André Menrath BSc BA

Breath sounds and their relationship to turn-taking in conversational speech

master's thesis

submitted to

Graz University of Technology

Supervisors

Ass.Prof. Mag.rer.nat dr. Barbara Schuppler
Univ.-Prof. Dipl.-Ing. Dr.techn. Gernot Kubin

Signal Processing and Speech Communication Laboratory

Graz, June, 2024

Affidavit

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly indicated all material which has been quoted either literally or by content from the sources used. The text document uploaded to TUGRAZonline is identical to the present master's thesis.

Signed by:	Andre Menrath
Date:	11.06.2024 16:13:34

This document is digitally signed!

This document bearing a qualified electronic signature has the same legal standing as a handwritten document in accordance with Article 25 (2) of Regulation (EU) No. 910/2014 of 23 July 2014 ("eIDAS-VO")

Validation: Information for checking the electronic signature can be found at: www.a-trust.at/pdf

date (signature)

Acknowledgements

Above all, I would like to thank Ass.Prof. Mag.rer.nat. Dr. Barbara Schuppler for her excellent supervision. Our regular meetings with intensive professional exchange were indispensable in order for me not to lose focus and to maintain a clear approach. I would also like to thank the team members of the FWF project on conversational speech, in whose monthly meetings I had the opportunity to discuss and reflect on the status of my work and received a lot of helpful advice. I would also like to thank Univ.-Prof. Dipl.-Ing. Dr.techn. Gernot Kubin, without whose willingness to supervise this thesis, at the time when Mrs Schuppler was not yet authorized to supervise the work independently, I would not have been able to do this topic.

I would particularly like to thank Caroline Hammer, without whose advice and help in organizing my time and resources I could never have hoped for such a free and easy completion. I would also like to give special thanks to Olivia Szewczykowski, who always made sure that I maintained a good balance between work and relaxation and did not get lost in too many details. I would also like to thank all my other friends who kept me motivated, encouraged me not to lose sight of this work, and believed in me to complete it despite my many other responsibilities.

Furthermore, it must be emphasized that the completion of this thesis would probably not have been possible without the possibility of educational leave offered by the Austrian state and without the unreserved support of my current employer, the Karl-Franzens University of Graz.

Abstract

This thesis aims to deepen our understanding of the impact of breath sounds on turn-taking in conversation. It provides an overview of the contexts in which audible breathing occurs and aims to identify features of breath sounds pertinent to their communicative function. This is based on the hypothesis that breath sounds are a practice that can be shaped by the speaker.

We conducted quantitative analyses using the Graz Corpus of Read and Spontaneous Speech (GRASS), where breath sounds were manually annotated. We extracted contextual, durational, and acoustic features of annotated breath sounds and employed classification algorithms to predict surrounding turn-taking labels, which were derived from annotations of Points of Potential Completion (PCOMP). These offered a fine-grained representation of the conversation, as they point out each turn-relevance place in time (TRP). For the classification tasks, we separately employed a random forest and two gradient boosting machine learning models. The model with the highest Matthews correlation coefficient (MCC) score was selected for the subsequent analysis. This involved the use of SHAP (SHapley Additive exPlanations) values to determine the importance and impact of the features. Firstly, the analysis highlighted the relation of contextual and durational features of breath sounds to turn-taking. This was used as a foundation for investigating the further impact of acoustic features, with a particular focus on those that have been proposed in the literature (e.g., the relative intensity). Additionally, the identification of further acoustic features was of interest. Our analysis demonstrated that features that are most likely related to the audibility of breathing, emerged as promising predictors of the speaker taking the turn. On the other end, backchannels (i.e., hearer-response tokens) were more frequently associated with residual breaths or longer breath sounds.

The findings support existing research and intuitive expectations, though the prediction scores were lower than anticipated, likely due to the highly spontaneous nature of the employed speech data. However, this thesis also points out methodological limitations in interpreting acoustic features, and thus establishes a foundation for future research on the relationship between audible breathing and turn-taking in conversation.

Keywords: *audible breathing, breath sounds, conversation, turn-taking, PCOMP*

Abstract (German)

Die vorliegende Arbeit verfolgt das Ziel, das Verständnis für die Bedeutung der hörbaren Atmung für das Turn-Taking in Umgangssprache zu vertiefen. Dies erfolgt auf Basis der Hypothese, dass Atemgeräusche eine vom Sprecher formbare Praxis darstellen.

Dazu führten wir quantitative Analysen mit dem Graz Corpus of Read and Spontaneous Speech (GRASS) durch, in dem hörbare Atemgeräusche manuell annotiert wurden. Wir extrahierten kontextuelle, zeitliche und akustische Features der annotierten Atemgeräusche und setzten Klassifikationsverfahren zur Vorhersage von Turn-Taking-Labels ein, welche aus Annotationen von Points of Potential Completion (PCOMP) abgeleitet wurden. Für die Klassifizierung wurden separat ein Random Forest und zwei Gradient-Boosting-Modelle zum maschinellen Lernen eingesetzt. Das Modell mit dem höchsten Matthews-Korrelationskoeffizienten (MCC) wurde für die anschließende Analyse herangezogen. Dabei wurden SHAP-Werte (SHapley Additive exPlanations) verwendet, um die Bedeutung und den Einfluss der Features zu bestimmen. Als erstes wurde die Relation zwischen kontextuellen und zeitlichen Features der Atemgeräusche und dem Turn-Taking analysiert. Dies diente als Grundlage für die Untersuchung der Auswirkungen zusätzlicher akustischer Charakteristika, mit besonderem Augenmerk auf diejenigen, die in der Literatur bereits beschrieben wurden. Dazu zählt beispielsweise die relative Intensität. Darüber hinaus war die Identifizierung weiterer akustischer Features von Interesse. Unsere Analyse zeigte, dass Features, die vermutlich mit der Hörbarkeit der Atmung zusammenhängen, sich als vielversprechende Prädiktoren erwiesen, ob der Sprecher das Wort ergreift. Auf der anderen Seite wurden Backchannels (d. h. Hörer-Antwort-Token) häufiger mit verbleibenden Atem oder längeren Einatemgeräuschen in Verbindung gebracht.

Die Ergebnisse bestätigen im Wesentlichen bereits bestehende Forschungsergebnisse und intuitive Erwartungen. Allerdings war die Genauigkeit der Prädiktionen geringer als erwartet, was vermutlich auf die hochgradig spontane Kommunikation in den verwendeten Daten zurückzuführen ist. Die vorliegende Arbeit demonstriert zudem die methodischen Limitationen bei der Interpretation akustischer Features und legt damit den Grundstein für zukünftige Studien zum Zusammenhang zwischen hörbarer Atmung und Turn-Taking in Umgangssprache.

Contents

1 Introduction

1.1 Personal motivation

If it were up to me and my personal experiences to name a situation in which breath sounds have a crucial function, I would point to their importance in making music together, especially in classical chamber music. As a professionally trained pianist, I am used to the fact that I cannot always see all my fellow musicians due to the constraints of my instrument. Audible breathing helps me first and foremost to coordinate and synchronize joint entries. In addition, accentuated entries of a single musician seem much more convincing to me if the whole body carries them along naturally, which includes not only movements that may seem more extensive than absolutely necessary for the execution of the music, but also strong and audible breathing. The scenario of making chamber music may be a niche one, and therefore not an important example for many, but spontaneous conversations via speech affect us all, whether with friends or at work, and are an indispensable part of our everyday lives.

Since I started working on this thesis, the situation arose again and again that I had to explain the topic to friends and colleagues. I told them that it was about getting an insight into the meaning of breath sounds in spontaneous conversations in Austrian German. After a short period of reflection, most of them told me that it was completely natural for them that these breath sounds also had a function in communication. Getting this feedback over and over again, I was even more surprised to find out through my literature research that what I wanted to address in my thesis was somehow still a quite unwritten chapter in science.

1.2 Aim of this thesis

The primary objective of this thesis is to gain a deeper understanding of the impact of audible breathing on turn-taking in conversation. This is achieved through an investigation of the contexts in which audible breathing occurs and the identification of acoustic features that may be pertinent to their communicative function. Our experiment series is designed to provide more insight into the limited number of acoustic features that have been discussed in the literature, specifically the relative intensity (see section 1.4). In addition, we are particularly interested in determining whether certain other more fine-grained acoustic features could assist in the depiction of communicative information derived from breath sounds. Furthermore, this thesis aims to provide an overview of the frequency of occurrence and the variations of breath sounds in different turn-taking scenarios, in order to compare them with results from the literature. To this end, we conduct quantitative analyses on the Graz Corpus of Read and Spontaneous Speech

(GRASS) (Schuppler et al., 2017). In GRASS, no respiratory activity is available; instead, breath sounds were manually annotated along with the speech. This thesis limits itself to scenarios where an audible breathing is present and does not take into account the communicative function that the sheer presence of an audible breathing already might contain.

1.3 State of research on the meaning of breathing in speech

In order to review the current state of research, we will also consider the results of recent studies that have examined respiratory activity in conversation, a topic that has been studied far more extensively than the acoustic domain of breathing. The respiratory activity is usually recorded via multiple belts wrapped around the speakers body (e.g., Heldner et al. (2019)). This continuous signal is as closely aligned as possible with the volume of air in the speaker's lungs. A possible relationship between the two domains could be that a short and at the same time strong inhalation or exhalation probably leads to a louder breath sound.

Differences between audible breathing and respiration A respiratory signal, on the one hand, and an audible breathing, on the other, can both contain information that the other cannot cover. For instance, breath holds are unlikely to be accompanied by sound, or audible breathing can be acoustically varied independently of the respiration activity. Furthermore, the annotations of audible breaths are not continuous; they have time intervals where they begin and end, which are interleaved with speech and other noises. In contrast, the respiratory signal is a continuous signal. When comparing findings derived from the analysis of audible breaths with those derived from respiratory activity, it is essential to consider these factors.

In order to keep these two distinct, yet closely related domains properly separated, in this thesis the term *inhalation* or *exhalation* will be used in reference to respiratory breathing activity, while the term *audible breathing*, and the suffixes *noise* and *sound* refer to the acoustic domain (e.g., *exhalation noise* or *breath sound*). Additionally, the terms *inbreath* or *outbreath* will be employed solely to denote audible breathing.

Breath sounds and speech pauses Breath sounds are sounds produced by the speaker's vocal tract. From this perspective, they share the same category as spoken words, singing, laughing, or smacking and coughing sounds. Until recent years, however, breath sounds were given little attention in speech analysis and were in fact often treated as pauses. Speech pauses that contain breath sounds were generally not distinguished from pauses that do not contain audible breathing. Evidence concerning the importance of including breath pauses as a separate category in the analysis of spontaneous conversations is increasingly corroborating. In a dataset of dyadic conversations in Danish, it was shown that the type of pause that most often has a communicative function is the breathing pause (Navarretta, 2020).

1.3.1 Presence and absence of breath sounds

Before we address the possible variety of meanings of breath sounds, it is helpful to have a general understanding of why, by their very presence, they are fundamentally important for the processing of speech. This is applicable to at least two cases: human perception, on the one hand, and applications such as automatic speech recognition, on the other. As early as 1992, it was shown that excluding filled pauses (e.g., *uhm* or *ah*) and breath sounds resulted in a substantial proportion of errors in automatic speech recognition (Butzberger et al., 1992). In speech synthesis, on the other end, experiments showed that modelling breath sounds improved the naturalness of speech perception. Moreover, a study suggested that when individuals listen to syntactically generated sentences in which breath sounds are not neglected, their ability to remember these sentences improves (Whalen et al., 1995). Their findings suggest that audible breathing may increase the listener's attention.

1.3.2 Potential functional diversity

Schegloff (1996) already posted that audible inhalation is more than a physiological prerequisite for imminent speech. He notes that breath sounds are practices that can be shaped. Fundamentally, speakers can control whether the breath sound can be heard more or less. Focusing on turn management, he hypothesized that a deep audible inhalation might be an indicator for an extended spate of talk to come, and thus suggested that the various breath sounds contribute to building units. He further points out that in spontaneous speech breathing and associated breath sounds can occur syntactically anywhere.

While one might assume that the functions of breath sounds are language-independent, they have been shown to have speaking-style dependent functions in Korean. By comparing the phonetic profile of Korean speech in formal and informal situations, Winter and Grawunder (2012) showed that breath sounds differ significantly in these two. In formal situations, associated with more politeness, louder and more noticeable inhalation sounds are used. Since the data processed in this thesis, however, contained only dialogues between people who have known each other for years, thus resulting in an informal speaking style, the formal style remains to be studied. Nevertheless, for our work it is helpful to keep the spectrum of possibilities in mind for interpreting the results.

1.3.3 Turn-taking management

In spontaneous conversations, different speakers often pass into one another seamlessly. How it is possible for these transitions to be so smooth has been the subject of linguistic research for a long time (e.g., Ishii et al., 2013; Rochet-Capellan and Fuchs, 2014; or Hara et al., 2018). Which and whether breathing also plays a role in these transitions, however, has not yet been adequately elucidated.

Ishii et al. (2014) investigated whether it is possible to predict the next speaker in spontaneous multiparty conversations with four speakers using features derived from the respiratory signal. Using predictive models, they found that in situations where the speaker changes, the time interval between the previous utterance and the onset of inhalation as well as the slope and the

duration of inhalation were the strongest indicators of who would speak next. As a reference point in time when their model makes a prediction, they used 350 ms before the next utterance. Based on their observations, they concluded that the next speaker takes a bigger breath towards speaking in turn-changing than listeners who will not become the next speaker. Since it is reasonable to assume that a bigger breath is most likely to be coupled with a more audible breath sound, this hypothesis has also been a starting point for this work. Also, of interest to this work are their findings that whenever a speaker inhales during speaking, the inhalation is compressed in time. This effect was confirmed by Rochet-Capellan and Fuchs (2014), but could not be shown in other studies (Włodarczak & Heldner, 2020).

One of the most common situations in spontaneous conversations are turn-taking events during a question-answer sequence. Torreira et al. (2015) studied the functions of breaths in this turn-taking context using a speech corpus containing several, 45-minute-long, dyadic conversations between Dutch male friends. They found that inhalations preceding a response typically begin promptly after the interlocutor has finished asking a question, and that they usually occur together with significantly delayed responses. Furthermore, short answers are predominantly executed on residual breath, whereas longer answers are more likely preceded by inhalations.

Between 2015 and 2020, Włodarczak and Heldner (2020) probably carried out the most comprehensive and sequential studies on breathing in conversation. Their research includes the relationship between breathing and turn-taking in conversations. However, they primarily focus on respiratory activity, which provides considerably more information than just the acoustics of breathing, such as breath holds. For a pause to be perceived by a listener, it must be longer if it includes audible breathing (Ćwiek et al., 2016; Heldner & Włodarczak, 2016). This suggests that an audible inbreath after a section of speech acts as a cue that the speaker intends to continue speaking, thereby holding the turn. The same authors later discovered indicators of the communicative functions of breath sounds through the comparison of *silent pauses* with and without audible breathing (Ćwiek et al., 2017). In their résumé Włodarczak and Heldner (2020), proposed a categorization of turn-taking events that combines the criterion of speaker change with whether the original speaker inhales before producing the next utterance.

1.3.4 Preference status

Different types of inbreaths could possibly serve as content indicators, such as indicating the preference status of the speaker. While such a relationship between preference and breath sounds has not yet been proven in face-to-face conversations, using telephone calls in English Torreira et al. (2015) found that disagreeing responses are more likely to be preceded by an audible inbreath than consenting responses. This may be due to at least two reasons. Firstly, the study with the face-to-face conversations was based on the respiration signal, whereas for the study based on the telephone conversations only the audible inhalation sounds were available. Secondly, the two settings of the conversations differ substantially for the interlocutors: if one is not visible to their conversational partner, it is reasonable to argue that other semiotic signals in speech are more important.

In this thesis the GRASS corpus is used which does not contain respiration data. However, it is essential to note that the conversations took place face-to-face in the same room. It is

important to consider that the acoustic environment during the conversation, as perceived by the speakers, and the recorded audio used for analysis, were much better than the quality of telephone signals. Furthermore, the interlocutors could see each other well and had clear visual cues about each other's breathing. Nevertheless, it is not possible to make a direct comparison of the aforementioned results with those presented in this thesis, as the data utilized in this thesis has not included specific information pertaining to the preference status.

1.4 Features of breath sounds

The absence of manifold acoustic features of audible breathing highlights the necessity for further investigation and analysis in this area, and emphasises the novel and original contribution of the present work. Nonetheless, some baseline work has been carried out.

Relative intensity Despite being frequent nonverbal vocalizations in spoken communication, inhalation sounds have received little attention in research regarding their acoustic features and variability (Trouvain et al., 2020). Trouvain et al. examined the intensity of breath sounds. They describe inbreaths as typically soft in intensity. Furthermore, they introduce a method for calculating a feature derived relatively, based on the difference in loudness between the breath sound and the two seconds surrounding it. This is achieved by subtracting the average of the decibel values taken two seconds before and after the pause from the decibel value of the entire breath sound.

Spectral Centroid Trouvain's study also provides insight into another feature - the *spectral centroid* of a breath sound, also known as the *centre of gravity*. When compared to the aspiration sounds of plosives, it is found that inhalation sounds typically have a spectral centroid below 2kHz, while aspiration sounds are typically above 2kHz and have a significantly higher intensity.

Formants Additionally, the study compares the vocal formants of speech between aspiration and inbreath sound, demonstrating that the formant values for aspiration noises exhibit more variability than those for inbreath sounds.

In their fundamental explorative analysis, Trouvain et al. (2020) employed the duration of the breath sound as a primary feature, in addition to the intensity. They highlight the extremely variable duration of the breath sound, noting that aspirations are typically much shorter than inbreath sounds.

2

Materials & Distribution of breath sounds

2.1 GRASS corpus

2.1.1 Recording scenario

In this thesis, meanings of breath sounds have been studied using the conversational part of the Graz Corpus for Read and Spontaneous Speech (GRASS) (Schuppler et al., 2017). It contains 19 conversations of two adults each, who have known each other for several years and have spent a large part of their lives in Graz. The different gender combinations in the pairs are distributed almost evenly. The conversations last approximately one hour while no experimenter was present. The participants were free to talk about anything they wanted, although they were given some incentives by some pictures to start the conversation. The audio sources used for this thesis were recorded with headset microphones (AKG HC-577L). Although it is an unnatural and at the same time unfamiliar situation to have a conversation with headsets in a studio environment, most of the conversations are nevertheless quite natural due to the sheer duration of the recordings, the fact that all these people already knew each other very well (couples, family members, friends and colleagues) and that they were neither observed nor eavesdropped on during the time of the recording.

