent restriction to a single interactional representation between filled pause and backchannel was observed. By contrast, instances of *eh* → *e* were treated as equivalent when annotated as filled pauses. The remaining occurrences in the WER-based output correspond to backchannel uses of *eh*. Interestingly, this substitution does not appear among the ten most frequent patterns in the Interaction-aware configuration, suggesting a slight redistribution of substitution types under the alternative optimization strategy.

Overall, substitution patterns are consistent across the two optimization strategies, further reinforcing the conclusion that differences between configurations are limited. The dominant error profile remains characterized by high omission rates and a smaller set of recurrent substitutions primarily involving non-lexical vocalizations.

6.3.5 Overlap Analysis

6.3.5.1 Global Test

overlap	Obs 0	Obs 1	Exp 0	Exp 1	match rate
no overlap	8727	2817	8840.86	2703.13	24.4%
overlap	2030	472	1916.13	585.86	18.9%

Table 6.14: Observed (Obs) and expected (Exp) frequencies for overlap per recognition outcome (all phenomena, both configurations combined). Match rate is also included.

Table 6.14 presents the distribution of recognition outcomes as a function of overlap. Tokens produced in non-overlapping contexts show a match rate of 24.4%, whereas tokens occurring in overlap display a lower match rate of 18.9%. Expected frequencies (also reported in Table 6.14) are well above the minimum threshold in all cells, confirming that the assumptions of the chi-square test are satisfied at the global level. The test confirms that this

difference is statistically significant ($\chi^2(1) = 34.85$, $p < 0.001$), indicating that overlap is associated with reduced recognition accuracy.

6.3.5.2 Interaction-specific Test

When examining interactional phenomena separately, the effect of overlap is found to be phenomenon-specific. Table 6.15 reports the observed and expected frequencies for each phenomenon.

phenomenon	**overlap**	**Obs 0**	**Obs 1**	**Exp 0**	**Exp 1**
BC	no overlap	4146	1336	4235.55	1246.45
BC	overlap	1539	337	1449.45	426.55
FP	no overlap	3248	98	3251.61	94.39
FP	overlap	369	7	365.39	10.61
OR	no overlap	66	76	67.16	74.84
OR	overlap	4	2	2.84	3.16
SR	no overlap	1267	1307	1265.08	1308.92
SR	overlap	118	126	119.92	124.08

Table 6.15: Observed (Obs) and expected (Exp) frequencies for overlap × recognition outcome by phenomenon.

phenomenon	**test**	χ^2	p**-value**
BC	Chi-square	32.30	< 0.001
FP	Chi-square	1.04	0.307
OR	Fisher	–	0.422
SR	Chi-square	0.04	0.849

Table 6.16: Association between overlap and recognition outcome by interactional phenomenon (adjusted across six follow-up tests).

As illustrated in Table 6.16, a significant association between overlap and recognition outcome emerges only for backchannels (BC) ($\chi^2 = 32.30$, $p < 0.001$). In contrast, no significant effects are observed for filled pauses (FP) ($p = 0.307$) or self-repairs (SR) ($p = 0.849$). For other-initiated repairs (OR), due to low expected frequencies, Fisher's exact test was computed; the association between overlap and recognition, however, remains

non-significant ($p = 0.422$). These results confirm that the global effect of overlap is largely driven by backchannel tokens. The implementation of the Holm correction procedure (Table 6.17) did not alter the behavior of the results: only backchannels remain statistically significant, whereas all other phenomena remain non-significant.

Phenomenon	p **(Holm)**
BC	7.92×10^{-8}
FP	.921
OR	.921
SR	.921

Table 6.17: Holm-adjusted p-values for the overlap effect by interactional phenomenon.

6.3.5.3 Optimization Test

Regarding optimization tests, as shown in Table 6.18, expected frequencies remain well above the minimum threshold for all cells in both optimization conditions. In particular, no expected count approaches the critical value of 5, thereby confirming that the assumptions underlying the chi-square approximation are fully satisfied. This ensures that the inferential results obtained from the Pearson chi-square statistics can be interpreted without concerns related to sparse data or distributional distortions.

optimization	**overlap**	**Obs 0**	**Obs 1**	**Exp 0**	**Exp 1**
Event-based	no overlap	4358	1414	4414.27	1357.73
Event-based	overlap	1013	238	956.73	294.27
WER-based	no overlap	4369	1403	4426.60	1345.40
WER-based	overlap	1017	234	959.40	291.60

Table 6.18: Observed (Obs) and expected (Exp) frequencies for overlap per recognition outcome by optimization strategy.

The association between overlap and recognition outcome was found to be statistically significant in both optimization configurations (Table 6.19. Separate chi-square tests reveal significant effects for the WER-based configuration ($\chi^2 = 17.7$, $p < 0.001$) and for the Interaction-aware configuration

($\chi^2 = 16.8$, $p < 0.001$). Although the magnitude of the chi-square statistics slightly differs, the direction of the effect is consistent: tokens produced in overlapping speech contexts show lower recognition rates compared to tokens produced in non-overlapping contexts, regardless of the decoding objective adopted.

Optimization	χ^2	p**-value**
WER-based	17.7	< 0.001
Interaction-aware	16.8	< 0.001

Table 6.19: Chi-square tests of overlap effect by optimization strategy.

As shown in Table 6.20, after correction the overlap effect remained statistically significant for both the WER-based ($p_{\text{Holm}} = 1.27 \times 10^{-4}$) and the Interaction-aware configuration ($p_{\text{Holm}} = 1.65 \times 10^{-4}$). This procedure therefore does not alter the inferential pattern previously observed. This confirms that the negative impact of overlap on recognition outcomes is robust and stable across model configurations.

Optimization	p **(Holm)**
Interaction-aware	1.65×10^{-4}
WER-based	1.27×10^{-4}

Table 6.20: Holm-adjusted p-values for the overlap effect by optimization strategy (adjusted across six follow-up tests).

6.4 Qualitative Error Analysis

6.4.1 Filled Pauses

The qualitative inspection of correctly recognized filled pauses (FP) must be interpreted in light of the very low mean match ratios observed quantitatively. Across interaction types and optimization strategies, FP consistently display the lowest performance, with values ranging between approximately 1% and 4% (see Section 6.3.3.2). This very small proportion of annotated hesitation markers preserved in the ASR output aligns with the qualitative analysis. Sure enough, correctly recognized cases are sparse and tend

to share specific characteristics. For instance, in PTD020 (Figure 6.9), the token *eh* is correctly matched in both configurations when it occurs at the onset of an intonational unit and precedes propositional content. In contrast, a subsequent nasal hesitation (*mh*) in utterance-final position is not recognized. Comparable *eh-eh* matches are observed in PTB007 (Figure 6.10) and PBC017 (Figure 6.11), again in contexts where the filled pause is prosodically isolated and relatively salient.

```
TOR001      | [cosa fai durante il tuo tempo libero?           ]
TOI008      |
event       |


TOR001      |
TOI008      |                                    [e::h è una b..→
event       |                                    [FP      ]


TOR001      |
TOI008      |bella domanda               ]
event       |


TOR001      |
TOI008      |
event       |


TOR001      |
TOI008      |diciamo che::: mh                             ]
event       |          [FP                                 ]
```

Figure 6.9: Example of a recognized FP from PTD020 (ParlaTO).

```
TOR004      |
TOI021      |e quindi lei si è esaurita un pochino                →
event       |

TOR004      |
TOI021      |e quindi lei si è esaur..]          [e::h            →
event       |                                   [    FP             →

TOR004      |
TOI021      |e::h                            ]
event       | FP                             ]

TOR004      |                           [e quindi hanno cambiato ..→
TOI021      |
event       |

TOR004      |tante case a ven..→
TOI021      |
event       |
```

Figure 6.10: Example of a recognized FP from PTB007 (ParlaTO).

```
BOI042      |[e poi da LÌ:
event       |
BOR009      |

BOI042      |         [e::h a ventun anni: venne fuori sto concor..→
event       |         [FP     ]
BOR009      |

BOI042      |concorso in azienda trasporti dico >porc(o) boia< pe..→
event       |
BOR009      |

BOI042      |per me era un sogno                                    →
event       |
BOR009      |
```

Figure 6.11: Example of a recognized FP from PBC017 (ParlaBO).

In BOC1005 (Figure 6.12), the form *ehm* is correctly transcribed as *ehm* in both configurations. It is important to underline that, as illustrated by the timestamps, this filled pause is segmentally isolated and acoustically well-defined. This suggests that longer vocalic hesitation forms may not be inherently problematic for the ASR system in this specific context. However, such successful recognitions remain numerically limited when considered against the total number of annotated filled pauses.

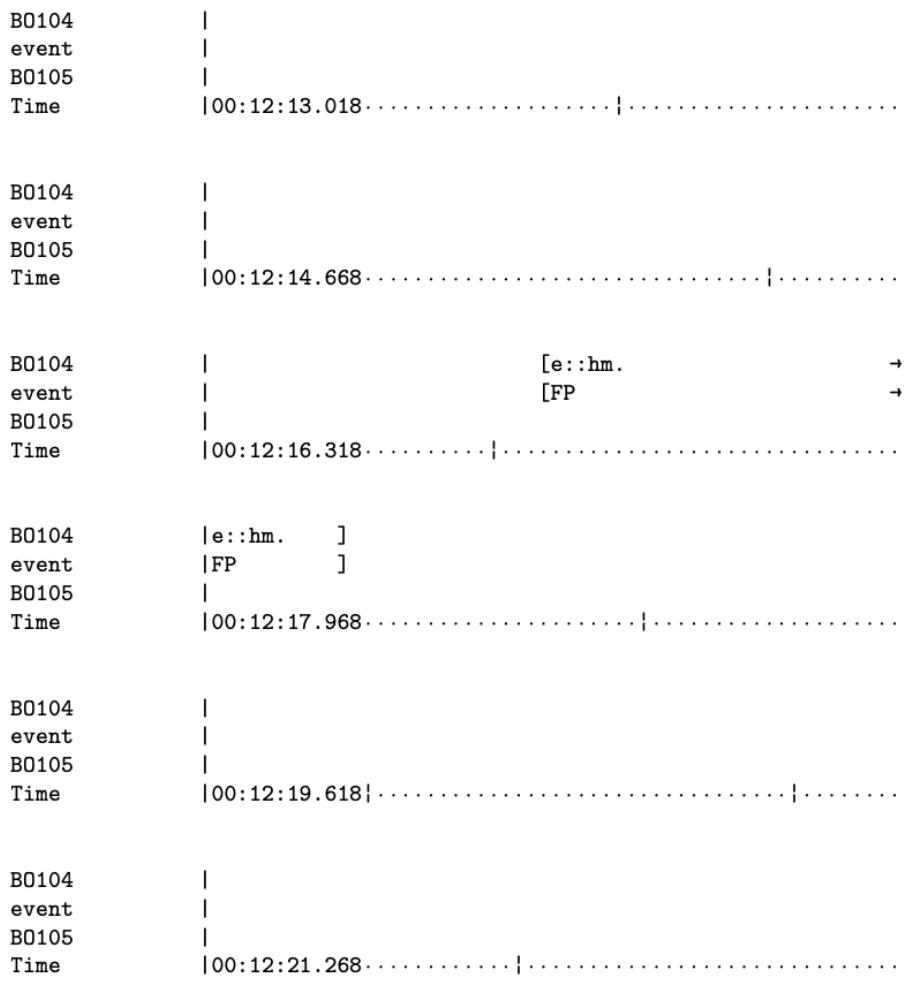

Figure 6.12: Example of a recognized FP from BOC1005 (KIP).

Only one instance of a nasal filled pause is matched through normalization

(*mh* recognized as *mmm* in KPS024, 6.13). This near absence of matched nasal forms coherently reflects the extremely low FP mean match ratios reported in Table 6.11. Nasal hesitations, maybe for being short and low in intensity, appear particularly vulnerable to omission.

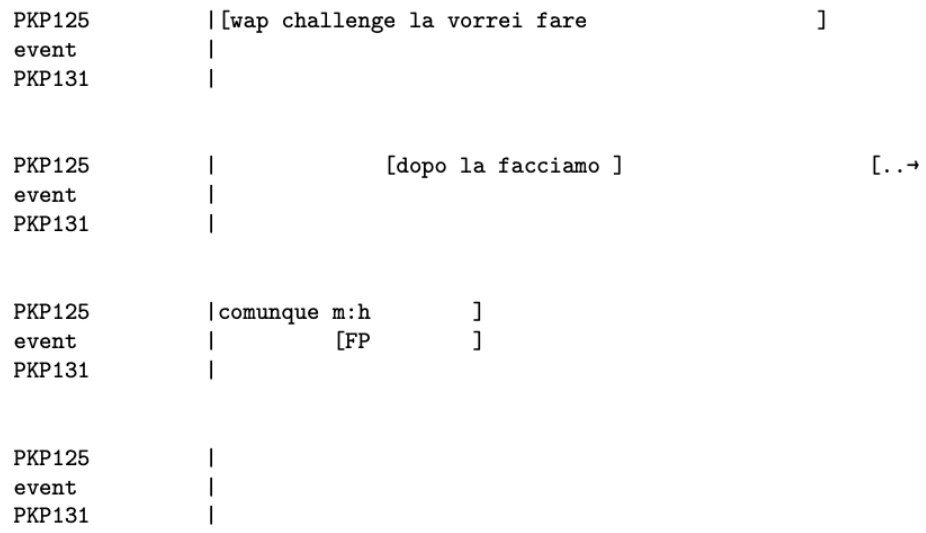

Figure 6.13: Example of a recognized FP from KPS024 (KIPasti).

The few asymmetries between decoding strategies (e.g., PTA016, matched only in the WER-based configuration, and BOC1003, matched only in the Interaction-aware configuration) do not suggest a systematic advantage of one optimization strategy over the other. Rather, they reinforce the broader quantitative finding: differences between configurations are marginal, whereas the phenomenon itself constitutes the primary source of difficulty.