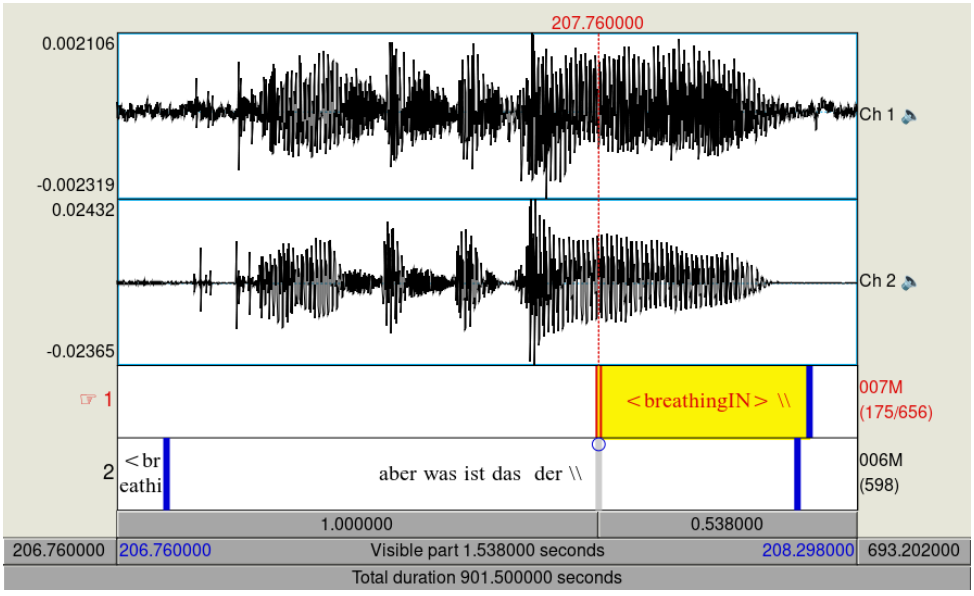


Figure 2.1: Interference by the interlocutor in overlapping speech. Ch 1 shows the speaker 007M and Ch 2 shows the speaker 008M. Note that the amplitude scaling is different.

The Signal-to-noise ratio (SNR) differs a lot between the different speakers (Schuppler et al., 2017). The lowest SNR was 35.8dB and the highest was 52.8dB. This is partly because the positioning of the headset microphone or the average volume of the speaker varies, and partly because the preamplifier was set individually for each speaker. Although the two participants were sitting at the corner of a table and their two chairs were initially more than one meter apart, the speech of the other person is still slightly hearable on the headset microphone. Also, since speakers were not observed, they tended to move closer to each other during the one-hour conversations. Therefore the interlocutor plays a crucial role as an interfering signal, especially with breath sounds, which are usually on average many times quieter than other components in speech. Figure 2.1 shows an example where the speaker's breathing sound has about the same amplitude as the other person's speech. This makes it impossible to extract reliable acoustic features for breath sounds that overlap with the speech of the conversation partner without prior suppression of the background noise. Attempts to suppress the interlocutor by applying Noisereduce (Sainburg, 2021), a time-varying non-stationary spectral noise gating algorithm, did not yield to sufficiently useful material. For this reason, investigations using acoustic features that are performed together on overlapping and non-overlapping breath sounds are very sensitive and must be interpreted with caution. Therefore, when interpreting breath sounds with overlapping speech, it is important to bear in mind that these features may partly reflect the activity of the interlocutor.

2.1.2 Orthographic transcriptions

For all conversations orthographic transcriptions were created manually with PRAAT (Boersma & Weenink, 2001). The speech was divided into annotation chunks, related to utterances, with a maximum length of four seconds. In addition, these transcriptions also contain very detailed annotations of laughter and other speaker noises like backchannels (e.g., *hm*), fillers (e.g., *eh*, *ah*, *oh*), broken words, overlapping speech, disfluencies, as well as inhalation and exhalation noises (Schuppler et al., 2017).

2.1.3 Turn-taking subset

While working on this thesis, 300 seconds of speech per speaker in 14 conversations were corrected as part of another project to have better breathing annotations available. These improvements have been made in the course of annotating *communicative functions*: *Inter-Pausal Units* (IPU) and *points of potential syntactic completion* (PCOMP), in separate tiers. The characteristics and definitions of the aforementioned annotation systems are presented and discussed in detail in section 2.4.

2.1.4 Definition of the term *Pause*

It can be difficult to determine when to use the term *pause*. It is essential to consider the annotation's purpose, which, in this thesis, is the investigation of turn-taking behaviour (Schuppler & Kelterer, 2021). When annotating the *inter-pausal units* (IPU) in the GRASS speech corpus,

annotators were instructed to divide into two segments solely if the silence within the same speaker was at least 150 milliseconds long. This duration was chosen to maintain consistency in the annotated data, which is difficult to achieve even with detailed instructions on what to annotate, and considering the goals of later analyses. In order to be consistent with other studies carried out on the GRASS corpus, when referring to a speech pause in this thesis, this means that a speaker does not produce any speech for at least 150 ms. Włodarczak and Heldner (2020) and other researchers have also used 150 ms as the minimum gap between two holds in their annotations of inter-pausal units in their studies, so it seems to be a common practice.

2.2 Distribution and state of breath annotations in GRASS

In order to extract acoustic features from all the breath sounds, it would be particularly necessary for them to be separately annotated (i.e., for the beginning and end of each breath to be known explicitly). Furthermore, to be able to analyse and draw conclusions about the communicative functions of audible breathing, it is crucial not to work with a dataset that omits a substantial portion of the breath sound annotations.

2.2.1 Orthographic tier – whole data

In the orthographic tier of the GRASS annotations, more than about half of all transcribed breath sounds are not explicitly time-aligned; instead, they are chunked together mostly with speech or other breath sounds. Table 2.1 shows how often which type of breathing annotation occurs in the entire corpus.

Table 2.1: Number of annotations in the orthographic tier of the whole GRASS corpus.

	Quantity
Total annotation chunks	53710
Annotated breath sounds	9252
Annotations with breath sounds	8981
Annotated inbreaths	8048
Annotated outbreaths	1204
Separately annotated breath sounds	4940
Separately annotated breath sounds without overlap	2637

In the breath sound annotations where the boundaries of the transcribed breath sounds were not explicitly set in time, the breath sounds are directly attached to speech in almost all cases, to laughter in a few cases, and to other noises in exceptional cases. In this group, with a total of 4041 annotations, about 45 percent begin with an inbreath and less than 3 percent with an exhalation noise, whereas 10 percent end with an inbreath and about 7 percent with an outbreath sound. It is noticeable that in the cases where the breath sound is found at the end of the annotation, these annotations consist on average of less than two words. These are, to put it another way, mainly very short utterances, most likely backchannels. In contrast, the annotations in which a breath sound is annotated in the middle, meaning that it is surrounded

by speech (the case for 569 instances), consist on average of 10 words and are 3 seconds long. It is also clear that there are considerably (about eight times) more inbreath annotations than outbreath annotations. Figure 2.2 shows how many audible breaths were annotated in speech pauses, separately for each speaker.

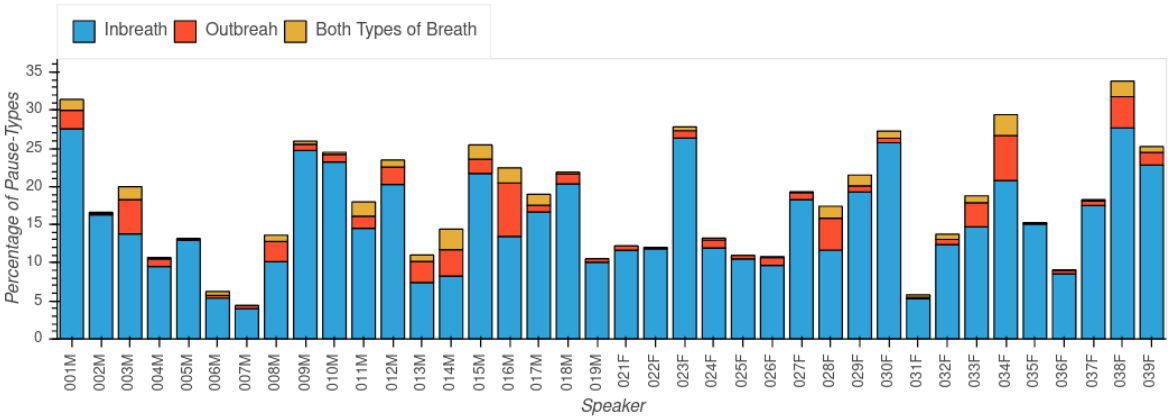


Figure 2.2: Distribution of pause types per speaker based on the annotations of the whole corpus. Note that these numbers are very imprecise due to the insufficient annotation of breath sounds. Silent pauses are at least 150ms long.

Although the annotations of the GRASS corpus were reviewed by one annotator other than the one who made the first transcription, who were all instructed in the same way, the different frequency of occurrence of the annotated breath sounds, as Figure 2.2 shows, could also be not only due to the different speakers, but also to the different annotators or the different signal-to-noise ratio. It must also be considered that breath sounds are usually much quieter than speech and transcribing and correcting them was probably not the main focus for the annotators.

The annotators might have tended to annotate the breath sounds separately or not depending on certain situations in the course of the conversation (e.g., possible categories or meaning of breath sounds). To compensate for this possible bias, we manually checked at least 5 minutes of each conversation for this work, with the goal of having at least about 30 consecutive, separately annotated breath sounds from each speaker. During this review some conversations turned out to have an above-average number of breath sounds that were not annotated at all. However, most of these corrections were not used in the subsequent experiments, as corrections of breath sounds at other time points in the conversations were added in the course of another annotation process while working on this thesis.

2.2.2 Subset of turn-taking annotations

In the course of the annotations of the IPUs and PCOMPs (see section 2.1.3) numerous missing breath sound annotations have been added, and the starting and ending alignments of the breath sounds have been adjusted. Within the IPU tier, the breath sounds that are attached directly to a speech utterance are mostly missing. The PCOMP tier does not miss such breath sound annotations, although groups of consecutive breath sounds without a gap are treated as a single annotation. Nevertheless, time aligned breath annotations for those cases have been added to

the word-tier within the time intervals where the IPU and PCOMP annotations were made. The PCOMP tier also misses breath sounds that are rather isolated from speech utterances. This is due to the annotators instructions where to focus on utterances and what is directly attached to them, not on speech pauses. For the orthographic tier, however, there was no such instruction, so it can be assumed that there is less bias in the likelihood of a breath sound being annotated whether it is close to speech or not.

2.2.2.1 Analysis of audible breath occurrences

The Table 2.2 shows the total quantity of available breath sound annotations in this reduced set of data within the PCOMP tier. Comparing the number of breath sound annotations to the entire dataset in Table 2.1, a large difference in the ratio of the number of inhalation to exhalation annotations is noticeable. In general, in the subset there are more than four times as many annotations of audible breathing per time as in the orthographic tier of the whole corpus. Within the timeframe of the subset, the orthographic tier contains 2010 breath sound annotations, after having been corrected during the annotation process of the communicative functions. When compared to the total number of annotations in the PCOMP tiers, which include instances of multiple breath sounds within one annotation, there are still at least 269 breath sounds that are not included in the PCOMP tiers.

Table 2.2: Number of breath sound annotations within the PCOMP tiers.

Breathing Annotations	Quantity	Quantity without noise, smack or laughter
Total	1586	1293
In	864	652
Out	579	529
In-Out	71	54
Out-In	60	48
Others	22	10

2.2.2.2 Classification of pause types based on audible breath occurrences

The individual percentage distribution of pause types per speaker based on this subset of annotations is very different from the entire set of annotations. Compared to a similar analysis by Trouvain et al. (2020), the percentage of silent pauses is quite similar. However, there are clearly more audible exhalations (annotated outbreath sounds) in speech pauses in the present dataset. These differences could also be explained in part by the problem of distinguishing between aspiration and exhalation, as discussed in section 2.2.3.

Figure 2.3 compares how often breath sounds occur during pauses in speech and shows that there are noticeable differences between the individual speakers. Speaker 014M, for instance, gives the impression of being a bit sniffly. This, of course, contributes to the louder breath sounds in general, which are then much easier to recognize. Given that all recordings of the GRASS corpus were made in February, the probability of voice-influencing diseases is considerable. For speakers 031F and 032F there are almost no outbreath sounds annotated at all. Random checks

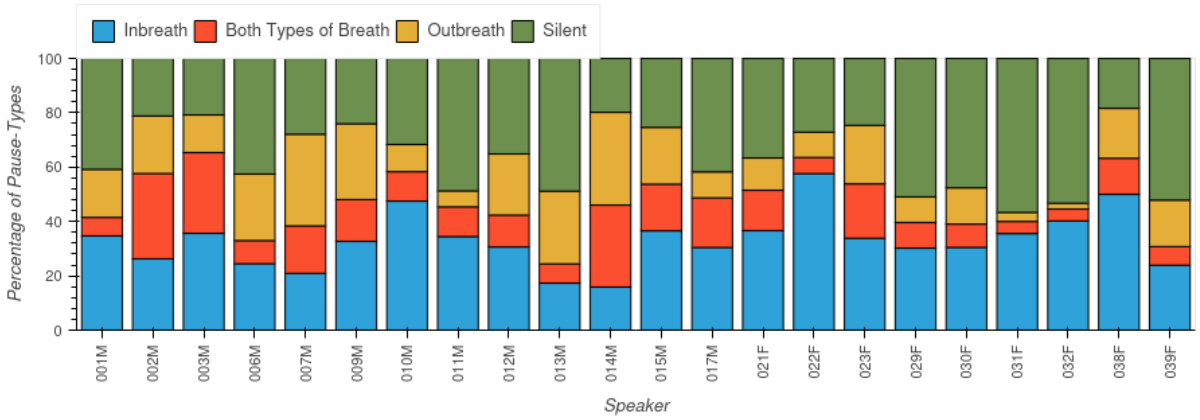


Figure 2.3: Distribution of pause types per speaker in the annotation subset of communicative functions, where orthographic breath transcriptions were also completed and corrected. Silent pauses are at least 150ms long.

of short time intervals suggest that the reason for this was the differences in attentiveness of the respective annotators (and their equipment quality) and not particular characteristics of the speakers.

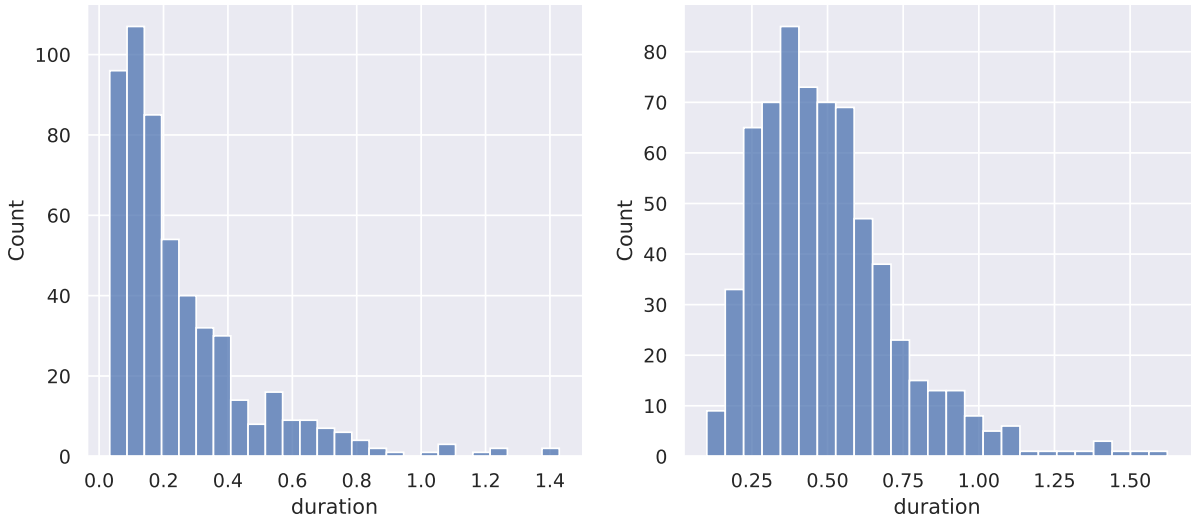
2.2.2.3 Further improvement of the breath annotations

An automatic annotation of breath sounds using classifiers or detectors, for example as employed by Fukuda et al. (2018), would have been a conceivable method to address the issues of the subjective biases of the human annotators. Especially for already known, but not separately annotated breath sounds, good results could be expected. Furthermore, the method allows for the detection of breath sounds that may have been overlooked during the manual annotation procedure. This approach was not pursued, however, as it would have distracted too much from the actual aim of the thesis and the current dataset appeared sufficient for the methodology of this thesis.

2.2.3 Outbreath vs. aspiration

As stated in the previous paragraph our dataset contains remarkably more audible exhalations than those of others. When randomly examining the annotation in the subset of communicative functions, many exhalation noises were noticed, which are arguably on the borderline of whether they should even be defined as such. In this dataset the majority of the annotated exhalation noises are very short (i.e., shorter than 0.2s). This distribution is shown in Figure 2.4(a). Note that in contrast the distribution of inbreath durations exhibits a markedly different pattern, as illustrated in Figure 2.4(b). It is nearly Gaussian-like, with a mean of slightly below half a second.

Whenever the exhalation noises are included in the analysis, it is important to keep in mind that it is often almost impossible to distinguish them from aspiration releases of plosives, or that they are often so low in volume that they can no longer be clearly recognized. Moreover, these were not the main focus of the annotators.



(a) Outbreaths (b) Inbreaths

Figure 2.4: Distribution of the duration of annotated breath sounds in the subset of communicative functions.

Figure 2.5 shows a Praat screenshot of an exhalation noise that cannot be recognized visually by the signals amplitude. During the analysis of the example, it is important to note that the author was uncertain about the presence of a distinct exhalation noise in this particular case.

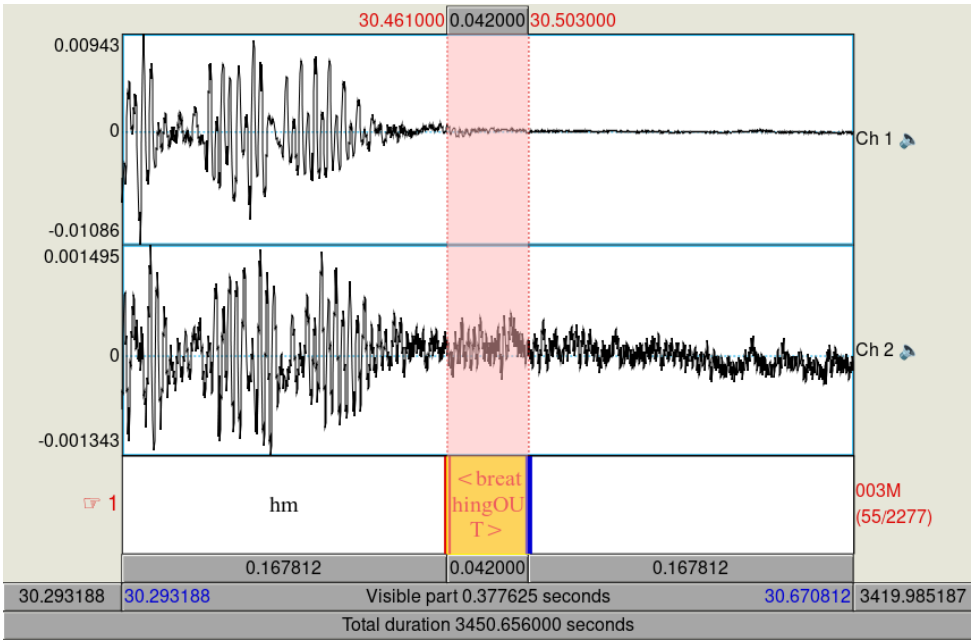


Figure 2.5: Example of a very short exhalation noise that cannot be recognized visually, and also acoustically it is probably a borderline case.

Another example where it is not easy to decide whether the noise should be labelled an aspiration or a exhalation is shown in Figure 2.6. The annotated *<breathingOUT>* occurs immediately following the plosive [t] in the utterance *was? ernsthaft?* (*what? really?*). The interlocutor is also speaking during the annotation of the exhalation noise, making it even more difficult to identify it as either aspiration or outbreath.

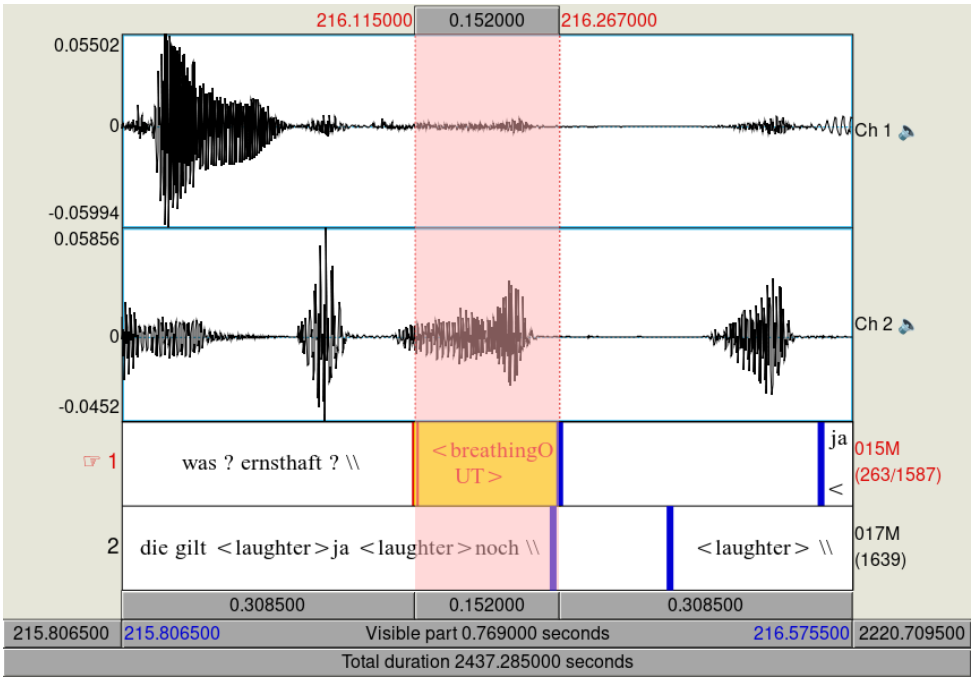


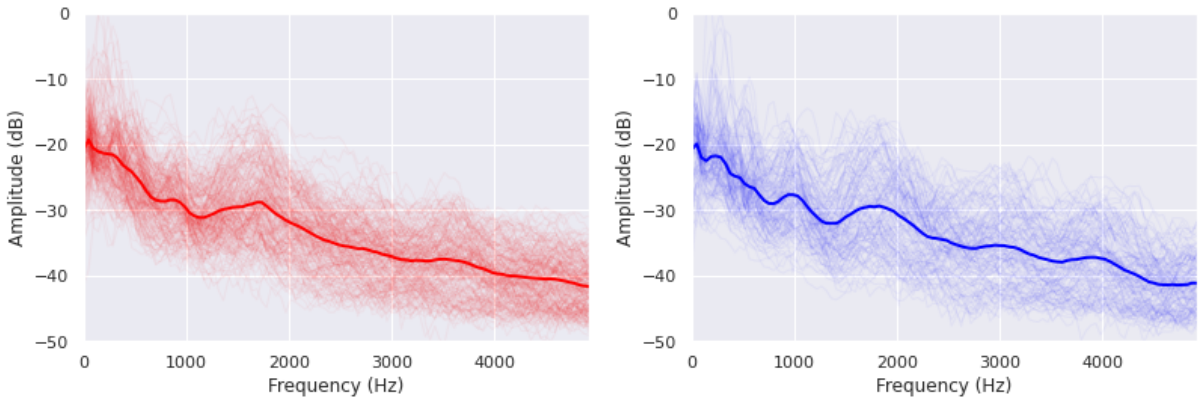
Figure 2.6: Borderline case between an aspiration and an outbreath directly after speech.

2.3 Spectral characteristics of breath sounds

2.3.1 Inhalation spectra

Werner et al. (2023) presented graphs that show the spectra of male and female inhalation noise. Reproductions of their graphs using inhalation noises from our dataset, but only samples that do not overlap with the interlocutor's speech, can be seen in Figures 2.7(a) and 2.7(b). The overall mean of all spectra calculated from GRASS shows a generally decreasing slope from low to high frequencies, which is of quite a similar shape as theirs. In the mean curves for both sexes, the prominent peaks located slightly below 2kHz are also clearly visible in our figures, although a peak around 500Hz is not observable in none of our data. Instead, the mean inhalation spectrum of all females shows a prominent peak at about 1kHz, which is not present in their figures. This peak is also visible in our mean inhalation spectrum of men, but it is significantly weakened.

The spectra presented by Werner et al. (2023) exhibited minimal variation, in stark contrast to our dataset. In general, the inhalation spectra from our individual samples showed a greater degree of diversity. This discrepancy can be attributed to several factors. Firstly, Werner et al. (2023) used data from two distinct female and male sources, both containing semi-spontaneous speech. In contrast, our data collection involved spontaneous conversations of pairs, featuring a mixed combinations of sexes. Secondly, the annotation methods differed between the two studies. Werner et al. (2023) employed an automated approach, detecting the onset and offset of noise in the audio signal for inhalation sound annotation. Conversely, our annotations were conducted manually by human annotators, allowing for a broader detection range compared to algorithms that may exhibit a tendency to identify similar inhalations. Furthermore, the datasets used by Werner et al. (2023) exclusively included inhalation sounds directly attached to speech, potentially overlooking breaths that are farer off. This restriction in their dataset composition



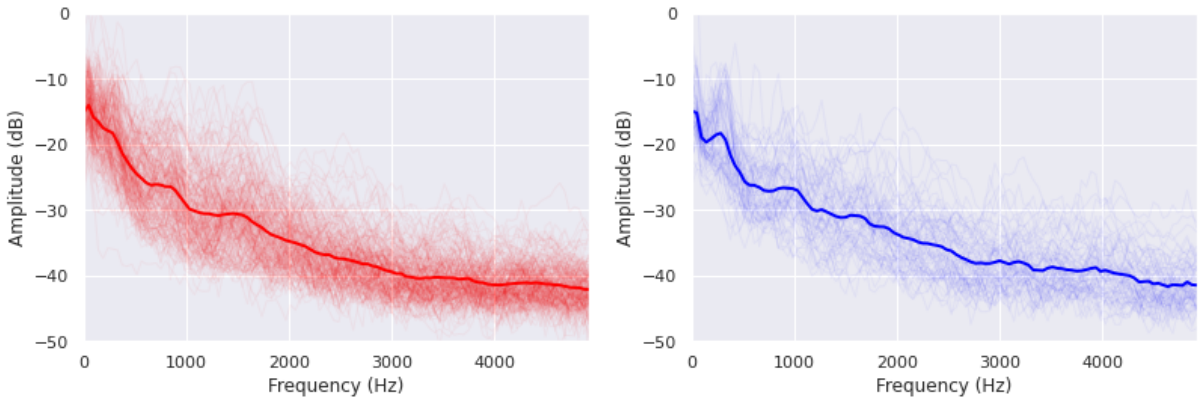
(a) Male (b) Female

Figure 2.7: All human inhalation spectra. The average spectrum is overlaid in bold.

could contribute to the observed differences in variation between our figures and theirs. An additional reason why our data shows more variation in the spectra could be that some speakers had a sniffly condition as described in section 2.2.2.2. Additionally, the relationship between semi-spontaneous and spontaneous speech and the shape of breath sounds is not yet known. It may be that even more spontaneous speech leads to more variation in breath sounds.

2.3.2 Exhalation spectra

Figures 2.8(a) and 2.8(b) show the exhalation spectra, computed using the same algorithm as the inhalation spectra presented in the preceding section. Compared to the inhalation spectra the exhalation spectra possess a similar overall shape, but there are two significant differences: the exhalation noises show less intense spectral peaks on average, and the mean spectral intensity curve rises more strongly towards the lower frequencies.



(a) Male (b) Female

Figure 2.8: All human exhalation spectra. The average spectrum is overlaid in bold.

2.4 Turn-taking annotations

In order to investigate the relationship between audible breathing and turn-taking, this thesis utilizes the turn-taking annotations available for the GRASS corpus. The annotation system used for these was designed with the aim of being suitable for both quantitative and qualitative analysis of conversation (Kelterer & Schuppler, subm.). Its main key concepts relevant to this work are as follows: (1) the annotations are based on what happens next in the conversation rather than on interpretation, (2) all continuously occurring utterances can be labelled, and (3) the set of labels is of reasonable size to be used in automated classification methods. To achieve this, turn-taking was annotated on two separate tiers: *Inter-Pausal Units* (IPU) and *Points of Potential Completion* (PCOMP), following the conversation analysis principles outlined by Ogden (2024).

2.4.1 Annotations of Inter-Pausal Units

Subsequent annotations of *Inter-Pausal Units* IPU are segmented at least by a silent pause of 150 ms. This threshold was chosen because it is longer than most voiceless plosive durations (Kelterer & Schuppler, subm.). The IPU annotation comprises three main categories: (1) the speaker continues speaking, (2) the interlocutor takes the turn, and (3) hearer response tokens of any type (e.g., backchannels or acknowledgements) that do not interrupt the other speaker's turn. These categories are also clearly reflected in the most upper levels of the decision tree for assigning a single IPU label, as described by Kelterer and Schuppler. These categories are further subdivided into seven specific labels, which are, however, too specific for our particular intended use here. Additionally, the IPU-tier contains time-aligned annotations for breath sounds, laughter and other non-speech sounds. When a breathing is timely not isolated from speech, it is mostly not being annotated as a separate chunk. Consequently, the IPU tier would be of limited use in this thesis.

2.4.2 Annotations of Points of Potential Completion

The *Points of Potential Completion* (PCOMP) annotations, which are available in GRASS, are on a separate tier from the IPU annotations and consist of eleven labels. It is important to note that the PCOMP annotation criteria only rely on syntactic, not prosodic, completeness. Although a PCOMP represents a point in time, the PCOMP annotations are made as time intervals, where the interval begins when the utterance starts at the beginning of a turn or immediately after the end of a preceding annotation. The decision tree (see Figure A.1) for assigning a single PCOMP label, as used by the annotators working on the extension of the GRASS corpus, differs significantly from that for IPU labels. The question of which speaker is continuing to speak is typically asked at the last stage of the decision tree. However, since PCOMP annotations combine syntactic and turn-taking criteria, they can ultimately be grouped in the same three main categories, namely change, hold, and hearer-response-token (HRT) as the IPU annotations. One limitation, however, is that the label `question` only has a tendency towards turn-taking, but does not reflect whether the question was actually answered and thus the turn was taken.

2.4.2.1 PCOMP label set

It is of utmost importance to acknowledge that the PCOMP labels, as mentioned above, are based on what actually happened in the conversation, rather than on the speaker's intention or our interpretation of it. In other words, they are forward-looking. An exception is the label `question` which is applied whenever the current syntactic structure is a question, regardless of which speaker continues. In contrast, on the IPU tier, the label `question` is only used on questions to which the speaker has actually responded and thus taken over the turn. It is not necessary to fully be familiar with the definitions of all PCOMP labels in order to be able to understand the experiments presented in this thesis, because most of them are not utilized directly, but in a grouped way (see section 3.2). A detailed overview of all labels and additional keywords is given in Table 2.3. Within the subset of PCOMP annotations that are directly preceded or followed by a breathing, approximately five percent of these annotations are marked with the uncertainty label. Furthermore, slightly more annotations have multiple labels.

2.4.2.2 Breathing and other additional annotations in the PCOMP tier

In addition to the eleven PCOMP labels, the same tier contains time-aligned annotations of all breath sounds, laughter and miscellaneous noises. In the case of breath sounds, however, it is necessary to consider that when several inbreath or outbreath sounds of the same speaker occur immediately after each other, they are usually annotated in a single time interval, with the label containing all of them. For example, an inbreath immediately followed by an outbreath would be annotated as a single interval with the annotation `<breathingIN><breathingOUT>`. There are also three additional events that can be included in a single breath sound annotation, see Table 2.2. For example, an inbreath that is interrupted by a smack is annotated as `<breathingIN><smack><breathingIN>`. Smack sounds are the most common, occurring in 143 cases ($\approx 9\%$). Miscellaneous noises, such as microphone wobble, occurs in 65 cases ($\approx 4\%$). Finally, the breath sound annotation occurs together with laughter in 65 cases.

2.4.2.3 Distribution of labels in the PCOMP tier

Figure 2.9 presents a histogram that illustrates the distribution of occurrences of all labels in the PCOMP tier of the entire dataset. The label `hold` is by far the most frequent, closely followed by breath sound annotations. The number of inbreath annotations is distinctly greater than that of outbreath annotations, which is logical given that exhalation is a natural phenomenon that occurs during speech at any time. The proportion of `hrt` annotations is half of that of `hold` ones, while the annotations of `cont`, `part`, and `change` are approximately one-third as numerous as `hold`s. The labels `question`, disruption and `laughter` occur with even less frequency. The remaining labels in the set are not included in the figure, as they occur with such rarity, occurring less than one hundred times in our data.

Table 2.3: List of PCOMP labels and their definitions, taken from Kelterer and Schuppler (subm.).

PCOMP label	Definition
`hold`	same speaker continues speaking after the PCOMP by starting a new sentence
`cont`	same speaker continues speaking after the PCOMP by continuing the same sentence with the addition of increments
`change`	other speaker continues speaking after the current speaker reaches a PCOMP
`part`	discourse particle uttered after a PCOMP or at the beginning of a turn
`q-part`	question particle (tag question) that transforms a declarative utterance into a question or is used to elicit some kind of listener feedback (e.g., a backchannel)
`question`	syntactic and/or prosodic question
`hes`	hesitation particle uttered after a PCOMP or at the beginning of a turn
`coll`	current speaker collaboratively finishes the previous speaker's sentence. This label is always combined with another label to indicate whether the same or the other speaker continues speaking after the PCOMP
`hrt`	hearer response token; usually short backchannels, continuers, acknowledgements, etc., that do not contain a (new) proposition of their own and do not take up the turn
`disruption`	current speaker does not reach a PCOMP, but interrupts themselves to rephrase and start a new sentence
`incomplete`	current speaker does not reach a PCOMP before the other speaker takes up the turn
`label_label`	combination of two of the labels above.
`@`	indicates uncertainty about a label; may also co-occur with a combined label to indicate the uncertainty between two specific labels

2.4.2.4 Frequency of PCOMP labels surrounding annotations of breath sounds

Figures 2.10 and 2.11 illustrate how often each PCOMP label gets either preceded or followed by an in- or outbreath respectively. Note that for these two figures, there is no limit for the time interval between the two sequential annotations (i.e., no instances are excluded). This is contrary to what was carried out in the experiments later (see section 3.2). For example, at first it might seem unintuitive that the label `change` gets followed by a breath more often than the label `hold`. But change means that at this point the speaker yielded the turn, and the following inbreath annotation might happen later, when the initial speaker takes the turn again. Another intuitively plausible detail that can be derived from the plot is that two audible outbreaths are rarely, if ever, heard in succession.

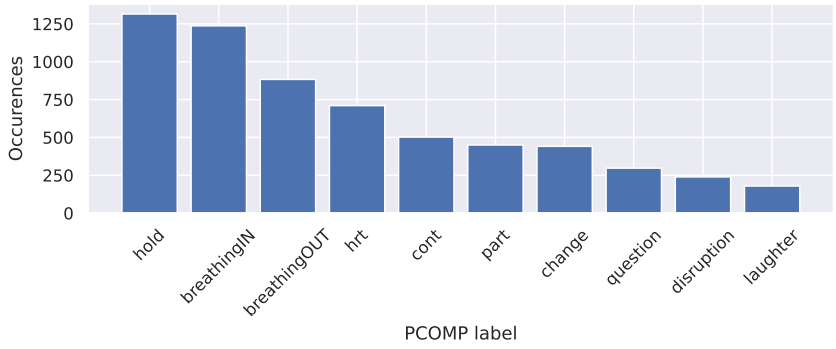


Figure 2.9: Unique annotations in the PCOMP tiers and their occurrences.