6.4.2 Backchannels

Figures 6.14, 6.15 and 6.16 present backchannels that were successfully recognized by Whisper. A common feature of these cases is their sequential positioning: they are not produced in overlap, but occur in the transition space between turns or at the clear onset of a new utterance.

In Figure 6.14 the backchannel (*eh no ti capisco*) is produced immediately after the prior speaker's completion, which makes it acoustically and

sequentially salient. Similarly, in Figure 6.15 the token *eh* appears at the beginning of the speaker's turn of the speaker's contribution (*eh no infa–però*), representing an example of incipient speakership.

```
TOR001          |
TOI008          |ma è un'avversione per i mezzi pubblici
event           |


TOR001          | [eh no ti capisco                    ]
TOI008          |
event           | [BC                                  ]
```

Figure 6.14: Example of a recognized BC in PTD020 (ParlaTO).

```
TO086           |
event           |
TO085           |[vabbè però sti cazzi dai::                          →


TO086           |             [eh no infa~ però::                     →
event           |             [BC         ]
TO085           |vabbè però..]


TO086           |eh no infa~ però::    ]
event           |
TO085           |                       [ma sì ma è in inghilter~ ((r..→


TO086           |
event           |
TO085           |((ride))                    ]
```

Figure 6.15: Example of a recognized BC from TOA3010 (KIP).

In Figure 6.16 the nasal vocalization (*mh*) is also produced immediately after the prior turn. Again, although it is short and interactionally lightweight, it is not masked by overlapping speech and is therefore successfully transcribed.

Given these first three examples, it could be affirmed that Whisper is more likely to preserve backchannels when they are prosodically integrated

```
PKP125        |[a me hanno regalato: m::h per il mio compleanno:     →
PKP130        |
event         |


PKP125        |a me hanno regalato: m::h per il mio ..]
PKP130        |                                          [m:h libri? →
event         |                                          [BC    ]


PKP125        |
PKP130        |m:h libri?        ]
event         |
```

Figure 6.16: Example of a recognized BC from KPS023 (KIPasti).

into a turn-initial position or when they occupy a recognizable transition space. Some cases of backchannel matches in overlapping contexts can also be found. In Figure 6.17, the backchannel *eh immagino [...]* is pronounced while the other speaker is still producing their utterance. Similarly, in Figure 6.18, the backchannel *eh vabbè* is overlapping with PKP125's ongoing turn.

```
TOR001        |mh mh           ]                          [[eh im..→
event         |BC              ]                          [BC     →
TOI008        |                 [anche quelle [son bellissime, prop..→


TOR001        |immagino, non son mai andata]         ]
event         |BC               ]
TOI008        |proprio belle]                      ]
```

Figure 6.17: Example of a recognized overlapping BC from PTD020 (ParlaTO).

Cases of backchannel alignment can also be observed. In Figure 6.19, two alignments occur in close succession. The sequence *bibliografia sì*, produced by BO089, functions as an aligning response to BO087's prior turn. However, it is uttered in overlap with *bibliografia* and is not preserved in the ASR output. By contrast, the subsequent token *magistrale*, also produced by BO089, is recognized and correctly transcribed. In this case, the alignment token is produced at the beginning of a new utterance, following the prior speaker's turn, and not in direct overlap. Although brief, it occupies a clearer sequential position and is integrated into a structurally well-defined turn.

```
PKP125          |
PKP131          |(#damme na foto a vede'?    ]
event           |


PKP125          |
PKP131          |
event           |


PKP125          |   [e:h ce l'ho sul telefono che al momento non pos[..→
PKP131          |
event           |


PKP125          |pos[so °(prendere)°]
PKP131          |    [[e:h va]bbè dopo     ]
event           |    [BC     ]
```

Figure 6.18: Example of a recognized overlapping BC from KPS024 (KIPasti).

```
BO087           |[[<bibliografia>],                    ]      [magist..→
BO089           |[[bibliografia (.) sì].                ]
event           |[BC                                   ]


BO087           |rale.                ]
BO089           |                       [magistrale (.) ho portato an..→
event           |                       [BC                ]


BO087           |
BO089           |anche::: [il reperto..→
event           |
```

Figure 6.19: Example of a recognized BC alignment from BOC1002 (KIP).

6.4.3 Self-repairs

Figures 6.20 and 6.21 present the distribution of self-repair event integrity across the two decoding strategies. In both configurations, the vast majority of self-repairs are only partially preserved. Fully preserved repairs represent a very small proportion of the total, while completely omitted repairs account for a limited but non-negligible share. The overall distribution is strikingly similar across the two optimization strategies. In both cases, par-

tial preservation clearly dominates, suggesting that the model tends to retain the reformulated segment of the repair while deleting or compressing the disfluent material. Differences between the WER-based and Interaction-aware configurations are minimal, indicating that optimization strategy does not substantially alter the structural treatment of repair sequences. This pattern points to a systematic behavior of the ASR model: rather than faithfully reproducing the full repair trajectory (typically composed of a repairable + hesitation marker + repair solution), Whisper frequently outputs a linearized version of the utterance, privileging lexical and grammatical continuity over interactional structure. This can be further observed in the examples provided below.

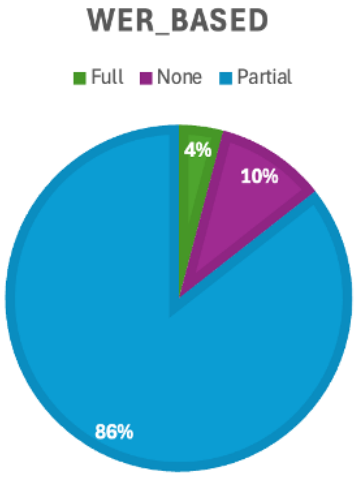

Figure 6.20: Distribution of self-repair events in the WER-based configuration, classified according to the event integrity (Full, Partial, None).

Table 6.21 illustrates a partially preserved self-repair. The repairable (*delle opere*) in the ASR output is directly condensed as *delle traduzioni*, which corresponds to the repaired element. Both the repairable and the truncated form (*pubbli-*) are omitted: this clearly shows the linearization and normalization behavior typical of ASR systems, that in this case produced a structurally reduced output. Another example of linearization is shown in Table 6.22, where Whisper deleted both the initial verb (*possono*), the hesitation marker (*eh*), and the truncated onset (*am-*), while directly retaining the final repair solution (*possono agire*). Whisper reconstructed a fluent clause by eliminating the repairable and the associated filled pause.

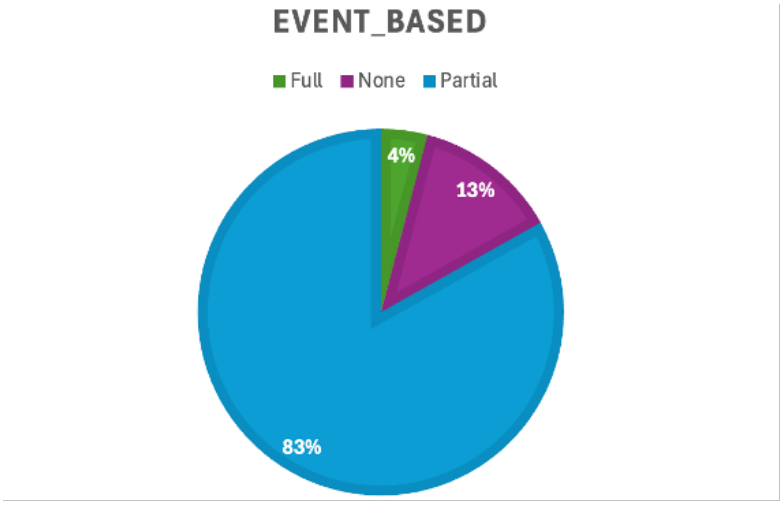

Figure 6.21: Distribution of self-repair events in the Interaction-aware configuration, classified according to the event integrity (Full, Partial, None).

The resulting transcript is syntactically coherent but discursively simplified, as the incremental nature of the production process is erased. A similar pattern emerges in the extract from Table 6.23. The entire first attempt (*che si è contratti*) was removed, while the corrected form (*che si è contratto*) was preserved. In this case, it seems that Whisper privileges grammatical agreement over the faithful reproduction of the repair sequence. The disfluency structure is suppressed, and only the resolved formulation survives in the output.

speaker_og	**speaker_whi**	**gold**	**whisper**	**annotation**
BO089	SPEAKER_B	delle	–	SR[89]
BO089	SPEAKER_B	opere	–	SR[89]
BO089	–	pubbli-	–	SR[89]
BO089	–	[PAUSE]	–	SR[89]
BO089	–	delle	delle	SR[89]
BO089	–	traduzioni	traduzioni	SR[89]
BO089	SPEAKER_B	pubblicate	pubblicate	SR[89]

Table 6.21: Self-repair sequence in BOC1002 (KIP).

In Table 6.24 the truncated numeral *(millenovec-)* and the preceding

speaker_og	speaker_whi	gold	whisper	annotation
TO007	–	possono	–	SR[32]
TO007	–	eh	–	FP[32]
TO007	–	am-	–	SR[32]
TO007	SPEAKER_A	possono	possono	SR[32]
TO007	SPEAKER_A	agire	agire	SR[32]

Table 6.22: Self-repair sequence in TOC1002 (KIP).

speaker_og	speaker_whi	gold	whisper	annotation
TO003	–	che	–	SR[31]
TO003	–	si	–	SR[31]
TO003	–	è	–	SR[31]
TO003	–	contratti	–	SR[31]
TO003	–	[PAUSE]	–	SR[31]
TO003	SPEAKER_A	che	che	SR[31]
TO003	SPEAKER_A	si	si	SR[31]
TO003	SPEAKER_A	è	è	SR[31]
TO003	SPEAKER_A	contratto	contratto	SR[31]

Table 6.23: Self-repair sequence in TOC1001 (KIP).

preposition (*dal*) are omitted, whereas the reformulated and phonologically complete form (*nel millenovecento...*) is retained. This, once again, reflects a strategy of selective preservation of the repair solution, particularly when the reformulated segment constitutes a correct lexical unit. The truncated element, lacking acoustic stability and lexical completeness, is discarded.

speaker_og	speaker_whi	gold	whisper	ann.
TOI021	–	dal	–	SR[16]
TOI021	–	millenovec-	–	SR[16]
TOI021	SPEAKER_A	nel	nel	SR[16]
TOI021	SPEAKER_A	millenovecento[...]	millenovecento[...]	SR[16]

Table 6.24: Self-repair sequence in PTB007 (ParlaTO).

Compared to the previous examples, the extract illustrated in Table 6.25 is slightly more complex from a structural point of view. Whisper partially reorganizes the sequence: while some tokens are preserved (*non*, *mai*, *pesato*),

others are substituted (*ho* → *mi*), inserted (*è*), or omitted. The second repetition of the repaired clause, with the correct verb, is largely deleted. Whisper therefore appears to reconstruct a grammatically coherent linear sequence, effectively collapsing the repair into a single clause. This confirms the tendency of ASR system towards normalization, favoring propositional coherence over incremental repair structure.

speaker_og	speaker_whi	gold	whisper	annotation
BOI101	SPEAKER_A	non	–	SR[25]
BOI101	SPEAKER_A	ho	–	SR[25]
BOI101	SPEAKER_A	mai	–	SR[25]
BOI101	SPEAKER_A	pesato	–	SR[25]
BOI101	SPEAKER_A	non	non	SR[25]
BOI101	SPEAKER_A	mi	mi	SR[25]
BOI101	SPEAKER_A	è	è	SR[25]
BOI101	SPEAKER_A	mai	mai	SR[25]
BOI101	SPEAKER_A	pesato	pesato	SR[25]

Table 6.25: Self-repair sequence in PBB029 (ParlaBO).

speaker_og	speaker_whi	gold	whisper	annotation
BOI076	SPEAKER_B	abito	abito	SR[24]
BOI076	SPEAKER_B	praticamente	praticamente	SR[24]
BOI076	–	[PAUSE]	-	SR[24]
BOI076	SPEAKER_B	vicino	vicino	SR[24]
BOI076	SPEAKER_B	al	al	SR[24]
BOI076	SPEAKER_B	centro	centro	SR[24]
BOI076	SPEAKER_B	cioè	cioè	SR[24]
BOI076	SPEAKER_B	abito	abito	SR[24]
BOI076	SPEAKER_B	scusa	scusa	SR[24]
BOI076	–	[PAUSE]	-	SR[24]
BOI076	SPEAKER_B	lavoro	lavoro	SR[24]
BOI076	SPEAKER_B	vicino	vicino	SR[24]
BOI076	SPEAKER_B	al	al	SR[24]
BOI076	SPEAKER_B	centro	centro	SR[24]

Table 6.26: Self-repair sequence in PBA024 (ParlaBO).

Unlike the previous examples, the repair sequence in Table 6.26 is entirely preserved in the ASR output. Both the initial formulation (*abito praticamente vicino al centro*) and the reformulated clause (*lavoro vicino al centro*), together with the repair markers (*cioè, scusa*), are retained. A possible explanation may lie in the structural and prosodic properties of the sequence. The repair unfolds fluently, without abrupt truncations or phonologically incomplete forms, and does not involve grammatical inconsistencies that would require restructuring (as observed in Table 6.26). The transition from the first formulation to the corrected one is sequentially clear and lexically well-formed. This suggests that when self-repairs are produced in a relatively fluent and structurally coherent manner, they are more likely to be preserved as interactional phenomena rather than reduced to a linearized, 'cleaned' version.