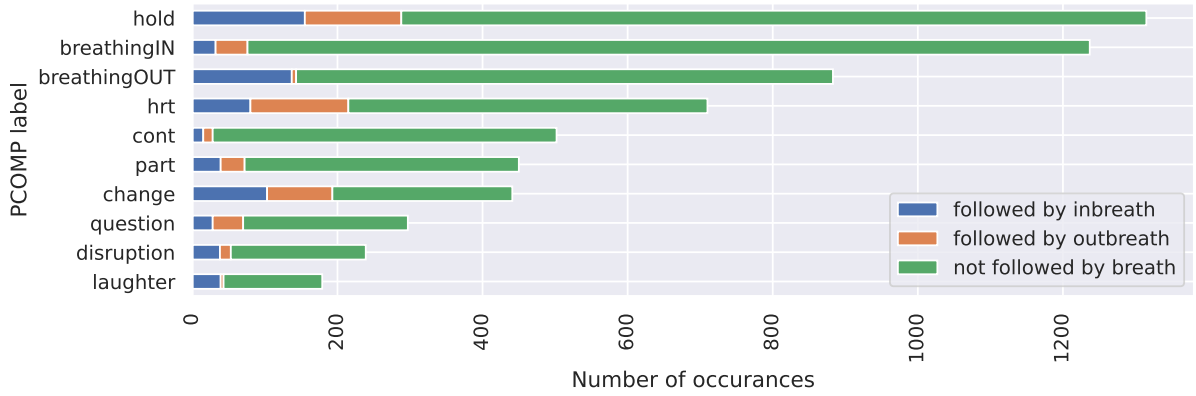


Figure 2.10: Distribution of the frequency with which PCOMP annotations are followed by a breath sound annotation.

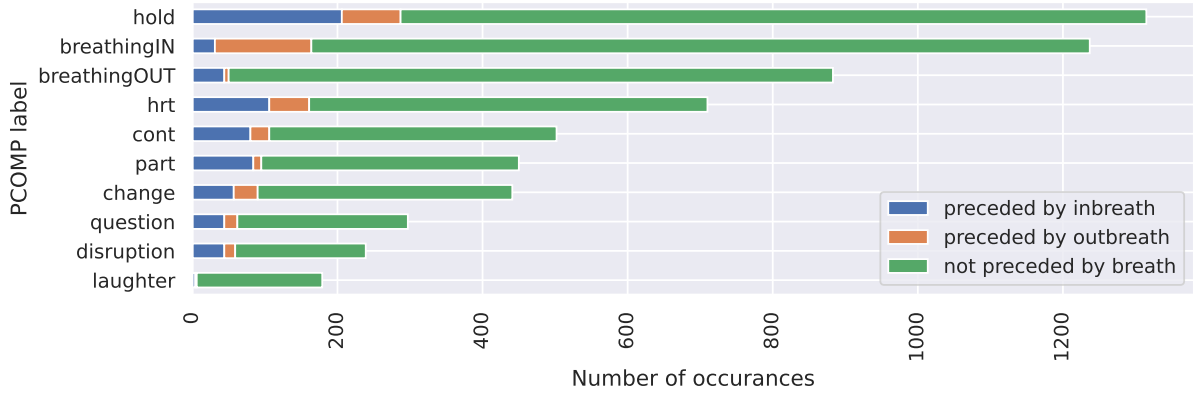


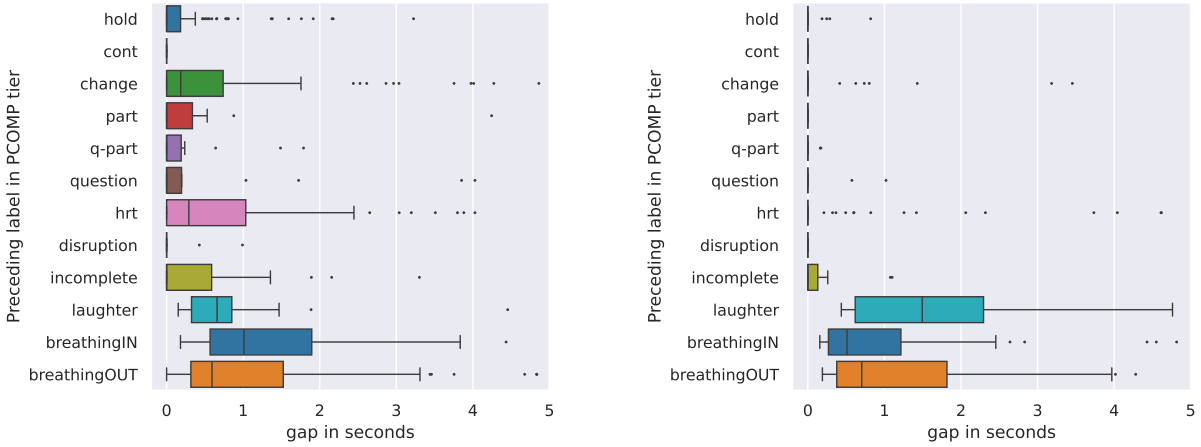
Figure 2.11: Distribution of the frequency with which PCOMP annotations are preceded by a breath sound annotation.

2.4.2.5 Distribution of the temporal gaps between breath sounds and surrounding PCOMPs

Figures 2.12 and 2.13 present boxplots with the distributions of the temporal gaps between the annotations of breath sounds and their surrounding PCOMPs, organized by inbreaths or outbreaths and preceding or subsequent annotations. Note that the PCOMP labels `hes` and `coll` are not included in the figures, as they occur very sparsely.

Before audible inbreath Figure 2.12(a) shows how the time interval gaps to the PCOMP annotations before inbreaths vary by the PCOMP annotations label. The temporal gap between an audible inbreath to a potentially preceding `cont` or `disruption` is typically zero. For the labels `hold`, `part`, `q-part`, and `question` that gap is shorter than $\approx 0,2$ seconds in 75% of instances. For both `change` and `incomplete`, which indicate that the speaker has just yielded the turn, the gap tends to be longer. In 25% of cases, the gap is longer than half a second, and in half of the instances where `change` is the preceding label the gap is at least $\approx 0,2$ seconds long. Note that PCOMPs with the label `question` also lead to a yielded-turn in the majority of cases, as it represents a very active way to hand over the word to the other speaker. Nevertheless, the gap is typically shorter than compared to that of potentially preceding PCOMPs `change` and `incomplete`.

Before audible outbreath As outlined in section 2.2.3, the majority of outbreaths are directly attached to speech. This is evident in Figure 2.12(b), which shows that considerable gaps before audible outbreaths can only be found in the context of laughter and other breath sounds. It is notable that the boxplot indicates that `incomplete` (i.e., the current speaker does not reach a PCOMP before the other speaker takes up the turn) is the only PCOMP label where an outbreath following speech is not that often directly attached. One possible explanation for this phenomenon may be that speakers intentionally plan these brief outbreaths. In situations where the interlocutor takes the turn before a potential point of grammatical completion is reached, this could result in a degree of surprise for the current speaker, which might delay the outbreath.



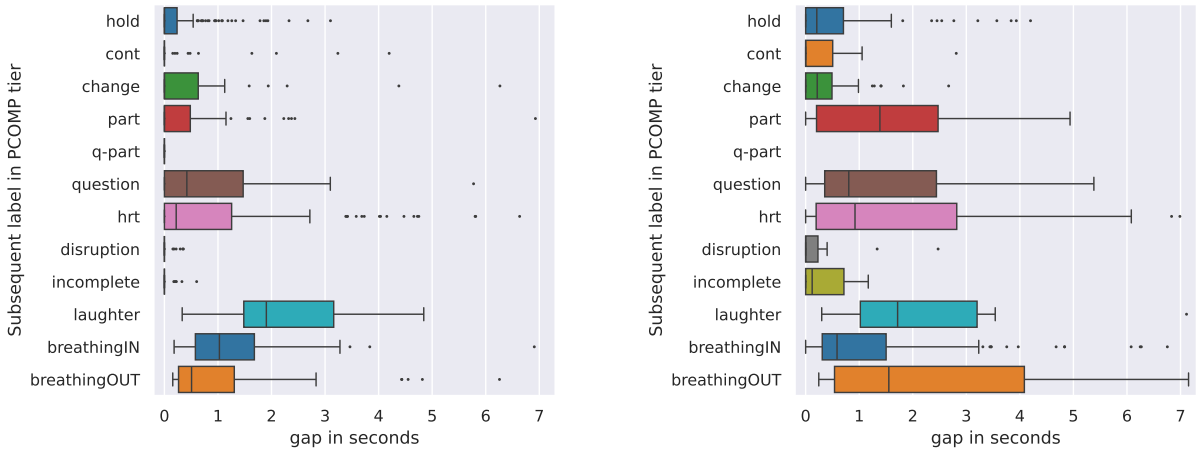
(a) Inbreath

(b) Outbreath

Figure 2.12: Distribution of the gap from the breath sound annotation to the preceding annotation depending on the preceding annotations label. Fliers beyond five seconds are not shown.

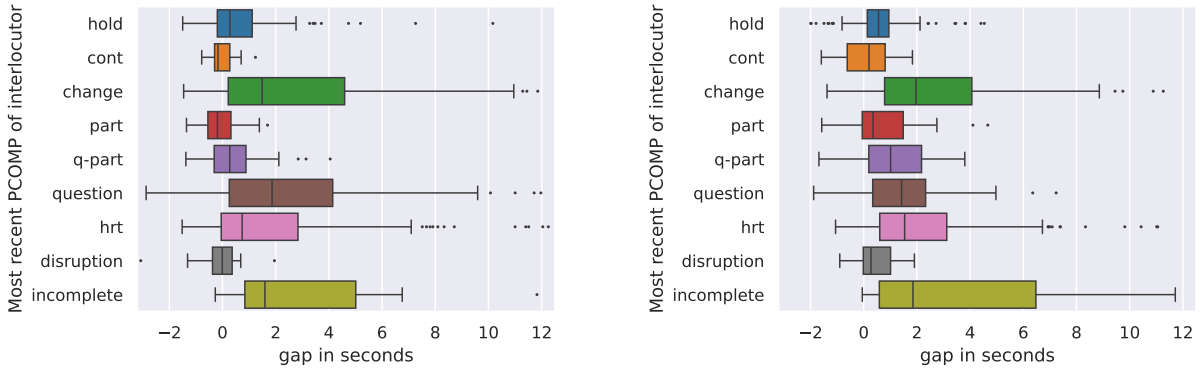
After audible inbreath Figure 2.13(a) shows that if the following PCOMP label after an audible inbreath is either `cont`, `q-part`, disruption or `incomplete`, the time interval between them is zero in more than 75% of the time. For `hold`, `part`, and change, the gap tends to increase. For the label `question`, the gap is typically the longest, with a duration exceeding 0.3 seconds in half of the cases. The gaps to subsequent `hrt` annotations are slightly shorter on average, but there is a considerable number of outliers between three and five seconds.

After audible outbreath Figure 2.13(b) demonstrates clearly that after an audible outbreath it is very unlikely for a `part`, `question`, or `hrt` to follow directly (that means within a gap shorter than 0.15 seconds), whereas the other labels are attached directly afterwards in more than half of the cases within that time. As a `q-part` annotation is almost never present subsequent to an audible outbreath, its statistics are not included in the boxplot. The median of the time interval to following annotations with `question` or `hrt` is ≈ 0.9 seconds, whereas for `part` annotations it is clearly longer: ≈ 1.4 seconds.



(a) Inbreath (b) Outbreath

Figure 2.13: Distribution of the gap from the breath sound annotation to the subsequent annotation depending on the subsequent annotations label. Fliers beyond seven seconds are not shown.



(a) Inbreath (b) Outbreath

Figure 2.14: The distribution of the difference between the start time of the breath sound annotation and the end time of the last ended interlocutor PCOMP annotation, grouped by the latter's PCOMP label. Fliers beyond twelve seconds are not shown.

Most recent PCOMP of the interlocutor before an audible inbreath It is also of interest to investigate the potential direct relationships between audible breathing and the interlocutor. To do this, we analysed the most recent PCOMP of the interlocutor, meaning the last PCOMP within the interlocutors tier that was completed before the current breath sound ended. Negative values are possible for the gap due to the time interval being measured as the

breath sound's start time minus the *interlocutor's PCOMPs end time*. (see also section 3.2). Figure 2.14(a) shows that the labels `change`, `question` and `incomplete`, which all indicate that the turn was yielded by the interlocutor, tend to have a clearly higher temporal gap to the start of the current speakers inbreath than all other labels.

Most recent PCOMP of the interlocutor before an audible outbreath Compared to the temporal gaps between a breath sound and the preceding and subsequent PCOMP of the same speaker, we observe in Figure 2.14(b) that the time gap distributions do not differ in regard to inbreath and outbreath towards the different labels of the last PCOMP annotation of the interlocutor.

3 Method

In order to study the relationship between breath sounds and turn-taking, it is possible to examine the phenomenon from multiple perspectives and at different levels of depth. One approach would be to consider all transitions, which would include all consecutive annotations, either at the IPU or PCOMP level (see section 2.4). This approach would allow for an investigation of the effects of the mere presence of a breath sound in all potential turn-taking and turn-keeping scenarios that can be represented by the present labels, in comparison to silent transitions (i.e., those without a breath sound). However, the annotations utilized would not be adequate for this purpose, as they lack the necessary completeness to perform such an analysis. Furthermore, the main objective of this thesis is to examine the variations of breath sounds and its implications for turn-taking in conversation. Consequently, a more narrow analysis has been conducted, focusing solely on instances where a breath sound is present, or at least where one was annotated in our data.

3.1 Methodological overview

The general approach chosen is to first extract context (mostly durational) and acoustic features of all annotated breath sounds. The dataset is then prepared for classification algorithms, which essentially involves outlier removal, label encoding, data normalization, and feature selection. Secondly, PyCaret (Ali, 2020), a high-level machine learning framework, was used to predict via random forest and gradient boosting classifiers the preceding and subsequent turn-taking labels (using categorically merged PCOMP labels) of the same speaker, as well as the most recent PCOMP annotation of the other speaker. For each prediction target, three differential feature sets were employed. The first feature set utilizes solely binary contextual and durational features. The second adds a relative intensity as well as an estimate for the breathing volume. The third adds selected other acoustic functional features. We then used SHapley Additive exPlanations (SHAP) (Lundberg & Lee, 2017) to analyse the importance of features and how they contribute to the results of the models. The flowchart in Figure 3.1 provides an overview of the procedure, from the initial selection of the dataset, to the extraction of features and targets from it, and the preparation of the dataset for classification experiments.

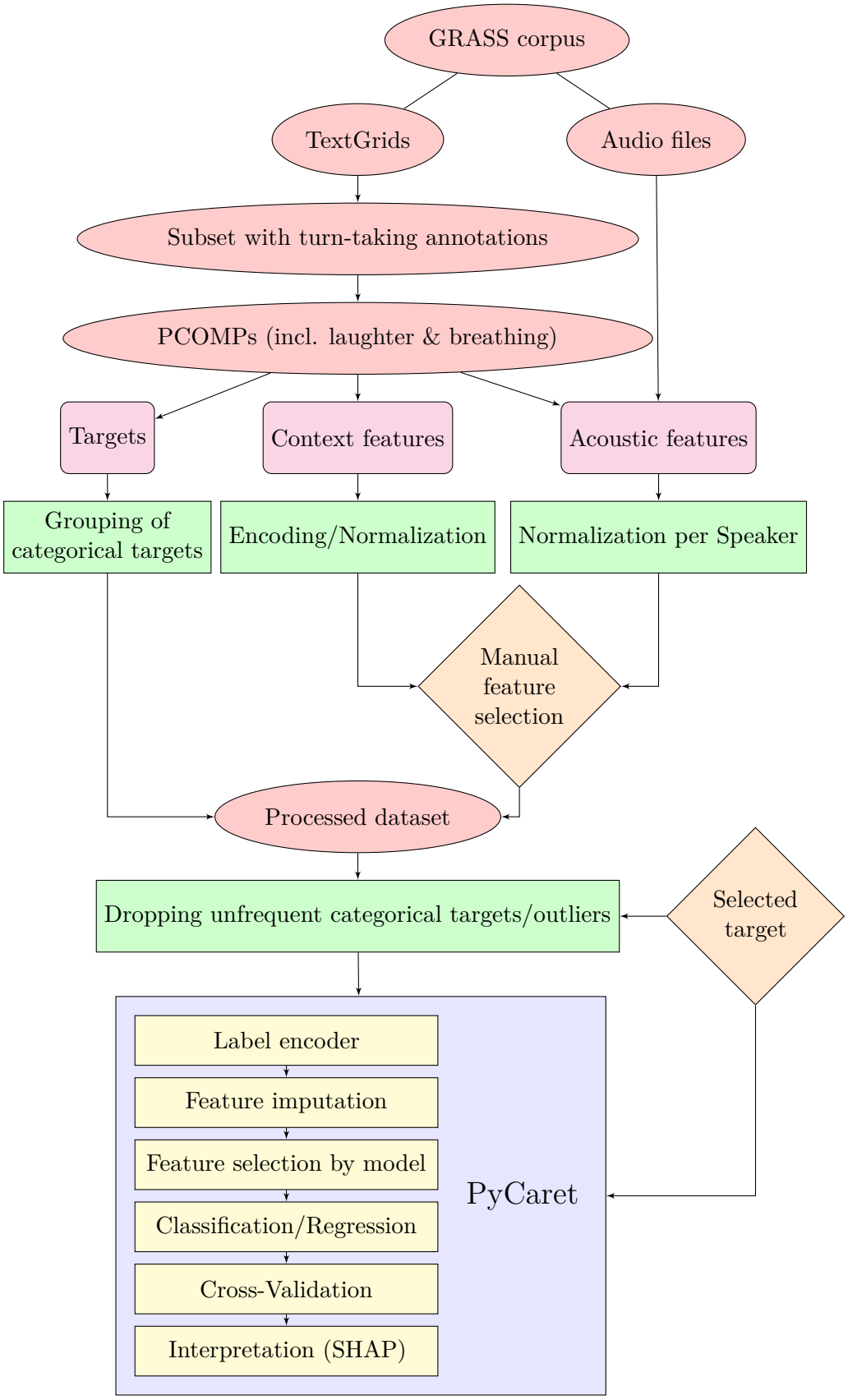


Figure 3.1: This overview presents the methodology employed in this thesis. The process begins with the extraction of information from the GRASS corpus, continues with data pre-processing, and concludes with the preparation for machine learning models and their application.

3.2 Targets

Grouping At the time of writing this thesis, there were few studies, particularly no comparable quantitative ones, that indicated the relationship between breath sounds and their surroundings. In our case, the PCOMP and the IPU annotations available in GRASS are used as a reference point. Therefore, the selection of our targets is exploratory in nature: we examine several potentially relevant and available interrelations. This includes preceding and subsequent PCOMP annotations of the same speaker. Furthermore, the preceding PCOMP annotation of the interlocutor is likely connected to the speaker's current breath sound. We defined the preceding PCOMP annotation of the interlocutor as the last one that ended before the current breathing annotation ended. However, we exclude the subsequent PCOMP annotation of the interlocutor from our experiments as it is highly unlikely that a connection can be found, due to the forward-looking nature of the PCOMP labels. Another target of interest is whether there is speech following a breath sound is part of a turn or whether it constitutes a hearer response token (backchannel). Moreover, we examine the temporal gap between the breathing and potential subsequent speech. An overview is given in Table 3.1.

Table 3.1: Overview of the different sources for the targets.

Target source	Type	Description
`pre_pcomp`	categorical	preceding annotation on the same PCOMP tier
`post_pcomp`	categorical	subsequent annotation on the same PCOMP tier
`interlocutor_last_pcomp`	categorical	most recent PCOMP annotation of the interlocutor
`post_gap`	float	durational gap till subsequent speech

3.2.1 Categorical targets

We aim to investigate the communicative functions of breath sounds by predicting one of the surrounding PCOMP annotations of a single breath sound annotation, both with and without considering acoustic features of the breathing. The size of the dataset is relatively small, comprising 1422 entries against eleven PCOMP labels, some of which have very low numbers of occurrences, as outlined in section 2.4.2.3. The utilization of all labels as a target, given the present class imbalance, would result in data sparsity, which would hinder the ability of machine learning models to effectively learn patterns and make accurate predictions. Therefore, to facilitate analysis, we clustered the target PCOMP labels into turn-taking categories, as described in section 2.4.1. Furthermore, two additional groups have been created for surrounding annotations of the current speakers PCOMP tier, containing the annotations for laughter and breathing. Table 3.2 provides a detailed depiction of the implementation of this mapping.

Simply using a single subsequent PCOMP label from either the speaker or interlocutor tier as the target to be predicted does not fully reflect the turn-taking process and all its variations from a breath sound's point of view. In particular, timing and overlapping speech make the situation complex and difficult to decompose or classify. A more comprehensive set of target classes using the combinations of PCOMP labels from a breath sound's point of view to cover the complex interaction of these labels is therefore difficult to find, and it is by no means certain that it would add any benefit to our method.

Table 3.2: Grouping of the annotations from the PCOMP tier to target classes. An asterisk indicates that those target groups were excluded from the dataset due to their infrequent occurrence. A dash indicates that these annotations were skipped and the successive annotation was used. The PCOMP labels are described in Table 2.3.

PCOMP annotation \ Target	pre_pcomp	post_pcomp	interlocutor_last_pcomp
hold	hold	speech	hold
cont			
part			
q-part			
hes			
coll			
disruption			
change	change		change
question			
incomplete			
hrt	hrt	hrt	hrt
laughter	laughter*	laughter*	—
breathing	breathing*	breathing*	—

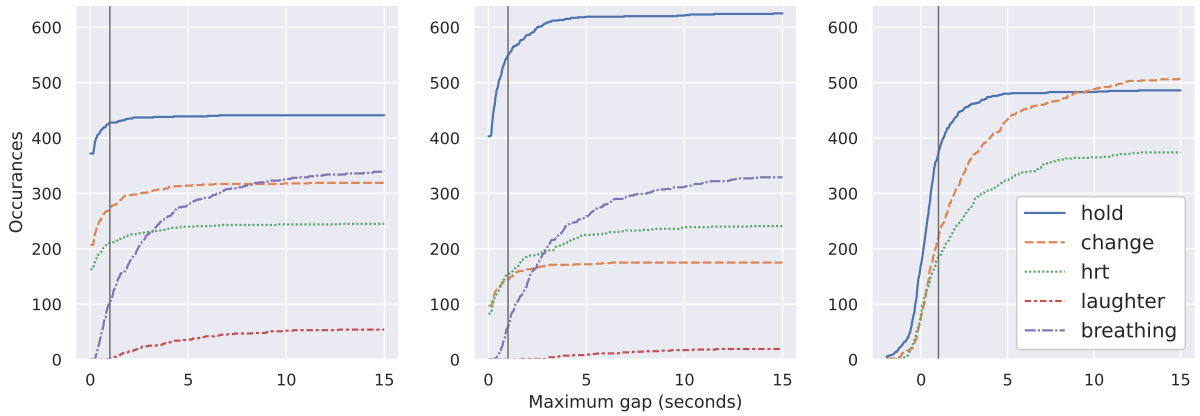
3.2.1.1 Dropping instances based on the target

Firstly, we dropped instances in our analysis where the target PCOMP annotations have an arbitrary gap of more than one second from the observed breath sound annotation. The reason for this is that surrounding PCOMP annotations with a larger time gap are more likely to have no causal relationship with the breath sound, and there is an increased chance that there may have been more actions influencing turn-taking that we have not accounted for. Secondly, we remove 125 data points that either don't have a PCOMP annotation before or after them, or are marked with uncertainty. The former are most likely to be breath sound annotations right at the end or at the beginning of the five-minute time interval for each conversation where the PCOMP annotations were made. The removal of the latter is to prevent complex and rare cases from skewing our results.

3.2.1.2 Target distribution

The removal of targets with a higher durational gap than 1s from the dataset results in a change to the distribution of the number of occurrences of the target values, as illustrated in Figure 3.2. The maximum gap is shown from zero to 15 seconds. While there are instances of targets with a greater distance, they are relatively uncommon and the distribution is almost unchanged. The occurrences of `breathing` either preceding or following the current instance exhibit a notable increase from zero, which is consistent with the expectation that subsequent breath sound is to be annotated within a single time interval. A similar pattern can be observed for `laughter`. This

can be attributed to the fact that laughter and breath sound annotations are often annotated within the same interval. Given the infrequency of occurrences of targets with `breathing` and `laughter` within the chosen conditions, these are excluded from the subsequent experiments as indicated in Table 3.2. For the *Preceding PCOMP* and the *Subsequent PCOMP* targets the group `hold` occurs roughly two times as often as the `change` and `hrt` ones when only including those with a durational gap of one second to the current breath sound, while for the *Last PCOMP of the Interlocutor* target the ratio is ≈ 3.4. Two other anticipated characteristics are also observable: (1) If the last breathing occurred more than two seconds ago, the probability of a hearer-response token occurring in the subsequent turn is slightly higher compared to other speech. (2) The last PCOMP annotation of the interlocutor being a turn-change becomes more likely the longer in the past it happened, which is a self-evident conclusion, given that the probability of the current speaker having already spoken before the current breath sound annotation is increased.



(a) Preceding PCOMP group (b) Subsequent PCOMP group (c) Interlocutors last PCOMP group

Figure 3.2: Number of occurrences of the grouped annotations of the PCOMP tiers as a function of the maximum temporal distance to the breath sound annotation. The vertical grey line at one second indicates our boundary, which we chose to still include them in our classification experiments.

3.2.2 Numerical targets

In addition to predicting the PCOMP group following an audible breath sound, it is of interest to ascertain whether it is possible to predict the time gap between an audible breathing and subsequent speech. In order to maintain consistency with the experiments where we are targeting the classes of subsequent speech, instances from our dataset were also excluded for this case where the time interval to the following speech was greater than one second. The distribution of this time gap is illustrated in Figure 3.3. The vast majority of the times the gap is shorter than 0.1 seconds. Gaps from 0.1 already are about ten times less frequent and their frequency steadily continues to decrease after that. It can be observed that in approximately 41% of cases, the gap is shorter than 0.1 seconds. Note that the probability of the temporal gap falling into the next 0.1-second bins is already 10 times less. The frequency of occurrences continues to decrease after that. Given the challenges inherent in applying regression machine learning methods to data that is distributed in this manner, we have chosen to refrain from pursuing this objective further within the context of this thesis. Nevertheless, the subsequent temporal gaps are differentiated according to the following PCOMP label in Figure 2.13.

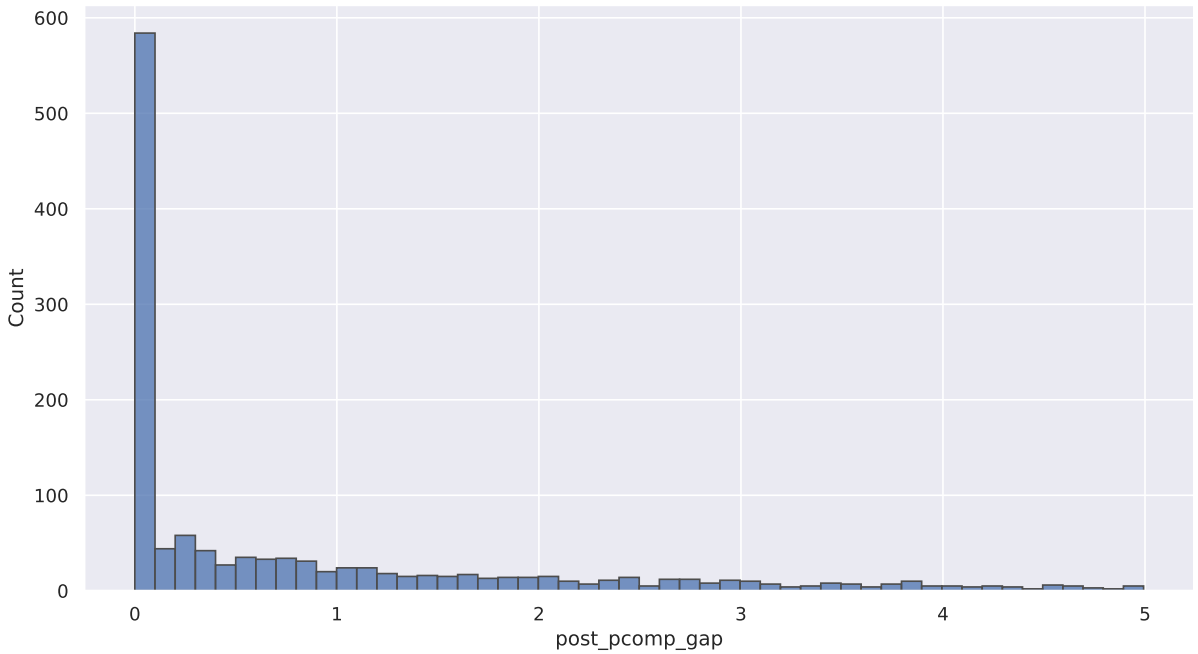


Figure 3.3: Distribution of the temporal gap from a breath sound annotation to subsequent speech. Fliers above five seconds are not shown.

3.3 Features

A central aim of this thesis is to identify features of audible breathing that may be predictive of turn-taking in a conversation. Given the limited existing discussions on this topic in the literature, as outlined in section 1.4, there were few indicators for which features to use. As a foundation, we extracted basic contextual features of the breath sounds' surroundings, as long as they were still related to the breathing itself. However, we only used those later that contained information which could be deduced before the end of the breathing annotation. Furthermore, we included those acoustic features that had already been discussed in the literature. In addition, a multitude of other acoustic features were generated with the intention of employing them in an exploratory manner. An overview of all the features ultimately utilized can be found in Table 3.3. It is of high importance to acknowledge that we refrain from including any PCOMP labels as input features in our machine learning models. This decision is determined by the strong relationship that exists between adjoining PCOMP labels, especially those taken from the respective other speaker.

Table 3.3: Overview of all used features or feature categories. An asterisk indicates that this is actually a feature category, or in other words a Low-Level Descriptor (LLD), and in this case the actual features are various functionals and delta's calculated from windowed frames of the PCM audio signal (see section 3.4.5).

Feature	**Type**	**Description**
duration	float	The duration of the breathing (annotation) in seconds.
ends_with_in	boolean	Whether the breathing annotation starts with an inbreath.
starts_with_in	boolean	Whether the breathing annotation ends with an inbreath.
pre_pcomp_gap	float	The start time of the breathing minus the end time of preceding PCOMP annotation in seconds.
pre_pcomp_duration	float	The duration of the preceding PCOMP annotation in seconds.
interlocutor_gap	float	The start time of the breathing minus the end time of most recent PCOMP annotation of the interlocutor in seconds.
interlocutor_last_pcomp_duration	float	The duration of the most recent PCOMP annotation of the interlocutor in seconds.
overlap	boolean	Whether the breathing annotation is in overlap with the interlocutor
rel_intensity	float	The relative intensity as described in section 3.4.5
volume	float	simply approximation of the breathing volume by multiplying the mean root-mean-square energy value by the duration
pcm_RMSenergy*	float	non-relative root-mean-square energy
pcm_psySharpness*	float	psychoacoustic Sharpness
pcm_spectralEntropy*	float	spectral entropy
pcm_spectralRollOff90*	float	90% percent spectral roll-off point
pcm_spectralCentroid*	float	spectral centroid
pcm_spectralFlux*	float	spectral flux
pcm_spectralHarmonicity*	float	harmonicity (mean of consecutive local min-max differences)
pcm_spectralSlope*	float	spectral slope over maximal frequency range
pcm_spectralKurtosis*	float	spectral kurtosis
pcm_spectralSkewness*	float	spectral skewness
pcm_spectralVariance*	float	spectral variance

3.4 Feature Extraction

3.4.1 Pre-preocessing of GRASS annotations

TextGridTools, a *TextGrid* processing and analysis toolkit for Python (Buschmeier & Włodarczak, 2013) was used to extract information from the raw annotations available in the GRASS corpus. At the time of working on this thesis, the breathing annotations in the PCOMP tier were the most complete and precise, in comparison to other annotation tier. Therefore, they were used as a basis to create a dataset in the form of a Pandas DataFrame (McKinney, 2010), where each breathing annotation was assigned a single entry (row). Note that the lack of knowledge regarding the precise boundaries of consecutive breath sounds (i.e., annotation as a single interval) does not necessarily represent a disadvantage. At present, there is no evidence to suggest that separate analysis would enhance the predictive power of these for the purposes of their meaning in conversational speech.

3.4.2 Extracting contextual and durational features

Contextual or so-called structural features were extracted from the immediate surround annotations. We were interested in understanding the communicative functions of breath sounds at the point in time at which they occur. Therefore, we did not include features that contain information that is only known after the breath. However, they have been used as targets in some experiments. The timing-related features therefore include the duration of the preceding PCOMP annotation, as well as the time interval gap from the end of that annotation to the start time of the current breathing annotation. Furthermore, the same information is extracted for the most recent PCOMP annotation of the interlocutor. The term *most recent* indicates that the PCOMP annotation of the interlocutor ends before the current breathing annotation ends.

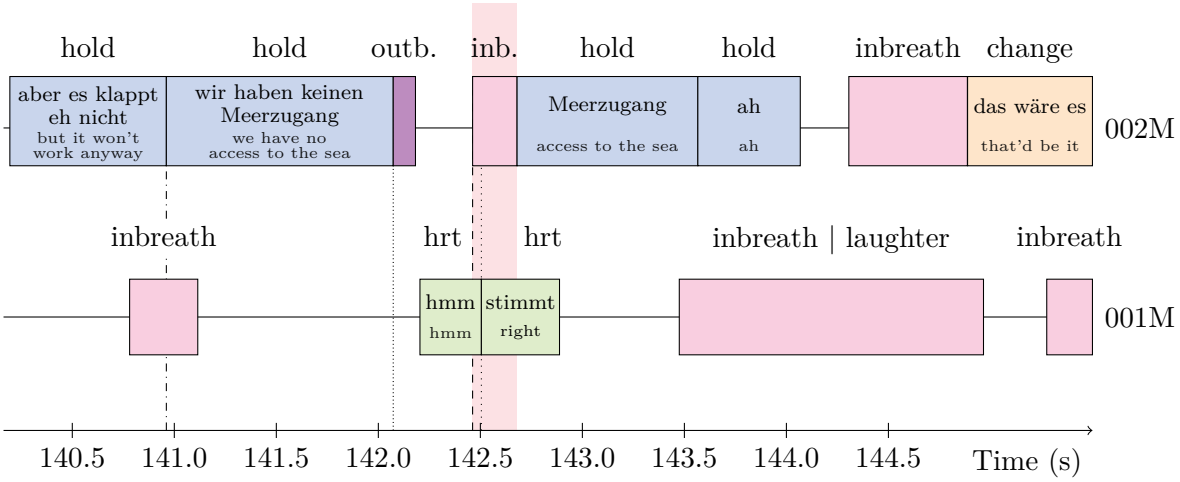


Figure 3.4: Temporal-contextual features illustrated by an example inbreath. The dashed vertical line represents the beginning of the inbreath, which is the focus of the analysis. The vertical loosely dotted line marks the end of the most recent PCOMP of the interlocutor. The vertical line with densely dotted dashes marks the end of the preceding PCOMP. The vertical dash-dotted line marks the beginning of the preceding PCOMP annotation. Values in seconds: `duration` $= 0.218$, `pre_pcomp_gap` $= 0.389$, `interlocutor_gap` $= -0.043$, `interlocutor_last_pcomp_duration` $= 0.301$.

Figure 3.4 illustrates an example of a breath sound annotation and its surrounding context, including the PCOMP and the word transcriptions. It demonstrates the extraction of temporal features, particularly highlighting the challenge of defining a concise yet comprehensive set of temporal features, especially those related to temporal intervals in relation to the interlocutor. For instance, the final inbreath annotation of speaker 001M, depending on whether it ends slightly before or after the interlocutor's utterance "das wäre es" ("that'd be it"), would have a significant impact on whether we would refer to that `change` or the preceding `hold` "ah" as the interlocutor's most recent PCOMP by the aforementioned definition.

In addition, we retrieved other information that is not utilized as features. This information was either beneficial for a more detailed manual examination of a specific point of interest or for analysis purposes. For example, we retrieved the IPU labels, the orthographic transcriptions of the surroundings, and the duration of the preceding turns.