6.4.4 Other-initiated Repair

In Figure 6.22, the other-initiated repair is first realized through the question *che cosa vuol dire?*, produced by BO104 in overlap with BO105's ongoing turn. The repair initiation occurs twice in close succession: first as *[che cosa] vuol dire?*, and then again as *[sì che cosa] vuol dire*, reinforcing the request for clarification. In the ASR output, however, only the first occurrence of the repair initiation is retained, while the second is omitted. This selective preservation may indicate that Whisper tends to avoid reproducing closely repeated repair initiators (or, more generally, questions), especially when they occur in rapid succession and in overlap. The second token sequence, although interactionally meaningful as an intensified or insistently reformulated repair initiation, is likely treated as redundant or disfluent material by the system. From an interactional perspective, the repetition strengthens the repair trajectory and signals persistent trouble in understanding. In the ASR transcript, by contrast, the deletion of the second initiation results in a structurally simplified sequence, where the persistence of the repair action is attenuated.

In this extract provided in Table 6.23, the other-initiated repair is realized through the clarification request *ma chi?*, produced by TO085 in overlap with TO086's ongoing turn. The repair initiation targets the referential ambiguity of *lui*, prompting TO086 to specify the intended referent (*Luca*). The sequence is structurally clear: a referential problem is identified and immedi-

```
BO104          |
event          |
BO105          |[validità significa che la prova valuta realmente, (..→


BO104          |
event          |
BO105          |(.) cioè che si intende valutare.                    →


BO104          |
event          |
BO105          |validità significa che la prova v..]        [una def..→


BO104          |     [[che cosa] vuol dire?                         ]
event          |     [OR                                            ]
BO105          |defini[zione u..]


BO104          |
event          |
BO105          |l'ho trovata anche una definizione un po' particola~..→


BO104          |                   [[sì che cosa] vuol dire,          ]
event          |                   [OR                                ]
BO105          |[nel senso che]                               ]
```

Figure 6.22: Example of two other-initiated repairs in BOC1005 (KIP).

ately resolved through specification. In the ASR output, however, both the repair initiation (*ma chi?*) and the subsequent specification (*Luca*) are omitted: the entire repair sequence therefore disappears from the transcript. The referential ambiguity introduced by *lui* remains unresolved, and the interactional work performed to restore clarity is no longer visible. This case illustrates a more radical form of suppression compared to partial omissions. The ASR system directly removes the repair trajectory altogether. The deletion of both the initiator and the repair solution suggests that short, overlapped, and low-intensity utterances are particularly vulnerable to omission. From an interactional perspective, the negotiation of reference, that is central to conversational organization, is erased.

In the excerpt illustrated in Figure 6.24, the token *mh?* produced by PKP115 functions as a minimal response and repair-relevant signal following PKP116's question (*vuoi un po'?*). Although short, the non-lexical initiator signals uncertainty or difficulty in processing the prior turn, potentially

```
TO086    |[perché lui in genere gli aneddoti non li butta a ca..→
TO085    |
event    |


TO086    |caso li dice::                                        →
TO085    |
event    |


TO086    |per..]     [[li le~]       ]                  [luca→
TO085    |           [[ma chi]?      ]
event    |           [OR             ]


TO086    |luca                ]
TO085    |
event    |
```

Figure 6.23: Example of an other-initiated repair mismatched in TOA3010 (KIP).

because of the sudden interruption of what PKP115 was recounting (Drew, 1997). In the ASR output, however, *mh?* is lost. This omission is consistent with the broader pattern observed for *mh*, as a filled pause or backchannel. Compared to vocalic hesitations such as *eh* nasal tokens like *mh* are typically shorter, lower in intensity, and less segmentally defined. These acoustic properties likely make them particularly vulnerable to suppression during decoding.

```
PKP116   |
PKP115   |ma (diciamo) che la mamma di sara è come la marty     →
event    |


PKP116   |                   [[vuoi un..]
PKP115   |ma (diciamo) che ..][[c(io)è ..]            [mh?    ]
event    |                                            [OR     ]


PKP116   |         [ne vuoi un po'?]
PKP115   |
event    |
```

Figure 6.24: Example of an other-initiated repair mismatched in KPN027 (KIPasti).

Figure 6.25, finally, shows an example of a matched other-initiated repair, which is realized through the question *che COSA?*, produced by PKP130 after PKP125's overlapped turn (*[le <mad> così le devi rifare]?*). The repair initiation functions as a request for clarification, targeting the referential or lexical content of the previous utterance. The subsequent turn by PKP125 (*le mad così le fai*) provides the reformulated element, which is then followed by a further elaboration by PKP130. In this case, the ASR system correctly transcribes the repair initiator. Unlike other short repair initiations that tend to be omitted, this one is preserved in the output. Several factors may account for this. First, the utterance is prosodically salient, as suggested by capitalization in the transcription, indicating emphasis or increased intensity. Second, it occurs outside direct overlap and forms a clearly bounded turn, which may facilitate its segmentation and recognition. The preservation of *che cosa?* therefore suggests that Whisper may be particularly sensitive to acoustic prominence, besides turn structure: when the repair initiator is prosodically marked and structurally independent, it may be more likely to survive decoding.

```
PKP125        |                                              [[le
PKP130        |TI RICORDI:?       ] [in due tutte e [due le volte:].→
event         |

PKP125        | così le devi rifare]?  ]
PKP130        |in due tutte e [due le ..]     [che COSA:?           →
event         |                               [OR                   →

PKP125        |             [le mad così le fai     ]
PKP130        |che COS..]                             [no: le mad l..→
event         |OR       ]

PKP125        |
PKP130        |le: la: proprio: la GRADUATORIA l'inserimento in gra..→
event         |

PKP125        |
PKP130        |graduatoria                                          →
event         |
```

Figure 6.25: Example of an other-initiated repair matched in KPS023 (KIPasti).

6.5 Summary

The results provide a convergent picture of how interactional phenomena are represented in ASR output under the two decoding strategies. While the distributional analysis of the gold transcripts confirms that these phenomena vary across conversational settings, threfore reflecting differences in institutional asymmetry, spontaneity and cognitive load, the evaluation of ASR performance reveals a striking stability across optimization configurations. Global transcription accuracy (WER) is highly comparable across strategies, with only a marginal reduction in mean error and dispersion under the Interaction-aware setting. Event-level analyses confirm the absence of systematic differences: mean match ratios do not significantly differ between configurations at conversation level, and only minimal fluctuations emerge across interaction types. Instead, recognition patterns are primarily phenomenon-dependent. Self-repairs tend to be comparatively well preserved, backchannels show moderate recognition, filled pauses are systematically suppressed, and other-initiated repairs remain vulnerable when short or overlapping. Substitution patterns are stable across configurations, and omissions constitute the dominant source of error. Overlap significantly reduces recognition probability, particularly for backchannels, but this effect is robust across both decoding strategies. Qualitative analyses further indicate a consistent normalization tendency: Whisper frequently linearizes repair sequences and suppresses prosodically weak or overlapping tokens, privileging grammatical coherence over interactional structure. Taken together, the findings suggest that decoding optimization exerts only incremental influence, while the structural and acoustic properties of interactional phenomena represent the primary determinants of their preservation in ASR output.

Chapter 7

Conclusions

7.1 Main Findings

This study set out to investigate whether decoding optimization can meaningfully influence the transcription of interactional phenomena in spontaneous Italian speech, and whether such adjustments affect global transcription accuracy. The results converge toward a remarkably consistent picture.

At the level of overall transcription accuracy (RQ1), decoding optimization exerts only a limited influence. Across configurations and subsets, Word Error Rate remains broadly comparable between the WER-based strategy, optimized exclusively to minimize transcription errors, and the Interaction-aware strategy, which incorporates event-level weighting to preserve interactional phenomena. While the Interaction-aware configuration yields a marginal reduction in mean WER and slightly lower dispersion in some cases, the differences remain small. Importantly, incorporating interaction-sensitive components into the objective function does not degrade global performance: this is particularly relevant, as it indicates that promoting the preservation of interactional phenomena does not come at the cost of overall transcription accuracy. Transcription accuracy may be driven more strongly by conversation-specific factors, such as acoustic quality, speaker variability, and overlap density, than by the choice of decoding objective.

At the event level, the impact of optimization is similarly incremental (RQ2). Conversation-level statistical testing reveals no systematic differences in mean match ratios between the two strategies. Although minor shifts emerge in substitution and insertion patterns, omission rates remain

virtually identical across conditions and consistently account for the majority of errors. This suggests that the suppression of interactional tokens is not substantially mitigated by introducing event-aware components into the loss function. In other words, while the Interaction-aware objective slightly reshapes the distribution of event-related errors, it does not fundamentally alter the system's treatment of conversational phenomena.

Recognition outcomes appear to be primarily phenomenon-dependent (RQ3). Clear asymmetries emerge across categories. Compared to other conversational phenomena, self-repairs show relatively high match ratios. However, qualitative analysis reveals that they are frequently linearized: the reformulated segment is preserved, while the disfluent material is deleted. Backchannels occupy an intermediate position, with moderate recognition rates that vary depending on sequential positioning and overlap. Filled pauses, by contrast, are systematically suppressed across interaction types and optimization strategies, displaying extremely low match ratios. Other-initiated repairs remain vulnerable, particularly when they are short, acoustically weak, or produced in overlap. These patterns indicate that structural and acoustic properties (e.g., brevity, prosodic salience, lexical completeness, sequential position) play a more decisive role than decoding optimization in determining whether an interactional event survives in the ASR output.

Conversational overlap further reinforces this interpretation (RQ4). Tokens produced in overlapping contexts exhibit significantly lower recognition probabilities compared than tokens produced in non-overlapping contexts, and this effect is especially pronounced for backchannels. Crucially, the overlap effect remains statistically significant under both decoding strategies. In practical terms, tokens produced in overlapping speech are consistently recognized less often than those produced in non-overlapping speech, regardless of whether the system is optimized for overall error minimization or for interactional phenomena preservation.

To summarize, these findings suggest that decoding optimization operates within relatively narrow margins. Adjusting decoding parameters and introducing event-sensitive components can slightly redistribute errors and modestly stabilize performance, but it does not override the model's underlying representational tendencies. Whisper consistently favors lexical continuity and grammatical coherence over the faithful preservation of incremental, interactional structure. Repair trajectories are compressed, short vocalizations are normalized or omitted, and prosodically weak tokens are frequently suppressed.

From a broader perspective, this study highlights that the visibility of interactional phenomena in ASR output is shaped less by superficial decoding choices and more by deeper modeling assumptions embedded in large end-to-end systems. Decoding configurations matter, but they operate against structural constraints inherent in the architecture and training objectives of the model. Consequently, meaningful improvements in the preservation of conversational phenomena may require deeper interventions (for instance, changes at the training or representation level). In this sense, decoding optimization can be understood as a methodological lever: it allows limited calibration of system behavior, however it cannot fully counteract the normalization bias that characterizes current large-scale ASR models.

7.2 Limitations and Future Work

Despite the methodological care adopted throughout this study, several limitations must be acknowledged.

First, the size of the dataset remains relatively modest. The gold-annotated portion amounts to 8 hours and 40 minutes of speech, while the optimization procedure was conducted on approximately 3 hours of conversational data (Subset A). Although sufficient for exploratory analysis, this size may limit the statistical power of the study, particularly for low-frequency phenomena, particularly other-initiated repairs. In addition, Subset A and Subset B were balanced primarily in terms of total duration rather than interactional composition: as a result, their distribution across interaction types is not perfectly symmetrical, which may partially affect generalizability. Furthermore, for time constraints, only the first 20 minutes of each conversation were annotated. However, this choice restricts the annotation to a fixed temporal window, which may have limited the capture of phenomena that occur later in the interaction, particularly those emerging in more developed phases. As a consequence, the reported frequencies and distributions should be interpreted as reflecting early-stage interactional behavior rather than the full conversational trajectory.

Second, the annotation process was performed by a single annotator. Although careful listening and clearly defined criteria were applied, the absence of inter-annotator agreement measures constitutes a limitation. Some phenomena, particularly filled pauses and short vocalic items, involve perceptual and prosodic judgments that may not be entirely objective. Future

work could strengthen the reliability of the findings by incorporating multi-annotator validation.

A further limitation concerns the form-based identification of interactional events in the quantitative phases. While manual annotation was functionally grounded, the Interaction-aware optimization and error analysis relied on predefined token lists. This approach allowed systematic computation but does not exhaustively capture the full structural and functional complexity of interactional phenomena, especially in the case of multi-unit or context-dependent sequences. A more structurally sensitive computational modeling of interactional events would represent a valuable extension.

With respect to optimization, the study focused exclusively on decoding-time parameter tuning, without modifying the internal architecture or training objective of the ASR model. As a consequence, improvements were constrained to inference-level adjustments, and deeper representational biases of the model could not be addressed. In addition, the search procedure was limited to 50 trials per configuration. Although early stopping and pruning improved efficiency, a broader exploration of the parameter space might yield different local optima. Finally, the weighting coefficients of the Interaction-aware loss function were chosen manually: alternative weighting schemes could potentially produce different outcomes.

Concerning evaluation metrics, Word Error Rate (WER), while standard in ASR research, is structurally ill-suited for capturing interactional organization. Although the Interaction-aware loss function was introduced to partially compensate for this limitation, both WER and the Mean Match Ratio ultimately operate at a form-preservation level, considering that they do not directly assess the sequential or pragmatic function of interactional phenomena. For instance, a self-repair that is partially preserved may still lose its interactional relevance, a nuance that quantitative metrics alone cannot fully capture.

The study is also limited by its exclusive focus on a single ASR system. Although the Whisper large-v3 model yielded the best empirical performance among the tested options, it remains possible that alternative or future models could display different behaviors with respect to interactional phenomena. The findings should therefore be interpreted as model-specific rather than universally generalizable across ASR systems.

A further limitation concerns the annotation scope of other-initiated repairs. Only short and minimal repair initiators (i.e., open initiators and restricted requests) were systematically annotated. Longer and more elabo-

rated repair sequences (e.g., restricted offers) were not included in the annotation, under the assumption that well-structured, lexically-rich and longer turns would be less susceptible to suppression by the ASR system. As a consequence, the statistical representation of other-initiated repair phenomena in this study likely underestimates their actual frequency within the corpus. The descriptive statistics reported should therefore be interpreted as referring specifically to minimal repair initiators rather than to the full spectrum of other-repair sequences.

From a theoretical perspective, the analysis documents a recurrent tendency of the ASR system toward linearization and normalization of conversational structure. However, the present study does not allow for a deeper explanation of whether this behavior stems from training data composition, architectural constraints, or unexplored decoding dynamics. The results describe the phenomenon empirically but do not fully uncover its underlying computational causes.

An important dimension that was not explored concerns the acoustic properties of interactional phenomena. The present study operates primarily at the textual and alignment level. Future research could integrate acoustic analysis to investigate whether preserved versus suppressed phenomena display systematic differences in duration, intensity, pitch contour, or spectral characteristics. Such an analysis could clarify whether recognition asymmetries are partially driven by measurable acoustic patterns.

Finally, while the present findings suggest that decoding-level optimization alone is insufficient to substantially alter the normalization tendencies of the model, the annotation work conducted for this study has resulted in a set of transcripts enriched with interactional annotations. Once integrated into the broader KIParla annotation framework, this material could support more robust experimental designs. Future work could explore fine-tuning strategies on conversational data explicitly annotated for interactional phenomena, as well as alternative weighting schemes assigning stronger emphasis to interactional preservation. Such experiments would allow researchers to test how far the model can be 'pushed' toward maintaining disfluent and incremental structures, thereby clarifying whether the suppression observed here reflects an adjustable optimization trade-off or a deeper architectural constraint.

Bibliography

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework, 2019. URL https://arxiv.org/abs/1907.10902.

Abdulla Almahmood, Hesham Al-Ammal, and Fatema Albalooshi. Enhancing speech-to-text transcription accuracy for the bahraini dialect. In *2024 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT)*, pages 508–514, 2024. doi: 10.1109/3ict64318.2024.10824280.

Carlos Arriaga, Alejandro Pozo, Javier Conde, and Alvaro Alonso. Assessing latency in asr systems: A methodological perspective for real-time use, 2025. URL https://arxiv.org/abs/2409.05674.

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. Whisperx: Time-accurate speech transcription of long-form audio, 2023. URL https://arxiv.org/abs/2303.00747.

Janet Baker, li Deng, James Glass, Sanjeev Khudanpur, Chin-Hui Lee, Nelson Morgan, and Douglas O'Shaughnessy. Developments and directions in speech recognition and understanding, part 1 [dsp education]. *Signal Processing Magazine, IEEE*, 26:75 – 80, 06 2009. doi: 10.1109/MSP.2009.932166.

Silvia Ballarè and Caterina Mauri. La creazione del corpus kiparla: criteri metodologici e prospettive future. *Rivista Italiana di Dialettologia*, 44, 01 2021.

Joaquín Bautista and Jordi Pereira. A dynamic programming based heuristic for the assembly line balancing problem. *European Journal of Operational Research*, 194(3):787–794, 2009. ISSN 0377-2217. doi: https://doi.

org/10.1016/j.ejor.2008.01.016. URL https://www.sciencedirect.com/science/article/pii/S0377221708001446.

J. Benesty, J. Chen, and Y. Huang. Linear prediction. In Jacob Benesty, M. Mohan Sondhi, and Yiteng Huang, editors, *Springer Handbook of Speech Processing*, pages 121–134. Springer-Verlag, Berlin, Heidelberg, 2007. ISBN 978-3-540-49125-5.

James Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. Algorithms for hyper-parameter optimization. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011. URL https://proceedings.neurips.cc/paper_files/paper/2011/file/86e8f7ab32cfd12577bc2619bc635690-Paper.pdf.

Peter Blomsma, Julija Vaitonyte, Gabriel Skantze, and Marc Swerts. Backchannel behavior is idiosyncratic. *Language and Cognition*, 16:1–24, 02 2024. doi: 10.1017/langcog.2024.1.

Loredana Cerrato and Maiapaola D'Imperio. Duration and tonal characteristics of short expressions in italian. In *Proceedings of International Congress of Phonetic Sciences*, Barcelona, 2003.

Massimo Cerruti and Silvia Ballarè. Modulo ParlaTO, 2020a. URL https://doi.org/10.60760/unibo/parlato.

Massimo Cerruti and Silvia Ballarè. ParlaTO: corpus del parlato di Torino. *Bollettino dell'Atlante Linguistico Italiano (BALI)*, 44:171–196, 2020b.

Akshay Chandrashekaran and Ian Lane. Automated optimization of decoder hyper-parameters for online lvcsr. In *2016 IEEE Spoken Language Technology Workshop (SLT)*, pages 454–460, 2016. doi: 10.1109/SLT.2016.7846303.

N. Christenfeld, S. Schachter, and F. Bilous. Filled pauses and gestures: It's not coincidence. *Journal of Psycholinguistic Research*, 20:1–10, 1991. doi: 10.1007/BF01076916.

Patricia Clancy, Sandra Thompson, Ryoko Suzuki, and Hongyin Tao. The conversational use of reactive tokens in english, japanese, and mandarin.

Journal of Pragmatics, 26:355–387, 09 1996. doi: 10.1016/0378-2166(95)00036-4.

Eve V. Clark. Conversational repair and the acquisition of language. *Discourse Processes*, 57(5-6):441–459, 2020. doi: 10.1080/0163853X.2020.1719795. URL https://doi.org/10.1080/0163853X.2020.1719795.

Francisco Cossavella and Jazmín Cevasco. The importance of studying the role of filled pauses in the construction of a coherent representation of spontaneous spoken discourse. *Journal of Cognitive Psychology*, 33(2):172–186, 2021. doi: 10.1080/20445911.2021.1893325. URL https://doi.org/10.1080/20445911.2021.1893325.

Ronald Cumbal, Birger Moell, José Lopes, and Olov Engwall. "you don't understand me!": Comparing asr results for l1 and l2 speakers of swedish. pages 4463–4467, 08 2021. doi: 10.21437/Interspeech.2021-2140.

S. K. Das and M. A. Picheny. Issues in practical large vocabulary isolated word recognition: The IBM tangora system. In Chin-Hui Lee, Frank K. Soong, and Kuldip K. Paliwal, editors, *Automatic Speech and Speaker Recognition: Advanced Topics*, pages 457–479. Kluwer Academic Publishers, Boston, MA, 1996.

K. Davis, R. Biddulph, and S. Balashek. Automatic recognition of spoken digits. *Journal of the Acoustical Society of America*, 24:637–642, 1952. URL https://api.semanticscholar.org/CorpusID:121505424.

Edson Antônio Gonçalves de Souza, Marcelo Seido Nagano, and Gustavo Alencar Rolim. Dynamic programming algorithms and their applications in machine scheduling: A review. *Expert Systems with Applications*, 190:116180, 2022. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2021.116180. URL https://www.sciencedirect.com/science/article/pii/S0957417421014998.

Xabier de Zuazo, Eva Navas, Ibon Saratxaga, and Inma Hernáez Rioja. Whisper-lm: Improving asr models with language models for low-resource languages, 2025. URL https://arxiv.org/abs/2503.23542.

Akshay Madhav Deshmukh. Comparison of hidden markov model and recurrent neural network in automatic speech recognition. *European Journal*

of Engineering and Technology Research, 5(8):958–965, aug 2020. doi: 10.24018/ejeng.2020.5.8.2077. URL https://doi.org/10.24018/ejeng.2020.5.8.2077.

Tobias Deußer, Abdul Mohsin Siddiqi, Lorenz Sparrenberg, Tobias Adams, Christian Bauckhage, and Rafet Sifa. Fusing speech and language models for dementia detection. In *2024 IEEE International Conference on Big Data (BigData)*, pages 3908–3914, 2024. doi: 10.1109/BigData62323.2024.10825055.

Christina Dideriksen, Riccardo Fusaroli, Kristian Tylén, Mark Dingemanse, and Morten H. Christiansen. Contextualizing conversational strategies: Backchannel, repair and linguistic alignment in spontaneous and task-oriented conversations. In *Annual Meeting of the Cognitive Science Society*, 2019. URL https://api.semanticscholar.org/CorpusID:198654482.

Mark Dingemanse and N. J. Enfield. Other-initiated repair across languages: towards a typology of conversational structures. *Open Linguistics*, 1(1), 2015. doi: doi:10.2478/opli-2014-0007. URL https://doi.org/10.2478/opli-2014-0007.

Mark Dingemanse and N.J. Enfield. Interactive repair and the foundations of language. *Trends in Cognitive Sciences*, 28(1):30–42, 2024. ISSN 1364-6613. doi: https://doi.org/10.1016/j.tics.2023.09.003. URL https://www.sciencedirect.com/science/article/pii/S1364661323002504.

Tanvi Dinkar, Chloé Clavel, and Ioana Vasilescu. Fillers in spoken language understanding: Computational and psycholinguistic perspectives, 2023. URL https://arxiv.org/abs/2301.10761.

Paul Drew. 'open' class repair initiators in response to sequential sources of troubles in conversation. *Journal of Pragmatics*, 28(1):69–101, 1997. ISSN 0378-2166. doi: https://doi.org/10.1016/S0378-2166(97)89759-7. URL https://www.sciencedirect.com/science/article/pii/S0378216697897597.

Homer Dudley. The vocoder. *Bell Laboratories Record*, 17:122–126, 1939.

Homer Dudley, R. R. Riesz, and S. A. Watkins. A synthetic speaker. *Journal of the Franklin Institute*, 227:739–764, 1939.

Asmaa El Hannani and Thomas Hain. Automatic optimization of speech decoder parameters. *IEEE Signal Processing Letters*, 17(1):95–98, 2010. doi: 10.1109/LSP.2009.2033967.

Giolo Fele. *L'analisi della conversazione*. Il Mulino, Bologna, 2007. URL `https://www.mulino.it/isbn/9788815119506`.

Harvey Fletcher. The nature of speech and its interpretation. *Bell System Technical Journal*, 1:129–144, 1922.

Bruno Galantucci, Benjamin Langstein, Eliyahu Spivack, and Nathaniel Paley. Repair avoidance: When faithful informational exchanges don't matter that much. *Cognitive Science*, 44, 10 2020. doi: 10.1111/cogs.12882.

Rod Gardner. *When Listeners Talk: Response tokens and listener stance*, volume 92 of *Pragmatics & Beyond New Series*. John Benjamins, Amsterdam/Philadelphia, 2001. ISBN 9789027251114. doi: 10.1075/pbns.92.

Yashesh Gaur, Walter S. Lasecki, Florian Metze, and Jeffrey P. Bigham. The effects of automatic speech recognition quality on human transcription latency. In *Proceedings of the 13th International Web for All Conference*, W4A '16, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450341387. doi: 10.1145/2899475.2899478. URL `https://doi.org/10.1145/2899475.2899478`.

Anne Sophie Ghyselen, Anne Breitbarth, Melissa Farasyn, Jacques Van Keymeulen, and Arjan van Hessen. Clearing the transcription hurdle in dialect corpus building: The corpus of southern dutch dialects as case study. *Frontiers in Artificial Intelligence*, 3, 4 2020. ISSN 26248212. doi: 10.3389/frai.2020.00010.

Sharon Goldwater, Dan Jurafsky, and Christopher D. Manning. Which words are hard to recognize? prosodic, lexical, and disfluency factors that increase ASR error rates. In Johanna D. Moore, Simone Teufel, James Allan, and Sadaoki Furui, editors, *Proceedings of ACL-08: HLT*, pages 380–388, Columbus, Ohio, June 2008. Association for Computational Linguistics. URL `https://aclanthology.org/P08-1044/`.

Charles Goodwin. Between and within: Alternative sequential treatments of continuers and assessments. *Human Studies*, 9(2-3):205–217, 1986. doi: 10.1007/BF00148127.

Eugenio Goria and Caterina Mauri. Il corpus KIParla: una nuova risorsa per lo studio dell'italiano parlato. *CLUB Working Papers in Linguistics*, 2:96–116, 2018.

Jan Gorisch and Thomas Schmidt. Evaluating workflows for creating orthographic transcripts for oral corpora by transcribing from scratch or correcting ASR-output. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6564–6574, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.582/.

Jonathan Gratch, Anna Okhmatovskaia, Francois Lamothe, Stacy Marsella, Mathieu Morales, R. J. van der Werf, and Louis-Philippe Morency. Virtual rapport. In *Proceedings of the 6th International Conference on Intelligent Virtual Agents*, IVA'06, page 14–27, Berlin, Heidelberg, 2006. Springer-Verlag. ISBN 3540375937. doi: 10.1007/11821830_2. URL https://doi.org/10.1007/11821830_2.

Alessandro Gregori. *Automatic Speech Recognition (ASR) and NMT for Interlingual and Intralingual Communication: Speech to Text Technology for Live Subtitling and Accessibility.* PhD thesis, alma, Ottobre 2021. URL https://amsdottorato.unibo.it/id/eprint/9931/.

Nyoman Gunantara. A review of multi-objective optimization: Methods and its applications. *Cogent Engineering*, 5(1):1502242, 2018. doi: 10.1080/23311916.2018.1502242.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. URL https://arxiv.org/abs/2106.09685.

IBM. Gaussian mixture model, 2024. URL https://www.ibm.com/think/topics/gaussian-mixture-model.

Aleksandr V Isačenko and Hans Joachim Schädlich. *Untersuchungen über die deutsche Satzintonation.* Akademie-Verlag, 1964.

Gail Jefferson. Side sequences. In David Sudnow, editor, *Studies in Social Interaction*, chapter 9, page 294–338. Free Press, New York, 1972.

Gail Jefferson. Notes on a systematic deployment of the acknowledgement tokens "yeah"; and "mm hm";. *Paper in Linguistics*, 17(2):197–216, 1984. doi: 10.1080/08351818409389201. URL https://doi.org/10.1080/08351818409389201.