3.4.3 Additional features

Overlap In conversational speech, it is common for utterances from different speakers to overlap in time. Furthermore, breathing also occurs frequently in overlap with the speech of the interlocutor. The presence of an overlap for an annotated breath sound is extractable from our raw data in two ways. Firstly, in the orthographic transcription tier, overlapping sections can be identified by double backslashes surrounding them. A total of 789 instances of overlap were identified using this method, representing $\approx 55\%$ of all cases. Secondly, an alternative approach involves examining the presence of speech in the interlocutor's forced alignment tier, which is also available in the TextGrid files of GRASS. This yielded 665 cases with overlap, representing $\approx 47\%$ of all cases.

This substantial difference in the number of overlaps between these two methods of detection is partly due to the discrepancies between the different annotation tiers themselves (see section section 2.2.2.1), and partly due to the technical constraints involved in mapping an interval in one tier to a point in time in another annotation layer. Initially, we decided not to interpret breath sounds in overlapping speech in general, but given the high number of cases and that this would undoubtedly result in a highly biased or selected sample of cases, we decided not to exclude them. However, it cannot be precluded that these interferences of the interlocutor on the acoustic features might actually lead to an improvement of the predictive power in the subsequent experiments. As discussed in section 2.1.1, this fact has to be taken into account when interpreting the results. We include a boolean factor `overlap` in our feature set to capture this information in the analysis.

Breathing type As previously stated, one breathing annotation interval may encompass multiple in- and outbreaths. This is the case with 129 datapoints, representing approximately $\approx 0.09\%$ of the entire dataset. In order to circumvent the potential issues associated with high-dimensional features, which may result in sparsity and the risk of overfitting, we decided to break down the available information into two binary features. The first binary feature determines whether the annotation ends with an inbreath, whereas the second determines whether the annotation begins with an inbreath.

3.4.4 Dropping instances with very short breathing duration

We decided to remove 39 data points where the duration of the annotated breath is less than 0.06 seconds. These are likely to be either breath sounds of very low intensity and therefore barely perceptible, or they are directly following speech, in which case they are often indistinguishable from aspiration sounds of plosives (see section 2.2.3). Furthermore, their short duration makes it impossible to calculate functionals of acoustic features.

3.4.5 Extraction of acoustic features

As part of the research for this thesis, it became increasingly evident that there has been hardly any explicit definition and analysis of acoustic features associated with breath sounds in the existing literature. To the best of our knowledge, only two acoustic features relating to this aspect have been explicitly discussed, *relative intensity* and *spectral centroid* (Trouvain et al., 2020), as outlined in section 1.4. We implemented the algorithm for computing the *relative intensity* based on the difference between the mean intensity two seconds prior and the breath sound itself in decibels, as described in (Trouvain et al., 2020), by utilizing the core audio signal processing tools from the python library *librosa* (McFee et al., 2023).

To calculate the spectral centroid and all other acoustic features, the software openSMILE (Eyben et al., 2010) was employed. openSMILE was configured to generate a feature set that is prevalent in the field, the Compare2016 feature set (Schuller et al., 2016). This resulted in the generation of over 6,000 features, which are based on specific Low-Level Descriptors (LLDs) and are derived from LLD functionals and LLD deltas. For example, a single feature is `pcm_fftMag_spectralCentroid_sma_de_minPos`. The underscores indicate the transitions between the sequence of calculation steps, starting from the first to the last. The initial `pcm` stands for the pulse-code modulated audio signal. Next, `fftMag` refers the magnitude of the Fast Fourier Transform and `spectralCentroid` is the actual low-level descriptor, or feature category: the spectral centroid. Subsequently, `sma` is a keyword for the appliance of a contour smoother (moving average filter) and `de` is another keyword for the delta between subsequent frames. Finally, `maxPos` denotes the applied functional: the position of the maximum value.

It should be noted that using this feature set has certain disadvantages. Primarily, the size of this feature set is considerably larger than that of our dataset. Secondly, it is probable that a significant number of those features are either unsuitable for describing breath sounds or that they are likely to be highly correlated with one another. Consequently, we had to invest substantial quantitative and qualitative analyses to identify the most relevant features, namely those that have the greatest predictive power on turn-taking, from this feature set, see section 3.6.

3.5 Feature normalization

Feature normalization is a critical stage in the machine learning pipeline. While tree-based and gradient boosting models are more resilient to non-normalised data inputs, we nevertheless evaluated different normalization methods. For contextual duration features, we arbitrarily chose to apply a standard non-centred scaler $z = x/s$, where x is the value of a sample, s is the standard

deviation of the training samples, and z is the resulting value. Given the strong variability observed in the distributions of a single acoustic feature across speakers, a standard scaler with centering was applied separately for each speaker to this feature subset: $z = (x - u_{speaker})/s_{speaker}$, where $u_{speaker}$ is the mean of the training samples from the current speaker and $s_{speaker}$ is the standard deviation of the training samples from the current speaker. It should be noted that some durational features exhibit considerable variation between speakers. However, in these cases, we chose not to remove the speaker bias, as it did not result in an increased performance of the models.

3.6 Feature selection

Why feature selection is important Feature selection is a pivotal task within the realm of machine learning, particularly when confronted with a feature set that surpasses the number of available data samples. As our available feature set is five times bigger than the number of data samples in our pre-processed dataset, it is crucial to pre-select features in order to avoid overfitting of the models. The challenge of selecting the best features has been the subject of numerous studies, including those conducted in the field of paralinguistic analysis. Pohjalainen et al. (2015), for example, compared different unsupervised feature selection methods using an almost identical feature set to ours, with the goal of achieving the best scoring binary classification of several different speaker traits (e.g., likeability, agreeableness, conscientiousness) using a standard k-nearest-neighbors classifier. However, our primary objective is not to achieve the highest possible prediction scores. Rather, our focus is on identifying and interpreting the most relevant features. Nevertheless, we are interested in ensuring that our classifiers perform at an acceptable level, as this will ensure that the interpretation of the features and their relative importance is not per se skewed.

Defining primary features Due to limitations in computational power and time, unsupervised approaches such as Sequential Forward Selection or even less resource-intensive methods like Heuristic Genetic Feature Selection could not be carried out. Instead, we commenced by determining and utilizing a fixed feature subset, which consisted of all contextual features and was expanded with the *relative intensity* and an estimator of the breathing volume as additional acoustic features (i.e., all features until `volume` in Table 3.3) depending on the experiment. This exclusive use of the aforementioned feature subset was employed as a primary benchmark. Furthermore, this approach already provides insights into the information conveyed solely by the presence, timing, and likeability of the interlocutor perceiving the breathing Note that the decision to pursue this specific extension including the *relative intensity* was driven by the fact that the (relative) intensity has been the most discussed feature in the literature and we aim at quantifying its effect on the communicative function of a breathing. In this stage, features exhibiting minimal impact or even negative feature permutation importance (see section 3.7.5) across all targets were removed, namely `starts_with_in` (whether the breathing annotation starts with an inbreath). Although the duration of the breathing annotations did not appear to have an impact on the a priori classification experiments, we retained it at this stage as it was of interest to discuss later on.

Narrowing down potentially important acoustic features using domain knowledge A further objective was to identify acoustic features that capture variations in breath sound that can be shaped by the speaker and contain communicative information. To achieve this, we selected several low-level descriptors (LLDs) from the openSMILE feature set, which we believed were suited best to describe the variations of these potential shapes of breath (see the features marked with an asterisk in Table 3.3). For instance, Mel Frequency Cepstral Coefficients (MFCCs) were not included in the analysis because they are inherently linked to the formants, and Trouvain et al. (2020) have demonstrated that the formants in inhalation noises exhibit minimal variability (see section 1.4). Subsequently, the most effective features within each LLD category were identified separately in the context of each target, using both *SHapley Additive exPlanations* (SHAP) analysis (Lundberg et al., 2020) and feature permutation importance (see section 3.7.5). The features with the highest mean absolute SHAP values and highest feature permutation importances were retained, unless their relative importance compared to the top-performing features from the fixed feature set was marginal. This process ensured that only potentially informative features were kept in the final feature set. However, this method does not consider the possibility that, in certain circumstances, the combination of specific features within the acoustic feature set may be the dominant influence. In other words, the drawback associated with this approach is particularly the risk of eliminating interdependent features. When combining the selected features from all the LLDs used, in cases where there were strong correlations between these top features (correlation coefficient greater than 0.6), only the best one was retained.

Final feature selection As the resulting feature set was still rather extensive (≈ 30 features), we employed an automated feature selection process, utilizing the feature importances of a Light Gradient Boosting Machine (five times cross-validated), to reduce the feature set to a maximum of twelve features. Note that the context features, the relative intensity, and the breathing volume estimate were forcefully kept in all cases. Firstly, this number of features yielded the optimal overall prediction scores. Secondly, increasing this the number of features further led to an highly increased chance of the models overfitting, as indicated by close to zero or even negative permutation feature importances.

3.7 Classification

3.7.1 Machine learning models

The decision was taken to utilize and compare multiple ensemble learning models, not because the highest prediction scores were the objective, but because the better the model performs, the more meaningful and robust are the interpretations that can be derived of it. As we are working in a novel field and therefore lacked knowledge regarding the anticipated order of prediction scores, a comparison of several methods also provides certainty. Note that the PyCaret library (Ali, 2020) greatly facilitates this process. We focused on three models for our analysis: (1) **XGBboost**: Extreme Gradient Boosting (Chen & Guestrin, 2016), an effective and widely used tree boosting algorithm that is sparsity-aware and employs a weighted quantile sketch of

the features for approximate tree learning; (2) **lightGBM** (Ke et al., 2017): Light Gradient Boosting Machine, a highly effective Gradient Boosting Decision Tree (GBDT) based machine learning algorithm that has been specifically designed for addressing high feature dimensions and large data sizes; and (3) **Random Forest** (Breiman, 2001), one of the most commonly used algorithms, particularly due to its robustness, flexibility and simplicity. All of the models in question share multiple characteristics. The same tools regarding interpretability can be used, mainly SHAP. All are well integrated in PyCaret, and can be used in conjunction with the *explainerdashboard* (Dijk et al., 2023). Furthermore, all of them are robust with respect to highly correlated features, which is beneficial in that it allows for the identification of meaningful features, even if they are similar, which is a main concern for us. Lastly, all of them support both classification and regression tasks.

3.7.2 Configuration

Parameters We have refrained from further tuning our models by explicitly setting parameters, as PyCaret and the respective Python libraries already set many reasonable defaults. The most important parameter values of these defaults were as follows:

- **XGBoost**
  - **booster:** `'gbtree'` – Gradient Boosting Tree.
  - **eta:** 0.3 – Step size shrinkage used in updates to prevent overfitting.
  - **gamma:** 0 – No minimum loss reduction required for further partitioning.
  - **max_depth:** 6 – Maximum depth of a tree.
  - **min_child_weight:** 1 – Minimum sum of instance weight needed in a child.
- **LightGBM**
  - **booster:** `'gbtree'` – Traditional Gradient Boosting Decision Tree.
  - **num_leaves:** 31 – Maximum tree leaves for base learners.
  - **max_depth:** -1 – No limit for maximum tree depth for base learners.
  - **learning_rate:** 0.1 – Boosting learning rate.
  - **n_estimators:** 100 – Number of boosted trees to fit.
  - **subsample_for_bin:** 200000 – Number of samples for constructing bins.
- **Random Forest**
  - **n_estimators:** 100 – The number of trees in the forest.
  - **criterion:** `'gini'` – The Gini importance function measures the quality of a split.
  - **max_depth:** 0 – No limit on the maximum depth of the tree..
  - **min_samples_split:** 2 – A minimum of two samples required to split an internal node.
  - **min_samples_leaf:** 1 – A minimum of one sample required to be at a leaf node
  - **min_weight_fraction_leaf:** 0.0 – No minimum weight fraction for leaf nodes specified.

Target imbalance Given the highly imbalanced distribution of the targets in our machine learning experiments, even after grouping them, we attempted to overcome this imbalance using the Synthetic Minority Oversampling Technique (SMOTE) (Lemaître et al., 2017). However, this did not result in any substantive differences in the overall scores, and thus it was not employed in the experiments.

Feature transformations A common challenge faced by many classification algorithms is the inability to cope with non-Gaussian features. A typical initial approach to address this issue is to apply a more robust scaling technique, such as removing the median and scaling the data according to the quantile range. Another more radical approach is to apply transformations to the features such that the transformed data can be represented by an approximate normal distribution. We experimented with the application of a Yeo-Johnson transformation (Yeo & Johnson, 2000), but as this did not affect the predictions, it was deemed not to be useful in the final analysis. This is plausible, as our chosen machine learning models should be able to cope well with non-gaussian feature distributions by design.

Dimensionality reduction of the feature set Note that in order to analyse the impact of certain features on the predictions, we refrained from employing techniques such as Principal Component Analysis for reducing the dimensionality of the feature set or automatically dropping features based on a variance threshold.

Automatic outlier removal We considered the possibility of automatically removing outliers, a feature offered by PyCaret through PCA linear dimensionality reduction using the Singular Value Decomposition technique in conjunction with an Isolation Forest Algorithm. This includes a configurable threshold for determining the percentage of outliers to be removed. However, we observed that this technique had no impact, so we ultimately decided not to use it.

3.7.3 Validation

In order to provide reassurance regarding the reliability of the results, we employed a ten times stratified K-Fold cross validation with a 70 to 30 ratio split between the training and the test set for each fold. Additionally, a K-fold iterator variant with non-overlapping groups was trialed, with the objective of ensuring that each conversation appeared exactly once in the test set across all folds. Nevertheless, this approach yielded no discernible differences in the metrics. Consequently, the stratified K-fold was retained, also because of its focus on considering class information in order to prevent the creation of folds with further imbalanced class distributions was of importance (see Figure 2.3).

3.7.4 Metrics

In our classification experiments, the distribution of targets is occasionally highly imbalanced, as can be seen in Figure 3.2. Furthermore, we consider all targets to be of equal importance. Therefore, we have chosen to focus on three metrics which take these factors into account:

F1 Score The F1 score is a metric that combines precision and recall, calculating the harmonic mean of both. Its value range is from 0 (worst) to 1 (best). For the binary classification case, it can be written as

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{3.1}$$

where precision and recall can be calculated via the *True Positives* (TP), *True Negatives* (TN), *False Positives* (FP), and *False Negatives* (FN) as

$$\text{Precision} = \frac{TP}{TP + FP} \quad \textit{and} \tag{3.2a}$$

$$\text{Recall} = \frac{TP}{TP + FN}. \tag{3.2b}$$

Balanced Accuracy The Balanced Accuracy is the arithmetic mean the of recall obtained on each class. Its main idea is to reduce the influence of imbalanced targets to the prediction score. Its value range is from 0 (worst) to 1 (best). For the binary classification it is defined as

$$\text{Balanced Accuracy} = \frac{1}{2}\left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\right). \tag{3.3}$$

Matthews Correlation Coefficient (MCC) The Pearson-Matthews Correlation Coefficient (MCC) is a balanced measure that also considers classes with a high imbalance in occurrence. Its range is within $[-1, 1]$, where $+1$ represents a perfect prediction, 0 a random prediction, and -1 an inverse prediction. For the binary classification it is defined as

$$\text{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}. \tag{3.4}$$

Since the MCC most fairly accounts for the prediction of all classes, while still being robust against an imbalanced dataset (Chicco & Jurman, 2023), it is the most appropriate metric among those mentioned above for determining the best model in our experiments. Consequently, we have utilized the model identified this way for our analyses and interpretation.

3.7.5 Feature importance

Various methods exist to assess the importance of a feature on the output of a tree-based machine learning model. Each method provides an insight from a different perspective. The following three methods were utilized to a greater or lesser extent in this thesis.

Gini importance The Gini importance, also known as mean decrease in impurity, measures the total decrease in node impurity that a feature provides across all trees in the model (Breiman, 2001). It can be calculated exclusively utilizing the trained model, without having to further utilize (testing) data.

Permutation feature importance This feature importance evaluates the impact of shuffling a feature's values and investigating that effect on the model performance, measured by an arbitrary metric (see section 3.7.4), thus capturing the feature's influence on the model's predictive power.

Mean absolute SHAP values The SHAP values derive from cooperative game theory and quantify the average absolute contribution of each feature to the prediction (usually utilizing the test-set). Their mean absolute values offer a consistent and theoretically grounded interpretation of feature influence (Lundberg & Lee, 2017).

Permutation feature importance is particularly useful for detecting overfitting, as it assesses the model's reliance on specific features by measuring performance changes when feature values are randomly shuffled. In other words, it shows whether the model has captured genuine patterns or merely memorized the training data. Accordingly, we employed this methodology to validate the selection of our features. For interpreting our results, we primarily focused on SHAP values because they not only quantify feature importance but also indicate the direction and magnitude of a feature's impact on the model's output (using the test data).

4 Results

The results are presented separately for each of the three classification targets: *subsequent PCOMP group*, *preceding PCOMP group*, and *most recent interlocutors PCOMP group* of an audible breathing.

4.1 Classification of the subsequent PCOMP group

This experiment was a binary classification task. The two target groups were `speech` (which included all PCOMP labels except `hrt`) and `hrt` (which directly represented the PCOMP label for hearer-response-tokens, or in other words backchannels). The rationale behind this grouping is the forward-looking nature of the PCOMP labels. We did not anticipate that the breathing would have an impact on what happens after the subsequent PCOMP. The dataset was cleaned by removing instances where the target was not defined and where the target would have been more than one second away. Afterwards the resulting dataset consisted of 693 instances of `speech` and 154 instances of `hrt` targets (see section 3.2.1).

4.1.1 Basic feature set

The metrics shown in Table 4.1 indicate that the overall prediction on this binary classification, which involves distinguishing between backchannels and other types of utterances, is slightly above chance. The gradient boosting machines perform notably better in correctly classifying at least some backchannels, as reflected in the higher but still considerably low MCC values, reaching ≈ 0.17.

Table 4.1: Metrics for classifying the subsequent PCOMP group (`speech` vs. `hrt`) using the basic feature set.