Gail Jefferson. Glossary of transcript symbols with an introduction. In Gene H. Lerner, editor, *Conversation Analysis: Studies from the First Generation*, chapter 2, page 13–31. John Benjamins, Amsterdam / Philadelphia, 2004. doi: 10.1075/pbns.125.02jef. URL https://benjamins.com/catalog/pbns.125.02jef.

F. Jelinek. Continuous speech recognition by statistical methods. *Proceedings of the IEEE*, 64(4):532–556, 1976. doi: 10.1109/PROC.1976.10159.

S Karpagavalli and Evania Chandra. A review on automatic speech recognition architecture and approaches. *International Journal of Signal Processing, Image Processing and Pattern Recognition*, 9:393–404, 04 2016. doi: 10.14257/ijsip.2016.9.4.34.

Dietrich Klakow and Jochen Peters. Testing the correlation of word error rate and perplexity. *Speech Communication*, 38(1):19–28, 2002. ISSN 0167-6393. doi: https://doi.org/10.1016/S0167-6393(01)00041-3. URL https://www.sciencedirect.com/science/article/pii/S0167639301000413.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of System Demonstrations of the 27th International Conference on Computational Linguistics*, pages 5–9, Santa Fe, NM, 07 2018.

K. F. Lee. *Large Vocabulary Speaker Independent Continuous Speech Recognition: The Sphinx System*. PhD thesis, Carnegie Mellon University, Pittsburgh, PA, 1988.

Christian Lehmann. Data in linguistics. *Linguistic Review - LINGUIST REV*, 21:175–210, 01 2004. doi: 10.1515/tlir.2004.21.3-4.175.

Andreas Liesenfeld, Alianda Lopez, and Mark Dingemanse. The timing bottleneck: Why timing and overlap are mission-critical for conversational

user interfaces, speech recognition and dialogue systems. 2023. URL https://osf.io/hruva.

Jhen-Ke Lin, Hao-Chien Lu, Chung-Chun Wang, Hong-Yun Lin, and Berlin Chen. Acoustically precise hesitation tagging is essential for end-to-end verbatim transcription systems, 2025. URL https://arxiv.org/abs/2506.04076.

Alianda Lopez, Andreas Liesenfeld, and Mark Dingemanse. Evaluation of automatic speech recognition for conversational speech in Dutch, English and German: What goes missing? In Robin Schaefer, Xiaoyu Bai, Manfred Stede, and Torsten Zesch, editors, *Proceedings of the 18th Conference on Natural Language Processing (KONVENS 2022)*, pages 135–143, Potsdam, Germany, 12–15 September 2022. KONVENS 2022 Organizers. URL https://aclanthology.org/2022.konvens-1.16/.

Bruce Lowerre. *The Harpy speech understanding system*, page 576–586. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1990. ISBN 1558601244.

Alfredo Maffi. *Studio e sviluppo di un framework per il riconoscimento vocale nell'ambito di sistemi Hands-Free*. PhD thesis, 2016. URL https://amslaurea.unibo.it/id/eprint/11001/.

Barbara Maroni and Francesco Arcidiacono. Conversational repair in italian families. *Studies in Communication Sciences*, 10:181–206, 2010.

Caterina Mauri, Silvia Ballarè, Eugenio Goria, Massimo Cerruti, and Francesco Suriano. KIParla corpus: A new resource for spoken Italian. In Raffaella Bernardi, Roberto Navigli, and Giovanni Semeraro, editors, *Proceedings of the Sixth Italian Conference on Computational Linguistics (CLiC-it 2019)*, pages 243–249, Bari, Italy, November 2019a. CEUR Workshop Proceedings. ISBN 979-1-280-13600-8. URL https://aclanthology.org/2019.clicit-1.37/.

Caterina Mauri, Eugenio Goria, and Silvia Ballarè. Modulo KIP, 2019b. URL https://doi.org/10.60760/unibo/kip.

Caterina Mauri, Silvia Ballarè, and Eleonora Zucchini. Modulo ParlaBO, 2024a. URL https://doi.org/10.60760/unibo/parlabo.

Caterina Mauri, Silvia Ballarè, and Eleonora Zucchini. Modulo KIPasti, 2024b. URL https://doi.org/10.60760/unibo/kipasti.

Max Planck Institute for Psycholinguistics. *ELAN (Version 7.0)*. The Language Archive, Nijmegen, 2025. URL https://archive.mpi.nl/tla/elan. Computer software.

Raveesh Meena, Gabriel Skantze, and Joakim Gustafson. The map task dialogue system: A test-bed for modelling human-like dialogue. In Maxine Eskenazi, Michael Strube, Barbara Di Eugenio, and Jason D. Williams, editors, *Proceedings of the SIGDIAL 2013 Conference*, pages 366–368, Metz, France, August 2013. Association for Computational Linguistics. URL https://aclanthology.org/W13-4059/.

Daniela Mereu, Francesco Cangemi, and Martine Grice. Backchannels are not always very short utterances. the case of italian multi-unit backchannels. *Journal of Pragmatics*, 228:1–16, 2024. ISSN 0378-2166. doi: https://doi.org/10.1016/j.pragma.2024.05.003. URL https://www.sciencedirect.com/science/article/pii/S0378216624000857.

Hansjörg Mixdorff and Hartmut R. Pfitzinger. Analysing fundamental frequency contours and local speech rate in map task dialogs. *Speech Communication*, 46(3):310–325, 2005. ISSN 0167-6393. doi: https://doi.org/10.1016/j.specom.2005.02.019. URL https://www.sciencedirect.com/science/article/pii/S0167639305000920. Quantitative Prosody Modelling for Natural Speech Description and Generation.

H. Murveit, M. Cohen, P. Price, G. Baldwin, M. Weintraub, and J. Bernstein. SRI's DECIPHER system. In *Proceedings of the Speech and Natural Language Workshop*, pages 238–242, Philadelphia, PA, 1989. Morgan Kaufmann.

Pilar Pernas and Margarita Borreguero Zuloaga. Cortesia e scortesia in un contesto di apprendimento linguistico: la gestione dei turni. In Marcello Pettorino, Antonietta Giannini, and Francescamaria Dovetto, editors, *La Comunicazione Parlata 3. Atti Del Congresso Internazionale (Napoli, 23-25 Febbraio 2009)*, volume I, pages 227–247. Università Degli Studi Napoli L'Orientale, Napoli, 2010.

Rohit Prabhavalkar, Takaaki Hori, Tara N. Sainath, Ralf Schlüter, and Shinji Watanabe. End-to-end speech recognition: A survey, 2023. URL https://arxiv.org/abs/2303.03329.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021. URL https://www.R-project.org/.

Lawrence R. Rabiner. Automatic speech recognition - a brief history of the technology development. 2004. URL https://api.semanticscholar.org/CorpusID:12721778.

Lawrence R. Rabiner and Biing-Hwang Juang. *Fundamentals of Speech Recognition*, volume 14. Prentice Hall, Upper Saddle River, NJ, USA, apr 1993. ISBN 0-13-015157-2.

Lawrence R. Rabiner and B.-H. Jung. Historical perspective of the field of asr/nlu. In Jacob Benesty, M. Mohan Sondhi, and Yiteng Huang, editors, *Springer Handbook of Speech Processing*, pages 521–537. Springer-Verlag, Berlin, Heidelberg, 2007. ISBN 978-3-540-49125-5.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR, 2023.

Monica Romero, Sandra Gómez-Canaval, and Ivan G. Torre. Automatic speech recognition advancements for indigenous languages of the americas. *Applied Sciences*, 14(15), 2024. ISSN 2076-3417. doi: 10.3390/app14156497. URL https://www.mdpi.com/2076-3417/14/15/6497.

Giovanni Rossi. Other-initiated repair in italian. *Open Linguistics*, 1:256–282, 05 2015. doi: 10.1515/opli-2015-0002.

T. Sakai and S. Doshita. Phonetic typewriter. *Journal of the Acoustical Society of America*, 33:1664–1664, 1961.

Michelina Savino. The intonation of backchannel tokens in italian collaborative dialogues. In Zygmunt Vetulani and Joseph Mariani, editors, *Human Language Technology Challenges for Computer Science and Linguistics*,

pages 28–39, Cham, 2014. Springer International Publishing. ISBN 978-3-319-08958-4.

Michelina Savino and Mario Refice. Acknowledgement or reply? prosodic features for disambiguating pragmatic functions of the italian token 'sì'. *2013 7th Conference on Speech Technology and Human - Computer Dialogue (SpeD)*, pages 1–6, 2013. URL `https://api.semanticscholar.org/CorpusID:18419478`.

Simona Sbranna, Eduardo Möking, Simon Wehrle, and Martine Grice. Backchannelling across languages: Rate, lexical choice and intonation in l1 italian, l1 german and l2 german. *Speech Prosody 2022*, 2022. URL `https://api.semanticscholar.org/CorpusID:248897851`.

R.W. Schafer. Homomorphic systems and cepstrum analysis of speech. In Jacob Benesty, M. Mohan Sondhi, and Yiteng Huang, editors, *Springer Handbook of Speech Processing*, pages 161–180. Springer-Verlag, Berlin, Heidelberg, 2007. ISBN 978-3-540-49125-5.

Emanuel Schegloff. *Discourse as an interactional achievement: Some uses of 'uh huh' and other things that come between sentences*, pages 71–93. Georgetown University Press, 01 1982.

Emanuel Schegloff. Sequence organization in interaction: A primer in conversation analysis. *Sequence Organization in Interaction: A Primer in Conversation Analysis I*, 1:1–300, 01 2007. doi: 10.1017/CBO9780511791208.

Emanuel Schegloff, Gail Jefferson, and Harvey Sacks. The preference for self-correction in the organization of repair in conversation. *Language*, 53: 361–382, 06 1977. doi: 10.2307/413107.

Loredana Schettino and Violetta Cataldo. Lexicalized pauses in italian. pages 189–192, 11 2019. doi: 10.36505/ExLing-2019/10/0047/000409.

R. Schwartz, C. Barry, Y.-L. Chow, A. Derr, M.-W. Feng, O. Kimball, F. Kubala, J. Makhoul, and J. Vandegrift. The BBN BYBLOS continuous speech recognition system. In *Proceedings of the Speech and Natural Language Workshop*, pages 94–99, Philadelphia, PA, 1989. Morgan Kaufmann.

Lorenzo Spreafico. Le pause piene nel parlato plurilingue. In *Lessico e lessicologia: atti del XLIV Congresso internazionale di studi della Società di linguistica italiana (SLI)*, number 56 in Pubblicazioni della Società di linguistica italiana, pages 355–368, Roma, 2012. Bulzoni. doi: 10.1400/202123. URL http://digital.casalini.it/10.1400/202123.

Eberhard Stock and Christina Zacharias. Deutsche satzintonation. *(No Title)*, 1973.

Rosanna Turrisi. *On Deep Learning strategies to address Automatic Speech Recognition (ASR) for dysarthric speech*. PhD thesis, Università degli Studi di Ferrara, 2021. URL https://hdl.handle.net/20.500.14242/169206. Tesi di dottorato.

Muhammad Umair, Julia Mertens, Saul Albert, and J. Ruiter. Gailbot: An automatic transcription system for conversation analysis. *Dialogue Discourse*, 13:63–95, 04 2022. doi: 10.5210/dad.2022.103.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023. URL https://arxiv.org/abs/1706.03762.

Alessandro Vinciarelli, Paraskevi Chatziioannou, and Anna Esposito. When the words are not everything: The use of laughter, fillers, back-channel, silence, and overlapping speech in phone calls. *Frontiers in ICT*, Volume 2 - 2015, 2015. ISSN 2297-198X. doi: 10.3389/fict.2015.00004. URL https://www.frontiersin.org/journals/ict/articles/10.3389/fict.2015.00004.

Vincenzo Norman Vitale, Loredana Schettino, and Francesco Cutugno. Modelling filled particles and prolongation using end-to-end automatic speech recognition systems: A quantitative and qualitative analysis. In Felice Dell'Orletta, Alessandro Lenci, Simonetta Montemagni, and Rachele Sprugnoli, editors, *Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 990–996, Pisa, Italy, December 2024a. CEUR Workshop Proceedings. ISBN 979-12-210-7060-6. URL https://aclanthology.org/2024.clicit-1.107/.

V.N. Vitale, L. Schettino, and F. Cutugno. Rich speech signal: exploring and exploiting end-to-end automatic speech recognizers' ability to model

hesitation phenomena. In *Proc. Interspeech 2024*, pages 222–226, 2024b. doi: 10.21437/Interspeech.2024-2029.

Dong Wang, Xiaodong Wang, and Shaohe Lv. An overview of end-to-end automatic speech recognition. *Symmetry*, 11(8), 2019. ISSN 2073-8994. doi: 10.3390/sym11081018. URL https://www.mdpi.com/2073-8994/11/8/1018.

Huan Wang, Jie Bin, Chunyan Gou, Lian Yang, Baolin Hou, and Mingwei Qin. Low-resource speech recognition by fine-tuning whisper with optuna-lora. *Applied Sciences*, 15(24), 2025a. ISSN 2076-3417. doi: 10.3390/app152413090. URL https://www.mdpi.com/2076-3417/15/24/13090.

Yu Wang, Leyi Lao, Langchu Huang, Gabriel Skantze, Yang Xu, and Hendrik Buschmeier. Investigating the representation of backchannels and fillers in fine-tuned language models, 2025b. URL https://arxiv.org/abs/2509.20237.

Nigel Ward and Wataru Tsukahara. Tsukahara, w.: Prosodic features which cue back-channel responses in english and japanese. journal of pragmatics 23, 1177-1207. *Journal of Pragmatics*, 32:1177–1207, 07 2000. doi: 10.1016/S0378-2166(99)00109-5.

Hiroyoshi Yamasaki, Jérôme Louradour, Julie Hunter, and Laurent Prévot. Transcribing and aligning conversational speech: A hybrid pipeline applied to french conversations, 2023.

Liu Yang, Elizabeth Shriberg, and Andreas Stolcke. Automatic disfluency identification in conversational speech using multiple knowledge sources. 09 2003. doi: 10.21437/Eurospeech.2003-332.

Kaisheng Yao, Dong Yu, Frank Seide, Hang Su, Li Deng, and Yifan Gong. Adaptation of context-dependent deep neural networks for automatic speech recognition. In *2012 IEEE Spoken Language Technology Workshop (SLT)*, pages 366–369, 2012. doi: 10.1109/SLT.2012.6424251.

B. Yegnanarayana. Artificial neural networks for pattern recognition. *Sādhanā*, 19(2):189–238, 1994. doi: 10.1007/BF02811896. URL https://link.springer.com/article/10.1007/BF02811896.

Victor H. Yngve. On getting a word in edgewise. In *CLS-70*, pages 567–577. University of Chicago, 1970.

Steve Young. Large vocabulary continuous speech recognition: A review. *IEEE Signal Processing Magazine*, 1996.

Dong Yu and Li Deng. *Automatic Speech Recognition: A Deep Learning Approach*. Signals and Communication Technology. Springer, 2015.

Vicky Zayats, Trang Tran, Richard Wright, Courtney Mansfield, and Mari Ostendorf. Disfluencies and human speech transcription errors. In *Interspeech 2019*, pages 3088–3092, 2019. doi: 10.21437/Interspeech.2019-3134.

Xu-Yao Zhang, Cheng-Lin Liu, and Ching Y. Suen. Towards robust pattern recognition: A review. *Proceedings of the IEEE*, 108(6):894–922, 2020. doi: 10.1109/JPROC.2020.2989782.

Appendix A

Metadata

Conversation ID	Type	Speakers	Languages	Participants	Module
PBB024	semi-structured interview	2	Italian-dialect	BOI091, BOR030	ParlaBO
PBB019	semi-structured interview	2	Italian	BOI057, BOR009	ParlaBO
PBB010	semi-structured interview	2	Italian	BOI030, BOR007	ParlaBO
PBC017	semi-structured interview	2	Italian	BOI042, BOR009	ParlaBO
PBB029	semi-structured interview	2	Italian	BOI101, BOR036	ParlaBO
PBA014	semi-structured interview	2	Italian	BOI076, BOR019	ParlaBO
PTA016	semi-structured interview	2	Italian	TOI033, TOR006	ParlaTO
PTD020	semi-structured interview	2	Italian	TOI008, TOR001	ParlaTO
PTB007	semi-structured interview	2	Italian-dialect	TOI021, TOR004	ParlaTO
PTD003	semi-structured interview	2	Italian	TOI048, TOR001	ParlaTO
PTA018	semi-structured interview	2	Italian	TOI031, TOR006	ParlaTO
TOD2016	semi-structured interview	2	Italian	TO071, TO080	ParlaTO
BOC1005	exam	2	Italian	BO104, BO105	KIP
TOC1002	exam	2	Italian	TO006, TO007	KIP
TOC1001	exam	2	Italian	TO002, TO003	KIP
BOC1007	exam	2	Italian	BO104, BO144	KIP
BOC1002	exam	2	Italian	BO087, BO089	KIP
BOC1003	exam	2	Italian	BO087, BO090	KIP
TOA1005	office hours	2	Italian	TO075, TO076	KIP
TOA1003	office hours	3	Italian	TO061, TO065	KIP
TOA3010	free conversation	2	Italian	TO085, TO086	KIP
KPS024	free conversation	2	Italian-dialect	PKP125, PKP131	KIPasti
KPC004	free conversation	2	Italian	PKP027, PKP028	KIPasti
KPN027	free conversation	2	Italian	PKP115, PKP116	KIPasti
KPN017	free conversation	2	Italian-dialect	PKP057, PKP061	KIPasti
KPS023	free conversation	2	Italian-dialect	PKP125, PKP130	KIPasti

Table A.1: Overview of the conversations included in the dataset, with interaction type, number of participants, languages, participant codes and corpus module.

Appendix B

Optuna Trials - WER-based

B.1 Configuration A

trial	wer	temperature	beam	best_of	no_speech	compression	cond_prev	patience	length_penalty
0	0.90	0.26	8	10	0.47	1.99	TRUE	0.38	0.21
1	0.60	0.62	3	5	0.26	2.56	FALSE	0.23	0.56
2	0.38	0.72	7	2	0.63	1.75	TRUE	0.30	0.60
3	0.40	0.33	9	8	0.72	2.28	TRUE	0.37	0.57
4	0.38	0.67	6	8	0.75	2.12	TRUE	0.39	0.75
5	0.36	0.24	4	7	0.21	2.49	TRUE	0.36	0.23
6	0.51	0.74	7	1	0.23	2.36	TRUE	0.43	0.58
7	0.69	0.06	4	10	0.76	1.96	TRUE	0.71	0.73
8	0.48	0.77	4	10	0.54	2.77	FALSE	0.74	0.46
9	0.39	0.44	9	5	0.27	1.92	TRUE	0.47	0.72
10	0.39	0.96	1	7	0.41	1.58	FALSE	0.62	0.20
11	0.37	0.10	5	7	0.38	2.43	TRUE	0.53	0.34
12	0.37	0.00	2	6	0.36	2.53	TRUE	0.57	0.34
13	0.36	0.18	5	3	0.34	2.50	TRUE	0.53	0.34
14	0.37	0.19	5	3	0.30	2.77	FALSE	0.64	0.31
15	0.37	0.39	4	4	0.21	2.62	TRUE	0.31	0.42
16	0.38	0.19	2	3	0.33	2.25	TRUE	0.51	0.28
17	0.37	0.55	6	1	0.46	2.44	TRUE	0.31	0.38
18	0.47	0.30	10	4	0.55	2.65	FALSE	0.23	0.27
19	0.37	0.17	3	6	0.21	2.15	TRUE	0.59	0.44
20	0.37	0.41	7	8	0.32	2.28	TRUE	0.47	0.26
21	0.36	0.14	3	6	0.20	2.12	TRUE	0.59	0.48
22	0.38	0.11	3	7	0.28	2.09	TRUE	0.67	0.49
23	0.39	0.27	5	4	0.20	1.81	TRUE	0.55	0.37

trial	wer	temperature	beam	best_of	no_speech	compression	cond_prev	patience	length_penalty
24	0.38	0.00	1	6	0.40	2.40	TRUE	0.45	0.30
25	0.38	0.50	4	9	0.26	2.23	TRUE	0.63	0.63
26	0.41	0.23	3	5	0.33	2.67	FALSE	0.79	0.50
27	0.40	0.12	2	3	0.26	2.50	TRUE	0.50	0.40
28	0.62	0.34	6	2	0.37	2.04	TRUE	0.41	0.24
29	0.76	0.26	5	7	0.45	2.35	TRUE	0.35	0.68
30	0.38	0.06	4	9	0.52	2.21	TRUE	0.58	0.22
31	0.41	0.18	3	6	0.21	2.16	TRUE	0.59	0.45
32	0.40	0.16	3	6	0.24	1.89	TRUE	0.55	0.52
33	0.39	0.24	2	5	0.29	2.03	TRUE	0.61	0.44
34	0.37	0.35	3	5	0.24	2.32	TRUE	0.48	0.53
35	0.36	0.34	4	4	0.67	2.32	TRUE	0.26	0.63
36	0.59	0.30	5	2	0.67	2.58	FALSE	0.24	0.62
37	0.37	0.47	6	4	0.60	2.54	TRUE	0.27	0.66
38	0.37	0.57	4	3	0.69	2.47	TRUE	0.37	0.78
39	0.46	0.34	8	4	0.80	2.39	TRUE	0.27	0.32
40	0.36	0.40	4	8	0.63	2.58	FALSE	0.34	0.58
41	0.38	0.40	4	8	0.65	2.70	FALSE	0.20	0.58
42	0.39	0.23	5	9	0.61	2.59	FALSE	0.34	0.56
43	0.37	0.07	4	7	0.59	2.32	FALSE	0.41	0.60
44	0.38	0.89	6	8	0.50	2.50	FALSE	0.28	0.68
45	0.39	0.30	4	7	0.72	2.79	FALSE	0.34	0.56
46	0.63	0.44	5	5	0.65	2.43	TRUE	0.66	0.71
47	0.37	0.55	7	9	0.56	1.51	FALSE	0.20	0.61
48	0.39	0.64	3	8	0.73	2.71	TRUE	0.43	0.35
49	0.41	0.37	2	1	0.35	2.60	TRUE	0.54	0.49

B.2 Configuration B

trial	wer	temperature	beam	best_of	no_speech	compression	cond_prev	patience	length_penalty
0	0.56	0.80	1	9	0.35	2.14	FALSE	0.27	0.28
1	0.39	0.30	10	2	0.23	2.28	TRUE	0.57	0.62
2	0.70	0.29	1	3	0.51	1.90	TRUE	0.47	0.65
3	0.52	0.22	9	5	0.60	1.54	TRUE	0.71	0.73
4	0.41	0.61	9	7	0.51	2.23	FALSE	0.48	0.52
5	0.44	0.53	6	10	0.55	1.77	TRUE	0.57	0.69
6	0.39	0.39	6	4	0.30	2.49	FALSE	0.39	0.53
7	0.44	0.25	10	10	0.61	1.53	TRUE	0.40	0.75
8	0.42	0.25	7	6	0.22	1.99	FALSE	0.24	0.54
9	0.44	0.82	8	1	0.37	1.59	TRUE	0.52	0.27
10	0.41	0.00	4	1	0.73	2.67	TRUE	0.79	0.37
11	0.40	0.40	5	3	0.21	2.48	FALSE	0.63	0.59
12	0.51	0.04	3	3	0.34	2.37	FALSE	0.37	0.47
13	0.41	0.42	7	4	0.28	2.74	FALSE	0.37	0.40
14	0.40	0.64	10	2	0.42	2.51	TRUE	0.63	0.62
15	0.42	0.13	3	5	0.30	2.26	FALSE	0.31	0.80
16	0.45	0.41	6	8	0.44	2.61	TRUE	0.45	0.45
17	0.47	0.72	8	4	0.27	2.42	FALSE	0.57	0.59
18	0.78	0.99	5	2	0.80	2.03	FALSE	0.21	0.20
19	0.48	0.49	8	6	0.42	2.31	TRUE	0.70	0.42
20	0.41	0.34	3	4	0.20	2.12	TRUE	0.42	0.56
21	0.39	0.63	10	2	0.44	2.54	TRUE	0.64	0.65
22	0.40	0.49	10	2	0.31	2.57	TRUE	0.53	0.67
23	0.40	0.12	9	1	0.27	2.45	TRUE	0.63	0.50

trial	wer	temperature	beam	best_of	no_speech	compression	cond_prev	patience	length_penalty
24	0.40	0.58	7	3	0.45	2.79	TRUE	0.73	0.60
25	0.40	0.70	9	2	0.39	2.64	TRUE	0.58	0.70
26	0.53	0.35	10	4	0.26	2.35	FALSE	0.67	0.62
27	0.76	0.17	8	2	0.33	2.22	TRUE	0.33	0.55
28	0.43	0.94	2	5	0.47	2.54	FALSE	0.78	0.79
29	0.59	0.75	4	1	0.36	2.11	FALSE	0.43	0.36
30	0.40	0.80	9	3	0.25	2.70	TRUE	0.52	0.65
31	0.39	0.64	10	2	0.41	2.50	TRUE	0.63	0.62
32	0.40	0.54	10	3	0.40	2.41	TRUE	0.61	0.56
33	0.40	0.66	1	2	0.54	2.30	TRUE	0.67	0.64
34	0.39	0.33	9	1	0.49	2.56	TRUE	0.59	0.72
35	0.78	0.44	10	4	0.32	2.46	TRUE	0.49	0.52
36	0.41	0.55	9	7	0.61	2.21	TRUE	0.66	0.67
37	0.50	0.61	8	2	0.48	2.17	TRUE	0.73	0.74
38	0.41	0.84	7	3	0.35	2.36	FALSE	0.55	0.49
39	0.44	0.47	10	5	0.54	2.60	TRUE	0.75	0.58
40	0.40	0.30	9	1	0.67	1.86	TRUE	0.50	0.68
41	0.40	0.35	9	1	0.50	2.53	TRUE	0.60	0.69
42	0.45	0.25	10	1	0.57	2.68	TRUE	0.58	0.72
43	0.40	0.30	10	2	0.24	2.57	TRUE	0.59	0.63
44	0.40	0.21	9	3	0.49	2.50	TRUE	0.55	0.71
45	0.40	0.57	8	1	0.52	2.42	TRUE	0.46	0.76
46	0.41	0.38	6	2	0.43	2.74	FALSE	0.66	0.53
47	0.40	0.46	9	3	0.39	1.62	TRUE	0.69	0.61
48	0.55	0.23	10	9	0.58	2.65	FALSE	0.64	0.64
49	0.40	0.52	6	1	0.66	2.32	TRUE	0.38	0.45