Model	F1	MCC	Balanced Accuracy
Extreme Gradient Boosting	**0.7701**	**0.1711**	0.5643
Light Gradient Boosting Machine	0.7678	0.1696	**0.5657**
Random Forest Classifier	0.7511	0.0754	0.5278

The Gini impurity based feature importance (see Figure 4.1), which was calculated exclusively from the fully trained model, shows a completely different order of the feature importances than those derived from the mean absolute SHAP values, which are, on the other hand, computed utilizing the training set (figure 4.2). Based on the Gini feature importance the most powerful features are, (1) whether the breathing annotation overlaps with the interlocutor and (2) whether

it ends with an inbreath. These two features have a clear assignment towards the output target, which is shown in Figure 4.4. Both the absence of overlap with the interlocutor and the end of the breathing annotation with an inbreath indicate that the target is `speech`. Despite their clear tendency, they are ranked at the bottom in terms of their overall impact on the model output in the SHAP feature importance plot.

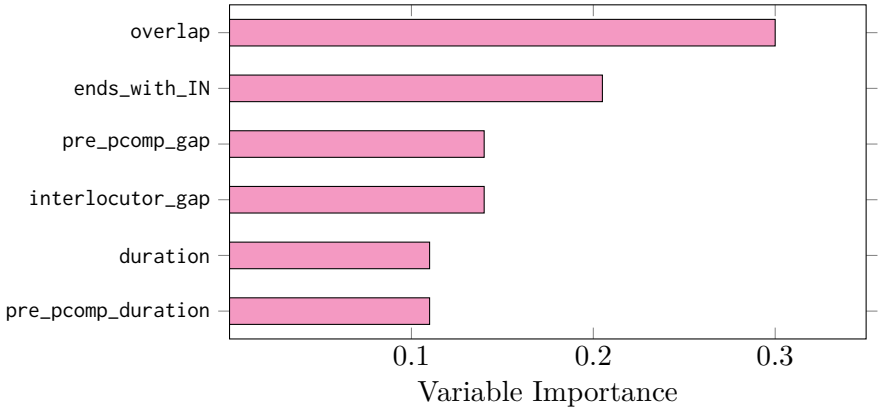


Figure 4.1: Gini impurity based feature importance for the target "subsequent PCOMP group" utilizing the basic feature set (i.e., no acoustic features at all).

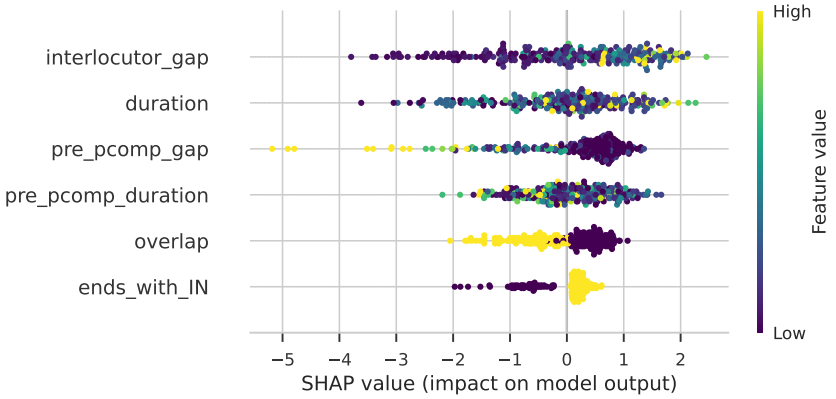


Figure 4.2: Impact of the model output based on the SHAP values for the target "subsequent PCOMP group" utilizing no acoustic features. Class mapping: backchannel: 0, speech: 1. The features are presented in a hierarchical order, with the most impactful features positioned at the top.

4.1.2 Including relative intensity and breathing volume estimate

The addition of the relative intensity and the breathing volume estimate to the basic feature set resulted in a 40% improvement in the best MCC score, that is a value of ≈ 0.29. However, this best value for the MCC is now reached by the LightGBM model (table 4.2 on the facing page). The balanced accuracy on the other hand only improved by 0.04 and the F1-Score even less by adding these two features.

Utilizing this feature set, the relative intensity is the most important feature, as indicated by both the Gini importance (figure 4.3) and the mean absolute SHAP values (figure 4.4). Note that the Gini importances changes fundamentally when the relative intensity and the breathing

Table 4.2: Metrics for classifying the subsequent PCOMP group using the basic feature set including the relative intensity and the breathing volume estimate.

Model	F1	MCC	Balanced Accuracy
Light Gradient Boosting Machine	**0.7989**	**0.293**	**0.6164**
Extreme Gradient Boosting	0.7812	0.2177	0.5905
Random Forest Classifier	0.7674	0.1519	0.5486

volume estimate are included in the feature set compared to when they are not. The overlap and whether the breathing annotation ends with an inbreath now have the lowest values, while they previously had the highest calculated Gini based importances. This can be explained by two facts. Firstly, the overall prediction scores are not particularly high. Secondly, the introduction of new features, particularly those that prove to be important, such as the relative intensity, can result in significant alterations to the tree-structure of the trained tree-based machine learning models. As the experiments with the other targets (section 4.2 and 4.3) showed similar results, we refrain from further depicting the Gini importances in the results.

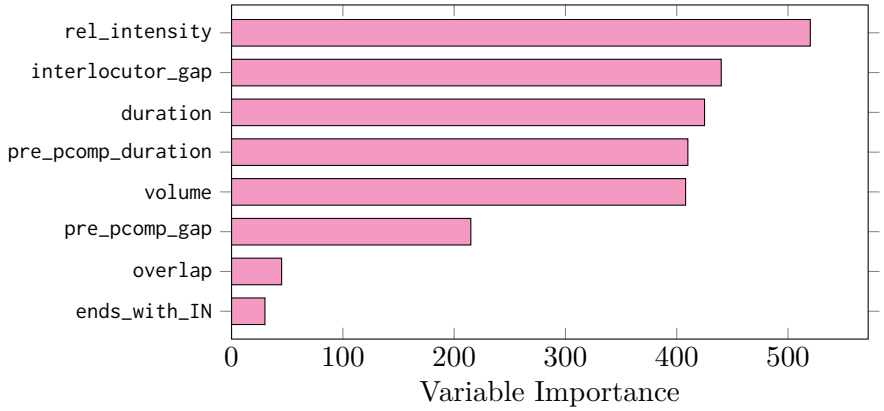


Figure 4.3: Gini impurity-based feature importance for the target "subsequent PCOMP group" when including the relative intensity and the breathing volume estimate, in addition to the basic contextual feature set.

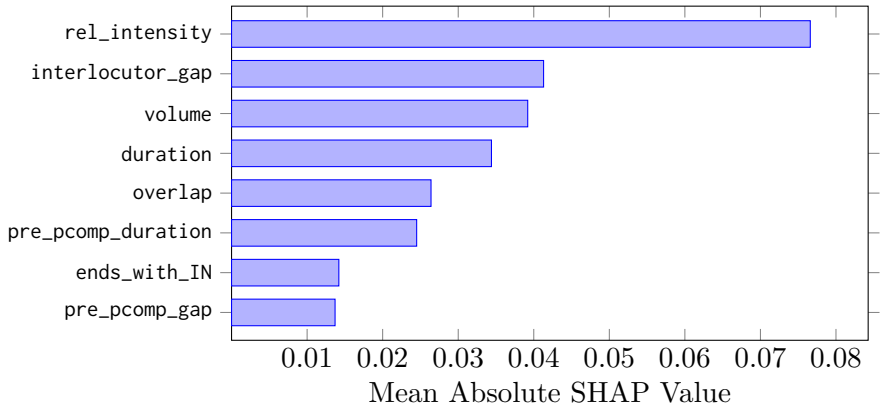


Figure 4.4: Mean absolute SHAP values for the target "subsequent PCOMP group" utilizing the basic feature set including the relative intensity and the breathing volume estimate.

The impact of the features on the model output is illustrated in Figure 4.5. The most notable aspect is that high values of the relative intensity serve as the most reliable indicator that no hearer-response-token will follow. Furthermore, the fact that the breathing is not overlapping with the interlocutor and that it ends with an inbreath has a clear but weak impact that the target subsequent PCOMP group is predicted as speech.

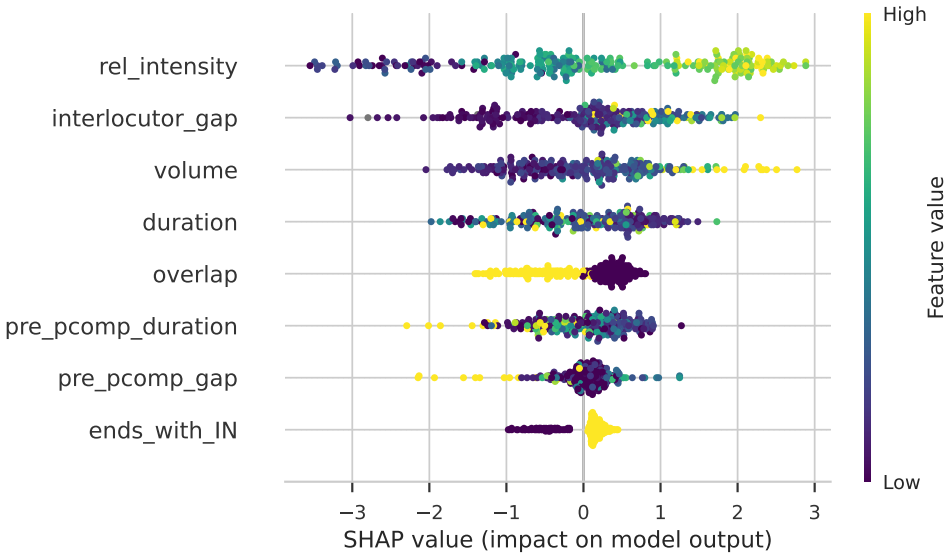


Figure 4.5: Impact of the model output based on the SHAP values for the target "subsequent PCOMP group" utilizing the basic feature set including the relative intensity and the estimate for the breathing volume. Class mapping: backchannel: 0, speech: 1. The features are presented in a hierarchical order, with the most impactful features positioned at the top.

The duration of breath sound is ranked moderately in terms of feature importance. However, its impact on the model output is not immediately clear. Figure 4.6 shows the relationship between the duration of breath sounds and their corresponding SHAP values. For durations up to approximately 200 ms, the SHAP values are negative, indicating a tendency towards an hrt prediction. The durations between about 0.2 and 1.0 seconds are predominantly indicative for speech to follow, although for durations around 800 milliseconds, the SHAP values are slightly below zero once more. As the durations extend further, ranging from approximately 1.0 to 2.2 seconds, the SHAP values fluctuate around zero and are predominantly negative. For even longer durations, the SHAP values rise above zero and continue to increase. The longest durations above 3.8 seconds are negative once more, although these are exceptional cases. Although the plot is coloured based on relative intensity, which is likely the feature with the highest interaction with duration according to the SHAP python package (Lundberg & Lee, 2017), the interaction effect between duration and intensity does not provide a distinct pattern.

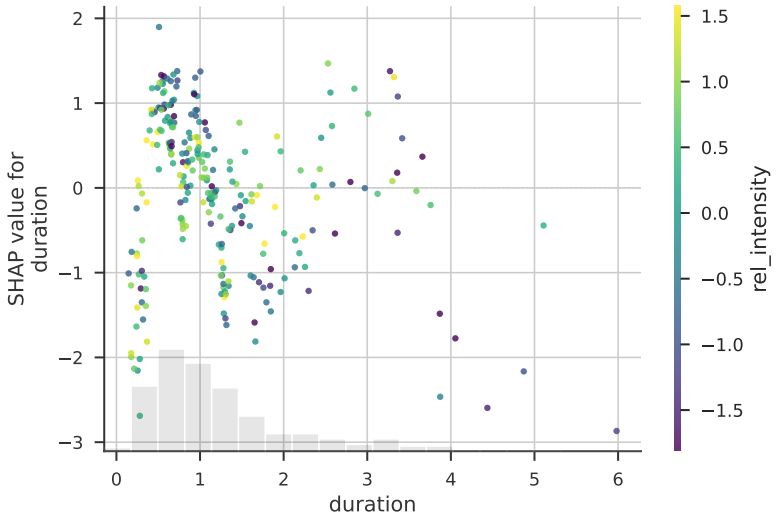


Figure 4.6: SHAP dependence plot for the breath sound's duration, colored by the relative intensity feature, which potentially exhibits the highest interaction. Target is the subsequent PCOMP group. Target mapping: `hrt` $= 0$, `speech` $= 1$.

4.1.3 Final feature set

As shown in Table 4.3, utilizing any selection of up to four additional acoustic functional features (see Table 3.3) for classifying the subsequent PCOMP group did not in fact lead to any further improved metric scores. Instead, the feature permutation importances for about half of the features exhibited values close to zero or even negative, indicating an overfitting of the model. This is further corroborated by LightGBM's warning that no further splits with positive gain could be found. Reducing the maximum depth of the tree, or adjusting the split criteria, for example, `n_child_weight` or `gamma` in XDGBoost or the equivalent parameters in LightGBM, see section 3.7.2) did not result in a notable improvement of the prediction. Consequently, we refrain from providing a detailed description of the results, in view of the limitations for the interpretation of SHAP values in this context.

Table 4.3: Metrics for classifying the subsequent PCOMP group using the full final feature set.

Model	F1	MCC	Balanced Accuracy
Light Gradient Boosting Machine	**0.7953**	**0.2873**	0.6092
Extreme Gradient Boosting	**0.7953**	0.2762	**0.6123**
Random Forest Classifier	0.7759	0.1891	0.5651

4.2 Classification of the preceding PCOMP group

This experiment was a multiclass classification task. The target groups were `hold` (the same speaker will continue speaking after the breath sound), `change` (the same speaker most likely just yielded the turn) and `hrt` (the speaker gave a hearer-response-token). The dataset was cleaned by removing instances where the target was not defined, instances where the preceding annotation was breathing or laughter (because these were very unfrequent), and instances where the target would have been more than one second away (see section 3.2.1). The final dataset for this experiment contained 427 `hold`, 272 `change` and 211 `hrt` instances as the target.

4.2.1 Basic feature set

For the classification of the preceding PCOMP group with solely contextual features, the Random Forest Classifier achieved the highest scores across all three metrics, with an F1 score of 0.5681, an MCC of 0.3228, and a Balanced Accuracy of 0.5449. As Table 4.4 indicates, the Random Forest Classifier has a slight advantage over the other models, but the differences are relatively modest.

Table 4.4: Metrics for classifying the preceding PCOMP group using the basic feature set.

Model	F1	MCC	Balanced Accuracy
Random Forest Classifier	**0.5681**	**0.3228**	**0.5449**
Light Gradient Boosting Machine	0.5475	0.2924	0.5286
Extreme Gradient Boosting	0.5273	0.2655	0.5088

4.2.2 Including relative intensity and breathing volume estimate

The addition of relative intensity and breathing volume estimate to the feature set did not demonstrably enhance the performance of the Random Forest Classifier. However, the XDGBoost and lightGBM models both outperformed it, as Table 4.5 shows. The XDGBoost model was slightly superior and achieved a MCC score of 0.3716 and a balanced accuracy of 0.5762.

Table 4.5: Metrics for classifying the preceding PCOMP group using the basic feature set including the relative intensity and the breathing volume estimate.

Model	F1	MCC	Balanced Accuracy
Extreme Gradient Boosting	**0.5963**	**0.3716**	**0.5762**
Light Gradient Boosting Machine	0.5937	0.3672	0.5721
Random Forest Classifier	0.5731	0.3322	0.5481

4.2.3 Final feature set

The addition of several acoustic functional features leads to the Random Forest classifier achieving scores that are further beyond those of the gradient boosting models presented in section 4.2.2. However, the gradient boosting models themselves did not either markedly improve or decrease in performance. The best MCC score of 0.3951 for classifying the preceding PCOMP group is clearly better than the best one reached classifying the subsequent (≈ 0.29, see section 4.1.2). This is particularly noteworthy given that the classification task involved three target groups, rather than two.

Table 4.6: Metrics for classifying the preceding PCOMP group using the full final feature set.

Model	F1	MCC	Balanced Accuracy
Random Forest Classifier	**0.61**	**0.3951**	**0.5886**
Extreme Gradient Boosting	0.5945	0.3685	0.5775
Light Gradient Boosting Machine	0.5911	0.3652	0.5725

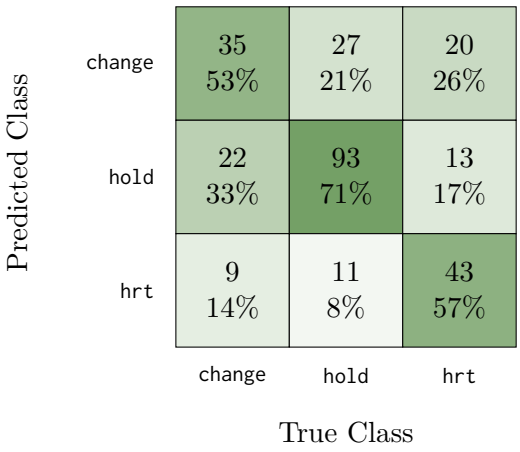


Figure 4.7: Confusion matrix for the target "subsequent PCOMP group" utilizing the full final feature set.

Figure 4.7 shows the confusion matrix for the prediction of the subsequent PCOMP group. It reveals that the most probable error is the false prediction of a `change` as an `hold`. The second most likely error is the classification of an `hrt` as a `change`. Note the probability of the class `hrt` being correctly predicted is greater than that of the `change` class, despite the fact that the dataset contained more instances of `change`. One potential explanation for the class `hold` exhibiting the highest rate of true positives is that this group had the most occurrences. Figure 4.9 shows the contribution and tendency of the features on each target.

4.2.3.1 Feature importances

The magnitude of the impact of each feature for classifying each target group can be derived from the mean absolute SHAP values, as illustrated in Figure 4.8. It is evident that the length of the preceding PCOMP annotation is the most crucial factor in classifying its PCOMP group. The relative intensity has the second-strongest overall influence, especially for the target group being `hold`. The third most influential variable is whether the breath sound overlaps with the interlocutor. The time gap to the end of the interlocutor's last PCOMP annotation also plays a relatively strong role in predicting both `hold` and `hrt`, whereas all other features have a much smaller overall impact.