B.3 Configuration C

trial	wer	temperature	beam	best_of	no_speech	compression	cond_prev	patience	length_penalty
0	0.56	0.52	3	1	0.47	2.38	FALSE	0.59	0.22
1	0.41	0.09	3	9	0.79	2.37	FALSE	0.61	0.31
2	0.41	0.66	3	9	0.57	2.17	FALSE	0.60	0.74
3	0.39	0.92	4	8	0.76	2.75	FALSE	0.23	0.65
4	0.51	0.20	1	6	0.63	2.13	TRUE	0.73	0.39
5	0.37	0.55	4	5	0.59	2.16	TRUE	0.37	0.66
6	0.37	0.55	10	3	0.25	2.68	FALSE	0.58	0.50
7	0.41	0.77	1	1	0.65	2.46	FALSE	0.50	0.69
8	0.42	0.65	3	6	0.43	2.69	FALSE	0.57	0.43
9	0.37	0.76	8	9	0.63	2.04	TRUE	0.47	0.46
10	0.46	0.32	10	3	0.24	1.62	TRUE	0.78	0.57
11	0.41	0.98	9	4	0.22	1.82	TRUE	0.37	0.50
12	0.70	0.36	8	3	0.33	1.89	TRUE	0.41	0.52
13	0.38	0.84	7	8	0.36	1.94	TRUE	0.48	0.38
14	0.39	0.72	7	10	0.70	2.56	FALSE	0.26	0.57
15	0.38	0.41	10	7	0.52	1.66	TRUE	0.68	0.80
16	0.40	0.54	8	3	0.32	2.00	FALSE	0.45	0.46
17	0.55	0.83	6	5	0.41	2.31	TRUE	0.52	0.30
18	0.49	0.64	9	2	0.53	2.59	FALSE	0.69	0.59
19	0.38	0.41	9	10	0.69	1.51	TRUE	0.33	0.35
20	0.40	0.22	6	7	0.47	2.05	TRUE	0.66	0.47
21	0.37	0.57	5	5	0.60	2.24	TRUE	0.33	0.64
22	0.38	0.60	5	4	0.59	2.80	TRUE	0.31	0.53
23	0.38	0.77	8	4	0.72	2.25	TRUE	0.44	0.64

trial	wer	temperature	beam	best_of	no_speech	compression	cond_prev	patience	length_penalty
24	0.38	0.45	7	2	0.64	1.81	TRUE	0.53	0.56
25	0.38	0.71	10	5	0.55	2.54	TRUE	0.27	0.44
26	0.73	0.59	5	7	0.28	2.25	FALSE	0.21	0.62
27	0.59	0.49	9	6	0.49	2.08	TRUE	0.46	0.73
28	0.41	0.85	6	2	0.39	2.45	FALSE	0.40	0.50
29	0.55	0.49	8	1	0.62	2.68	FALSE	0.56	0.29
30	0.38	0.34	4	4	0.68	2.43	TRUE	0.64	0.69
31	0.37	0.56	4	5	0.60	2.24	TRUE	0.34	0.70
32	0.38	0.01	5	5	0.51	2.34	TRUE	0.33	0.75
33	0.38	0.61	2	8	0.61	2.23	TRUE	0.29	0.24
34	0.38	0.70	4	9	0.77	2.08	TRUE	0.25	0.60
35	0.38	0.52	2	6	0.57	1.98	FALSE	0.42	0.70
36	0.39	0.92	3	3	0.73	2.17	TRUE	0.61	0.67
37	0.41	0.27	5	7	0.45	2.36	FALSE	0.36	0.40
38	0.38	0.77	3	4	0.55	1.87	TRUE	0.55	0.80
39	0.56	0.56	4	5	0.68	2.64	FALSE	0.48	0.55
40	0.37	0.67	7	6	0.65	2.51	TRUE	0.62	0.73
41	0.39	0.66	7	6	0.66	2.74	TRUE	0.73	0.75
42	0.38	0.63	6	8	0.62	2.51	TRUE	0.60	0.71
43	0.39	0.70	7	9	0.80	2.62	TRUE	0.64	0.67
44	0.39	0.46	10	6	0.60	2.41	TRUE	0.59	0.76
45	0.38	0.58	8	5	0.71	2.49	TRUE	0.38	0.62
46	0.37	0.67	9	3	0.65	2.21	TRUE	0.34	0.72
47	0.40	0.75	6	4	0.74	2.30	FALSE	0.51	0.77
48	1.12	0.39	5	10	0.54	2.13	TRUE	0.79	0.48
49	0.52	0.81	7	1	0.59	2.74	TRUE	0.62	0.35

B.4 Configuration D

trial	wer	temperature	beam	best_of	no_speech	compression	cond_prev	patience	length_penalty
0	0.58	0.61	2	6	0.23	1.63	TRUE	0.76	0.77
1	0.39	0.55	2	4	0.39	1.54	TRUE	0.36	0.67
2	0.37	0.35	8	6	0.35	2.46	TRUE	0.33	0.72
3	0.40	0.69	10	3	0.45	1.65	TRUE	0.31	0.65
4	0.49	0.96	7	10	0.69	1.69	FALSE	0.57	0.67
5	0.39	0.78	8	7	0.45	1.87	TRUE	0.41	0.61
6	0.49	0.40	6	3	0.59	2.63	FALSE	0.70	0.42
7	0.52	0.54	10	3	0.58	1.54	TRUE	0.47	0.71
8	0.93	0.08	2	6	0.42	2.08	TRUE	0.26	0.47
9	0.91	0.17	7	3	0.70	1.86	FALSE	0.41	0.80
10	0.36	0.38	4	9	0.25	2.61	FALSE	0.22	0.24
11	0.38	0.32	4	9	0.24	2.61	FALSE	0.22	0.22
12	0.39	0.30	4	8	0.32	2.41	FALSE	0.20	0.22
13	0.39	0.41	4	1	0.30	2.38	FALSE	0.57	0.35
14	0.38	0.00	8	10	0.33	2.77	TRUE	0.31	0.56
15	0.39	0.22	5	8	0.22	2.34	FALSE	0.30	0.31
16	0.38	0.44	9	5	0.54	2.20	FALSE	0.38	0.53
17	0.39	0.17	6	8	0.37	2.56	TRUE	0.49	0.35
18	1.06	0.30	1	7	0.29	2.75	TRUE	0.26	0.41
19	0.43	0.82	5	1	0.77	2.49	FALSE	0.56	0.28
20	0.39	0.48	3	5	0.51	2.21	TRUE	0.34	0.58
21	0.39	0.34	4	9	0.20	2.65	FALSE	0.20	0.20
22	0.38	0.24	3	9	0.27	2.52	FALSE	0.24	0.26
23	0.40	0.37	5	9	0.36	2.68	FALSE	0.24	0.25

trial	wer	temperature	beam	best_of	no_speech	compression	cond_prev	patience	length_penalty
24	0.38	0.12	7	7	0.26	2.46	FALSE	0.29	0.47
25	0.37	0.61	3	10	0.34	2.29	FALSE	0.20	0.32
26	0.37	0.70	3	10	0.34	2.29	FALSE	0.43	0.38
27	0.78	0.66	3	10	0.42	2.11	FALSE	0.62	0.38
28	0.42	0.80	1	10	0.34	2.28	FALSE	0.45	0.30
29	0.66	0.63	3	10	0.46	2.02	FALSE	0.77	0.34
30	0.37	0.71	2	8	0.21	2.29	FALSE	0.66	0.43
31	0.37	0.73	2	9	0.20	2.29	FALSE	0.71	0.43
32	0.38	0.59	2	9	0.28	1.97	FALSE	0.52	0.38
33	0.38	0.94	1	9	0.24	2.28	FALSE	0.71	0.50
34	0.38	0.74	3	10	0.30	2.17	FALSE	0.63	0.27
35	0.38	0.87	2	8	0.40	2.40	FALSE	0.72	0.33
36	0.44	0.52	4	10	0.37	2.33	FALSE	0.35	0.38
37	0.37	0.59	3	7	0.23	1.80	FALSE	0.39	0.45
38	0.60	0.65	1	9	0.34	2.11	FALSE	0.44	0.29
39	0.37	0.90	2	10	0.28	1.99	FALSE	0.53	0.50
40	0.41	0.49	5	9	0.32	2.56	FALSE	0.79	0.41
41	0.37	0.70	2	8	0.22	2.22	FALSE	0.66	0.44
42	0.38	0.72	2	8	0.21	2.22	FALSE	0.74	0.37
43	0.37	0.68	3	10	0.24	2.42	FALSE	0.66	0.32
44	0.56	0.78	2	9	0.26	2.06	FALSE	0.61	0.45
45	0.56	0.57	4	7	0.30	2.34	FALSE	0.68	0.24
46	0.37	0.86	3	8	0.23	2.15	FALSE	0.27	0.41
47	0.40	0.99	2	6	0.45	2.25	FALSE	0.58	0.53
48	0.38	0.63	4	4	0.63	1.92	TRUE	0.31	0.63
49	0.38	0.74	1	9	0.39	1.74	FALSE	0.74	0.48

Appendix C

Optuna Trials - Interaction-aware

Trial status codes:

- **C**: Complete trial
- **P**: Pruned trial

C.1 Configuration A

trial	state	($\mathcal{L}$)	temperature	beam	best_of	no_speech	compression	cond_prev	patience	length_penalty
0	C	1.14	0.49	4	8	0.76	2.48	TRUE	0.45	0.50
1	C	0.64	0.25	6	7	0.79	2.39	TRUE	0.35	0.33
2	C	0.62	0.95	7	8	0.76	1.85	TRUE	0.51	0.77
3	C	0.68	0.70	2	6	0.33	2.05	TRUE	0.41	0.40
4	C	0.87	0.38	5	7	0.33	1.65	TRUE	0.45	0.65
5	C	0.64	0.37	7	7	0.66	2.61	FALSE	0.29	0.46
6	P	0.85	0.93	6	9	0.35	2.75	FALSE	0.21	0.33
7	C	0.66	0.84	8	1	0.32	2.53	TRUE	0.44	0.72
8	P	0.76	0.99	5	1	0.68	2.08	FALSE	0.68	0.27
9	C	0.61	0.61	2	1	0.73	2.03	FALSE	0.64	0.20
10	C	0.60	0.02	1	3	0.52	1.55	FALSE	0.79	0.21
11	C	0.58	0.01	1	3	0.52	1.55	FALSE	0.80	0.21
12	C	0.59	0.03	1	4	0.54	1.55	FALSE	0.80	0.21
13	C	0.58	0.08	10	4	0.49	1.80	FALSE	0.80	0.33
14	C	0.59	0.19	10	4	0.46	1.79	FALSE	0.67	0.58
15	C	0.58	0.16	10	3	0.42	1.82	FALSE	0.59	0.36
16	P	0.64	0.11	3	5	0.59	2.28	FALSE	0.75	0.42
17	C	0.59	0.35	9	3	0.22	1.70	FALSE	0.72	0.29
18	C	0.59	0.27	4	2	0.57	1.87	FALSE	0.56	0.57
19	C	0.58	0.00	8	5	0.44	1.95	FALSE	0.60	0.27
20	P	0.64	0.11	9	10	0.65	1.52	FALSE	0.74	0.40
21	P	0.60	0.15	10	3	0.43	1.72	FALSE	0.57	0.33
22	C	0.59	0.08	10	4	0.39	2.19	FALSE	0.70	0.36
23	C	0.58	0.23	9	2	0.47	1.67	FALSE	0.63	0.25

trial	state	$(\mathcal{L})$	temperature	beam	best_of	no_speech	compression	cond_prev	patience	length_penalty
24	C	0.58	0.29	9	2	0.50	1.64	FALSE	0.80	0.27
25	C	0.59	0.22	8	2	0.49	1.61	FALSE	0.65	0.24
26	C	0.58	0.49	9	5	0.60	1.92	FALSE	0.75	0.30
27	C	0.57	0.05	7	2	0.54	1.73	FALSE	0.51	0.24
28	P	0.82	0.30	7	2	0.54	1.73	FALSE	0.52	0.24
29	P	0.72	0.65	4	1	0.63	1.51	TRUE	0.53	0.51
30	P	0.84	0.05	6	2	0.37	1.98	FALSE	0.49	0.47
31	P	0.60	0.08	8	4	0.47	1.79	FALSE	0.62	0.23
32	P	0.60	0.15	9	3	0.54	1.59	FALSE	0.37	0.31
33	C	0.58	0.22	7	4	0.56	1.74	FALSE	0.68	0.36
34	C	0.59	0.44	7	3	0.57	1.70	TRUE	0.67	0.36
35	P	0.61	0.21	5	6	0.62	1.89	TRUE	0.70	0.24
36	P	0.59	0.32	6	2	0.70	1.64	FALSE	0.49	0.39
37	P	0.60	0.43	7	6	0.56	2.27	FALSE	0.41	0.44
38	P	0.83	0.23	3	1	0.28	1.77	TRUE	0.28	0.29
39	P	0.69	0.56	6	3	0.51	1.66	FALSE	0.56	0.26
40	P	0.62	0.00	8	5	0.40	2.12	FALSE	0.76	0.22
41	C	0.57	0.07	7	4	0.47	1.73	FALSE	0.73	0.33
42	P	0.59	0.14	7	4	0.45	1.57	FALSE	0.65	0.30
43	C	0.58	0.05	6	8	0.51	1.69	FALSE	0.77	0.20
44	P	0.63	0.82	7	2	0.58	1.85	FALSE	0.72	0.71
45	C	0.58	0.06	5	4	0.47	2.77	FALSE	0.68	0.33
46	P	0.59	0.19	4	3	0.53	2.00	TRUE	0.62	0.38
47	C	0.58	0.12	8	1	0.61	1.75	FALSE	0.45	0.27
48	P	0.61	0.26	7	5	0.40	2.46	FALSE	0.73	0.34
49	P	0.80	0.18	2	3	0.49	1.57	FALSE	0.78	0.22

C.2 Configuration B

trial	state	($\mathcal{L}$)	temperature	beam	best_of	no_speech	compression	cond_prev	patience	length_penalty
0	C	0.64	0.00	6	2	0.74	2.72	FALSE	0.75	0.57
1	C	0.75	0.54	5	5	0.29	2.56	TRUE	0.74	0.67
2	C	0.66	0.19	2	9	0.42	1.68	TRUE	0.31	0.38
3	C	0.65	0.10	5	1	0.20	2.60	TRUE	0.31	0.71
4	C	0.64	0.77	8	5	0.49	2.72	TRUE	0.36	0.42
5	P	0.77	0.93	10	4	0.57	2.69	TRUE	0.50	0.56
6	P	0.92	1.00	9	6	0.25	1.81	FALSE	0.55	0.36
7	P	1.16	0.28	5	8	0.65	2.46	FALSE	0.50	0.72
8	P	0.90	0.57	2	5	0.60	1.50	FALSE	0.74	0.58
9	P	0.76	0.40	10	5	0.30	2.16	TRUE	0.56	0.46
10	C	0.62	0.76	8	7	0.45	2.26	TRUE	0.21	0.26
11	P	0.65	0.78	8	7	0.43	2.23	TRUE	0.25	0.21
12	C	0.64	0.73	7	10	0.47	2.30	TRUE	0.20	0.24
13	P	0.69	0.72	7	10	0.38	2.33	TRUE	0.23	0.22
14	P	0.66	0.65	7	10	0.50	2.10	TRUE	0.43	0.29
15	C	0.62	0.87	3	8	0.36	1.99	TRUE	0.21	0.30
16	C	0.63	0.86	3	8	0.35	1.94	TRUE	0.39	0.31
17	P	0.67	0.43	3	7	0.37	1.99	TRUE	0.29	0.31
18	P	0.64	0.89	1	8	0.57	1.83	FALSE	0.64	0.48
19	C	0.64	0.66	4	3	0.69	2.41	TRUE	0.42	0.37
20	P	0.83	0.81	4	7	0.53	2.06	TRUE	0.28	0.27
21	P	0.64	0.90	3	8	0.34	1.93	TRUE	0.37	0.31
22	P	0.64	1.00	1	9	0.31	1.87	TRUE	0.20	0.34
23	C	0.63	0.83	3	6	0.42	1.75	TRUE	0.38	0.43

trial	state	$(\mathcal{L})$	temperature	beam	best_of	no_speech	compression	cond_prev	patience	length_penalty
24	C	0.63	0.65	4	6	0.42	1.64	TRUE	0.34	0.41
25	P	0.64	0.84	2	6	0.46	1.75	TRUE	0.44	0.45
26	C	0.63	0.70	6	7	0.39	1.51	FALSE	0.28	0.52
27	P	0.65	0.60	6	9	0.25	1.51	FALSE	0.26	0.79
28	P	1.00	0.49	8	7	0.38	2.20	FALSE	0.25	0.52
29	P	0.64	0.71	6	3	0.75	1.61	FALSE	0.33	0.61
30	P	0.64	0.94	9	9	0.54	2.07	FALSE	0.20	0.54
31	C	0.63	0.80	6	7	0.41	1.72	FALSE	0.27	0.25
32	C	0.62	0.76	6	7	0.33	1.73	FALSE	0.26	0.24
33	P	0.64	0.73	5	8	0.33	1.59	FALSE	0.30	0.20
34	P	0.63	0.01	7	7	0.26	1.58	FALSE	0.23	0.24
35	P	0.64	0.50	6	9	0.21	2.44	FALSE	0.34	0.34
36	P	0.66	0.61	8	8	0.46	1.68	FALSE	0.30	0.64
37	P	0.74	0.76	5	6	0.80	1.99	FALSE	0.23	0.50
38	P	0.65	0.94	9	4	0.28	2.61	FALSE	0.32	0.40
39	P	0.65	0.66	7	4	0.37	1.80	FALSE	0.80	0.28
40	P	0.75	0.70	4	5	0.31	1.87	FALSE	0.47	0.58
41	C	0.61	0.79	6	7	0.41	1.57	FALSE	0.27	0.25
42	C	0.63	0.78	5	7	0.45	1.55	FALSE	0.27	0.23
43	P	0.65	0.87	6	6	0.49	1.68	FALSE	0.23	0.27
44	P	1.12	0.53	7	8	0.40	2.30	FALSE	0.56	0.35
45	C	0.61	0.27	8	1	0.35	1.63	TRUE	0.31	0.29
46	C	0.62	0.34	8	1	0.35	1.65	TRUE	0.31	0.29
47	P	0.64	0.28	9	1	0.23	1.64	TRUE	0.32	0.22
48	P	0.66	0.24	10	1	0.33	1.74	TRUE	0.36	0.33
49	C	0.62	0.34	8	1	0.28	1.57	TRUE	0.30	0.26

C.3 Configuration C

trial	state	($\mathcal{L}$)	temperature	beam	best_of	no_speech	compression	cond_prev	patience	length_penalty
0	C	0.60	0.08	6	4	0.28	2.03	TRUE	0.54	0.52
1	C	0.61	0.03	3	6	0.38	1.54	FALSE	0.48	0.27
2	C	0.68	0.02	8	4	0.39	1.89	FALSE	0.69	0.61
3	C	1.30	0.32	5	7	0.69	1.69	TRUE	0.56	0.63
4	C	0.61	0.72	8	7	0.23	2.25	FALSE	0.42	0.74
5	P	1.22	0.11	7	10	0.60	2.17	FALSE	0.20	0.58
6	P	0.62	0.78	3	2	0.40	2.21	FALSE	0.28	0.35
7	C	0.61	0.17	4	10	0.27	1.93	FALSE	0.49	0.44
8	P	0.69	0.07	7	5	0.28	2.23	FALSE	0.30	0.69
9	P	0.69	0.06	5	2	0.30	2.59	FALSE	0.39	0.64
10	P	0.62	0.47	1	4	0.53	2.70	TRUE	0.77	0.49
11	P	0.62	0.31	10	7	0.40	1.50	TRUE	0.62	0.21
12	C	0.61	0.28	2	5	0.80	1.52	TRUE	0.51	0.35
13	P	0.74	0.96	3	1	0.48	1.91	TRUE	0.60	0.22
14	P	0.66	0.50	4	8	0.20	2.47	TRUE	0.39	0.38
15	C	0.61	0.20	6	3	0.34	1.77	TRUE	0.50	0.53
16	P	0.65	0.41	1	6	0.49	2.40	FALSE	0.71	0.29
17	P	0.90	0.60	10	4	0.35	1.99	TRUE	0.44	0.43
18	C	0.59	0.19	3	9	0.45	1.69	FALSE	0.65	0.53
19	C	0.61	0.19	6	8	0.58	1.72	TRUE	0.67	0.52
20	P	0.62	0.39	4	8	0.46	2.04	TRUE	0.78	0.78
21	P	0.63	0.20	6	9	0.58	1.71	TRUE	0.68	0.54
22	P	0.61	0.15	7	8	0.67	1.78	TRUE	0.63	0.48
23	C	0.60	0.25	5	9	0.57	1.63	TRUE	0.56	0.56

trial	state	($\mathcal{L}$)	temperature	beam	best_of	no_speech	compression	cond_prev	patience	length_penalty
24	C	0.61	0.25	5	9	0.64	1.63	TRUE	0.56	0.56
25	C	0.61	0.58	4	9	0.53	1.83	TRUE	0.55	0.68
26	C	0.60	0.40	8	3	0.74	2.12	FALSE	0.64	0.43
27	C	0.60	0.39	9	10	0.77	1.62	FALSE	0.71	0.44
28	P	0.62	0.60	2	1	0.72	2.33	FALSE	0.64	0.41
29	C	0.59	0.34	9	3	0.73	2.03	FALSE	0.59	0.49
30	P	0.75	0.35	9	3	0.74	2.07	FALSE	0.75	0.40
31	C	0.59	0.26	9	3	0.69	2.13	FALSE	0.61	0.49
32	C	0.62	0.45	9	3	0.71	2.12	FALSE	0.60	0.50
33	C	0.59	0.34	8	2	0.65	2.11	FALSE	0.65	0.46
34	P	0.97	0.28	10	2	0.65	2.32	FALSE	0.67	0.47
35	C	0.59	0.35	9	2	0.62	1.98	FALSE	0.73	0.59
36	C	0.59	0.12	8	1	0.43	1.85	FALSE	0.74	0.60
37	P	0.96	0.00	8	1	0.62	1.97	FALSE	0.74	0.60
38	C	0.60	0.11	8	2	0.68	1.89	FALSE	0.74	0.65
39	P	0.70	0.54	7	1	0.54	1.86	FALSE	0.76	0.72
40	C	0.59	0.14	9	2	0.65	2.09	FALSE	0.79	0.58
41	C	0.60	0.13	9	2	0.65	2.09	FALSE	0.79	0.59
42	C	0.59	0.24	8	2	0.62	2.23	FALSE	0.70	0.61
43	P	0.62	0.23	10	4	0.61	2.27	FALSE	0.80	0.66
44	C	0.61	0.06	9	2	0.63	2.20	FALSE	0.70	0.57
45	C	0.59	0.31	7	2	0.70	2.24	FALSE	0.71	0.63
46	P	0.68	0.33	7	2	0.55	2.26	FALSE	0.72	0.62
47	P	0.61	0.30	8	4	0.77	2.46	FALSE	0.77	0.71
48	P	0.85	0.46	8	2	0.67	2.16	FALSE	0.66	0.67
49	P	0.62	0.71	7	5	0.70	2.33	FALSE	0.69	0.58

C.4 Configuration D

trial	state	($\mathcal{L}$)	temperature	beam	best_of	no_speech	compression	cond_prev	patience	length_penalty
0	C	0.82	0.65	6	8	0.69	2.57	FALSE	0.66	0.46
1	C	0.84	0.29	1	9	0.58	1.84	FALSE	0.79	0.50
2	C	0.61	0.10	3	9	0.30	1.89	FALSE	0.68	0.58
3	C	0.70	0.38	3	2	0.64	2.73	FALSE	0.20	0.49
4	C	0.84	0.53	10	1	0.69	2.00	FALSE	0.41	0.33
5	C	0.61	0.44	5	3	0.54	2.61	TRUE	0.60	0.37
6	C	0.64	0.75	10	1	0.73	1.87	TRUE	0.26	0.67
7	C	0.61	0.67	2	8	0.24	2.03	TRUE	0.55	0.33
8	P	1.10	0.06	8	6	0.36	1.76	FALSE	0.58	0.30
9	P	0.70	0.12	5	6	0.27	2.42	FALSE	0.53	0.38
10	C	0.61	0.94	3	10	0.41	1.61	TRUE	0.80	0.79
11	C	0.63	0.97	3	10	0.40	1.56	TRUE	0.79	0.79
12	C	0.61	0.95	4	10	0.44	1.52	TRUE	0.70	0.64
13	C	0.62	0.21	1	8	0.33	2.26	TRUE	0.70	0.80
14	C	0.62	0.86	7	4	0.20	1.67	FALSE	0.80	0.63
15	C	0.59	0.01	3	7	0.47	2.17	TRUE	0.45	0.20
16	C	0.60	0.03	4	7	0.49	2.20	TRUE	0.44	0.21
17	C	0.59	0.01	4	5	0.48	2.24	TRUE	0.42	0.22
18	P	0.63	0.23	6	5	0.48	2.33	TRUE	0.35	0.21
19	P	1.06	0.01	8	5	0.57	2.10	TRUE	0.33	0.27
20	C	0.59	0.17	2	4	0.78	2.46	TRUE	0.46	0.24
21	C	0.61	0.17	2	4	0.78	2.49	TRUE	0.47	0.26
22	P	0.61	0.29	4	4	0.62	2.36	TRUE	0.40	0.20
23	C	0.60	0.01	2	7	0.53	2.20	TRUE	0.49	0.25

trial	state	$(\mathcal{L})$	temperature	beam	best_of	no_speech	compression	cond_prev	patience	length_penalty
24	C	0.59	0.16	1	3	0.46	2.49	TRUE	0.32	0.40
25	C	0.59	0.16	1	3	0.45	2.76	TRUE	0.31	0.41
26	C	0.60	0.34	2	3	0.37	2.55	TRUE	0.36	0.34
27	P	0.63	0.50	1	2	0.80	2.64	TRUE	0.22	0.42
28	P	0.91	0.22	2	2	0.62	2.40	TRUE	0.29	0.30
29	P	0.73	0.11	1	7	0.53	2.44	TRUE	0.52	0.47
30	P	0.61	0.42	3	6	0.74	2.68	TRUE	0.38	0.53
31	C	0.59	0.17	1	3	0.45	2.54	TRUE	0.30	0.44
32	C	0.60	0.16	1	4	0.43	2.53	TRUE	0.26	0.44
33	C	0.60	0.27	2	3	0.57	2.29	TRUE	0.45	0.55
34	C	0.59	0.07	1	2	0.35	2.50	TRUE	0.32	0.38
35	P	0.72	0.58	2	3	0.40	2.13	FALSE	0.26	0.46
36	C	0.59	0.32	3	4	0.46	2.58	TRUE	0.63	0.51
37	P	0.61	0.32	1	4	0.51	2.60	TRUE	0.64	0.51
38	P	0.85	0.37	3	1	0.70	2.80	FALSE	0.65	0.59
39	P	0.75	0.23	2	5	0.29	2.68	TRUE	0.61	0.50
40	P	0.61	0.41	5	2	0.65	2.70	TRUE	0.57	0.57
41	C	0.59	0.13	3	3	0.44	2.36	TRUE	0.48	0.36
42	C	0.60	0.14	1	3	0.46	2.38	TRUE	0.51	0.35
43	C	0.59	0.27	2	4	0.41	2.45	TRUE	0.20	0.46
44	C	0.60	0.26	2	4	0.41	2.45	TRUE	0.20	0.45
45	C	0.60	0.33	1	4	0.51	2.58	FALSE	0.75	0.40
46	C	0.59	0.20	2	1	0.32	2.64	TRUE	0.25	0.73
47	P	0.61	0.08	4	5	0.38	2.49	TRUE	0.24	0.52
48	C	0.60	0.38	6	2	0.55	1.94	FALSE	0.74	0.47
49	P	0.90	0.46	2	3	0.42	2.55	TRUE	0.29	0.31