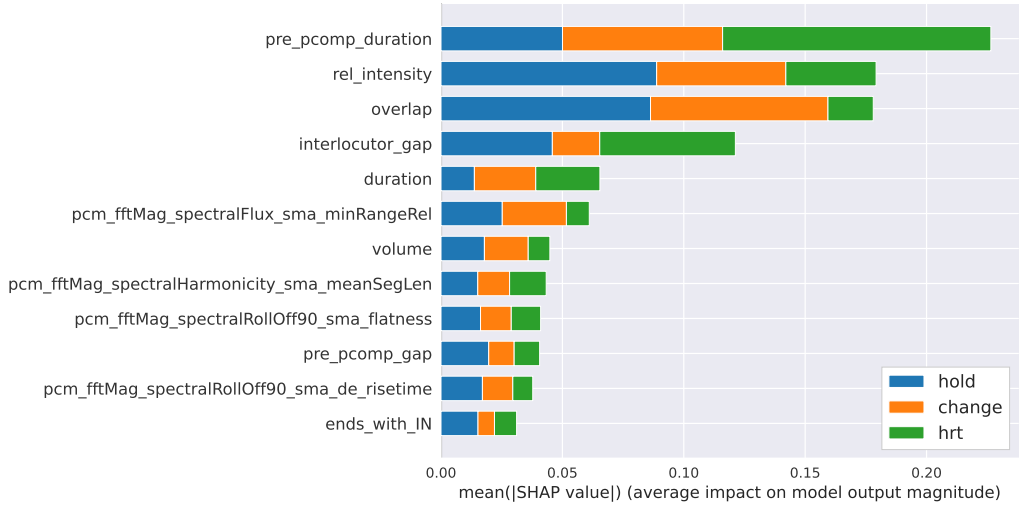


Figure 4.8: Mean absolute SHAP values for the target "subsequent PCOMP group" utilizing the full final feature set.

4.2.3.2 Effects of the most important features

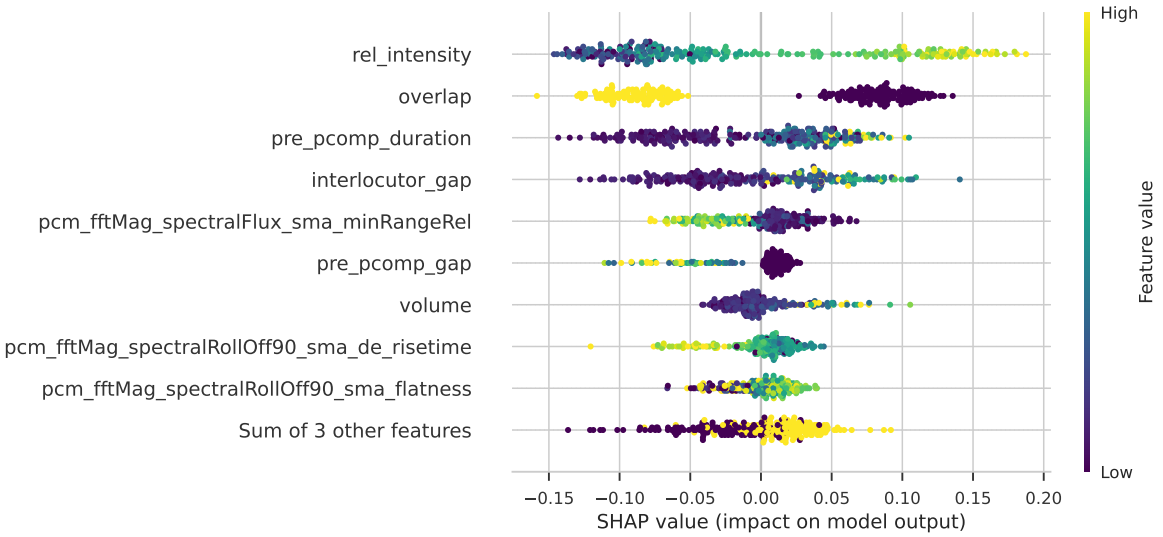
Previous PCOMP duration The duration of the previous PCOMP annotation is a feature with a special status, as it directly represents a characteristic of the target class in this experiment. The shorter the duration of the previous PCOMP annotation, the more likely it is to be a hearer-response token (hrt), which is as expected, given that an hrt usually contains fewer words on average than the other PCOMP annotations. In principle, the effect of this feature is reversed with regard to the classification of the other two classes. Nevertheless, in certain instances, short durations may also potentially support the detection of these groups.

Relative intensity Figure 4.9(a) illustrates that a high relative intensity has the greatest impact on the probability of the current speaker continuing to speak after a breath sound. The relationship between the relative intensity and its corresponding SHAP value is quite linear in nature. The relative intensity is also a highly ranked feature for the classification of the other targets, change and hrt. Higher values for the relative intensity increase the likelihood of the prediction for both change and hrt, as shown in Figures 4.9(b) and 4.9(c).

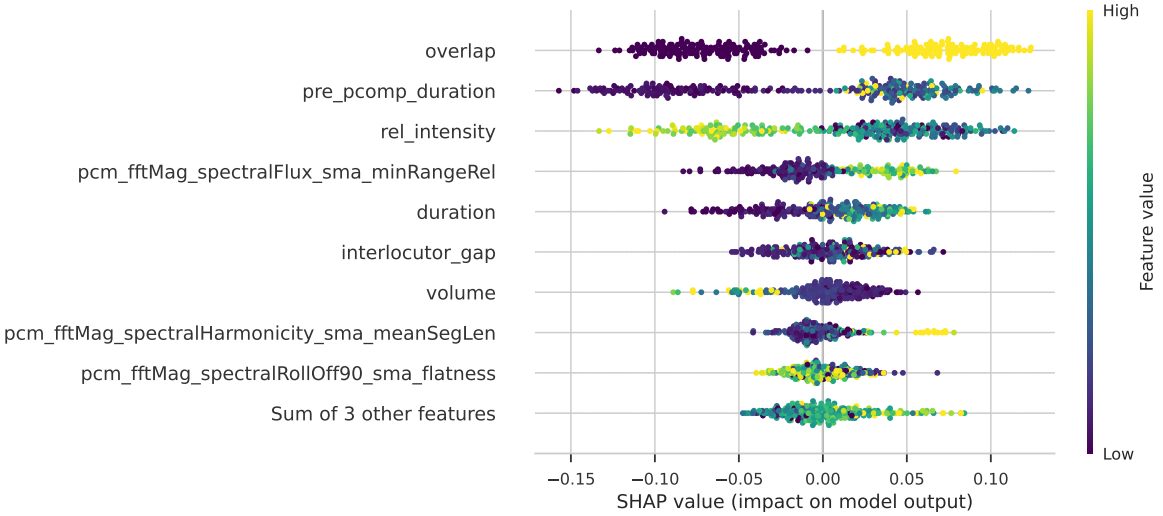
Overlap The second most impactful feature is whether the breath sound is in overlap with the interlocutor. Note that SHAP values have a magnitude greater than 0.05 in almost all cases. They are always positive when overlap is present and negative when not. The breath sound being in overlap with the interlocutor leads to an increased chance of the preceding PCOMP annotation being predicted as a change, as shown in Figure 4.9(b).

Duration Figure 4.8 shows that the duration of the breath sound has a comparatively minor impact on the prediction of the class hold, in comparison to the other two classes. However, the Figures 4.9(b) and 4.9(b) show that for the classes change and hold, higher values for the duration indicate a change being the preceding PCOMP group, while lower values indicate an hrt.

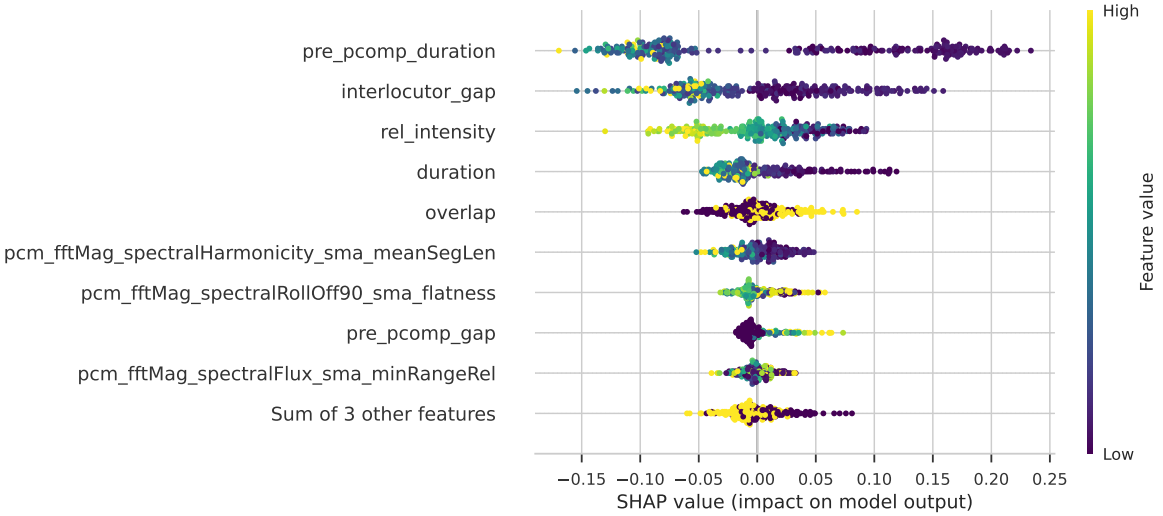
Interlocutor gap The feature importance overview depicted in 4.8 reveals that the temporal gap from the breath sound's start to the end of the most recent PCOMP annotation of the interlocutor has the least impact on the target group change. Figure 4.9(b) shows no direct relationship between this feature and the target class change, except that very high values support it. This is consistent with our expectations, given that the longer the interlocutor has not said anything, the more likely the speaker will eventually yield the turn. Furthermore, a high value for this feature is indicative of the target PCOMP group being an hold in favour of an hrt.



(a) Target class: `hold` — Preceding PCOMP group.



(b) Target class: `change` — Preceding PCOMP group.



(c) Target class: `hrt` — Preceding PCOMP group.

Figure 4.9: Beeswarm plots for the preceding PCOMP group for all target classes. The features are presented in a hierarchical order, with the most impactful features positioned at the top.

4.3 Interlocutors most recent PCOMP group

This experiment was a multiclass classification task for predicting the most recent PCOMP group of the interlocutor, as defined in section 3.2.1. The target classes were `hold`, `change` and `hrt`. After the cleaning process, the dataset of this experiment contained 373 `hold`, 221 `change` and 186 `hrt` instances.

4.3.1 Basic feature set

The experiment classifying the interlocutor with only the durational and contextual features (basic feature set) yielded to the least successful result compared to the other two targets (sections 4.1 and 4.2). All three machine learning models performed about equally poorly, with an F1 score of ≈ 0.49, an MCC of ≈ 0.2, and a balanced accuracy of ≈ 0.45.

Table 4.7: Metrics for classifying the interlocutor's most recent PCOMP group using the basic feature set.

Model	F1	MCC	Balanced Accuracy
Extreme Gradient Boosting	**0.4912**	**0.2002**	**0.4596**
Random Forest Classifier	0.4906	0.1994	0.4515
Light Gradient Boosting Machine	0.4869	0.1951	0.4546

4.3.2 Including relative intensity and breathing volume estimate

Adding the relative intensity and the estimate for the breathing volume to the feature set, the Random Forest Classifier improved notably, compared to the other models when they did not utilize those features. The F1 score improved by 0.04 to ≈ 0.53, the balanced accuracy improved to ≈ 0.49, and most notably, the MCC improved by ≈ 0.06 and reached ≈ 0.27.

Table 4.8: Metrics for classifying the interlocutor's most recent PCOMP group using the basic feature set including the relative intensity and the breathing volume estimate.

Model	F1	MCC	Balanced Accuracy
Random Forest Classifier	**0.5313**	**0.267**	**0.4914**
Extreme Gradient Boosting	0.5182	0.242	0.4811
Light Gradient Boosting Machine	0.51	0.227	0.4709

4.3.3 Final feature set

The Random Forest Classifier was identified as the most optimal model when utilizing the final feature set, resulting in an MCC score that was ≈ 0.1 higher than that achieved when acoustic functional features were not utilized. This increase represents the biggest benefit in performance via the final feature set across all experiments. The metric values of this model were: MCC ≈ 0.37, the F1 was ≈ 0.59, and the balanced accuracy was ≈ 0.56. Nevertheless, the other models also demonstrated comparable performance, as shown in Table 4.9.

Table 4.9: Metrics for classifying the interlocutor's most recent PCOMP group using the full final feature set.

Model	F1	MCC	Balanced Accuracy
Random Forest Classifier	**0.5946**	**0.3704**	**0.5561**
Light Gradient Boosting Machine	0.5817	0.3433	0.5447
Extreme Gradient Boosting	0.582	0.3432	0.5487

Figure 4.10 shows the confusion matrix for the prediction of the interlocutor's most recent PCOMP group. The 63% of the instances of `hold` and `hrt` get classified correctly, but only 52% of the `change` instances. Note that the probability of the class `hrt` being correctly predicted is greater than that of the `hold` class, despite the fact that the dataset contained more than twice as instances of `hold`. The most likely confusion is between `change` and `hold`.

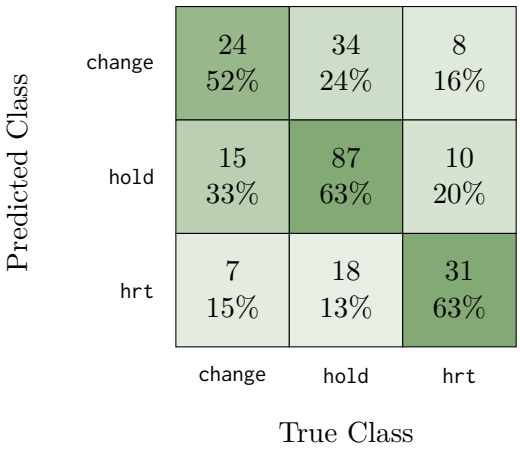


Figure 4.10: Confusion matrix for the target "most recent interlocutors PCOMP group" utilizing the full final feature set.

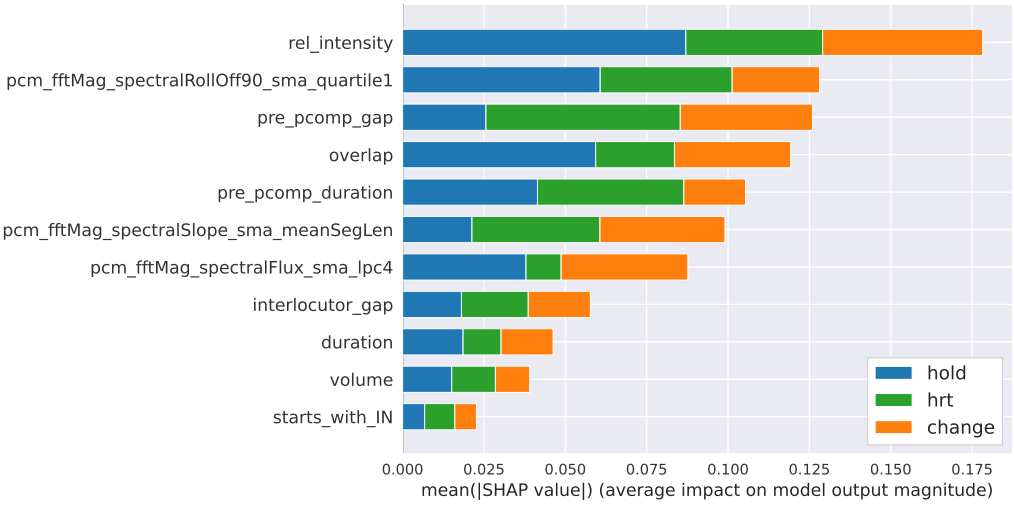


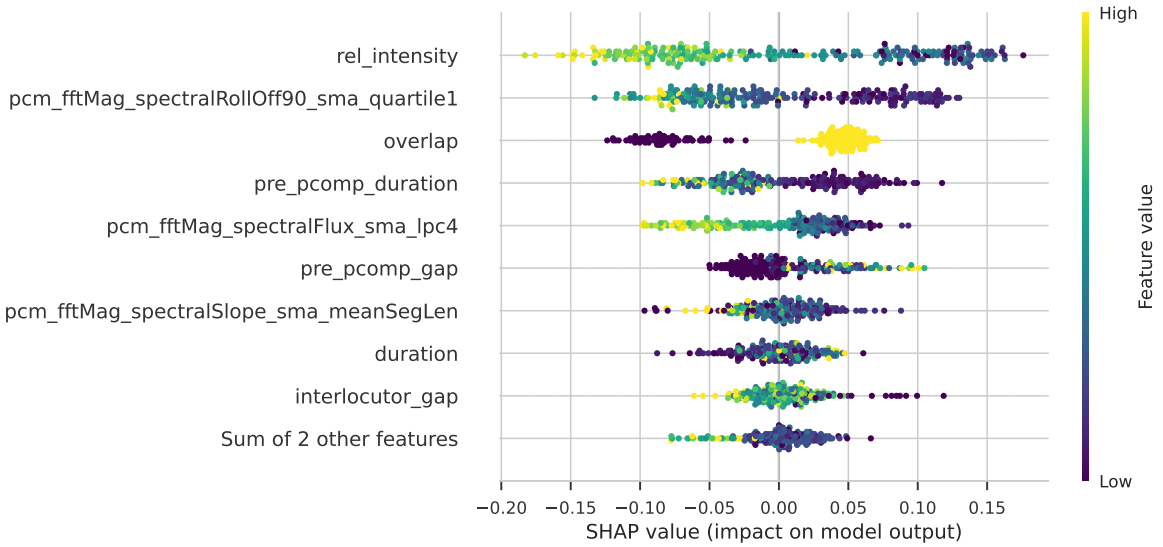
Figure 4.11: Mean absolute SHAP values for the target "most recent interlocutors PCOMP group" utilizing the final feature set.

Feature importances Figure 4.11 illustrates the importance of each feature in classifying the target groups, as indicated by the mean absolute SHAP values. The relative intensity of breathing has the greatest influence on the classification result, and this by a clearly identifiable lead. In addition to the context features `overlap`, `pre_pcomp_gap` and `pre_pcomp_duration`, the acoustic functional features `spectralRollOff90_sma_quartile1` (the first quartile of the spectral roll-off at 90 percent), `spectralSlope_sma_meanSegLen` (the mean segment length of the spectral slope), and `spectralFlux_sma_lpc4` (the 4th linear prediction coefficient of the spectral flux) also have a comparable impact on the models output.

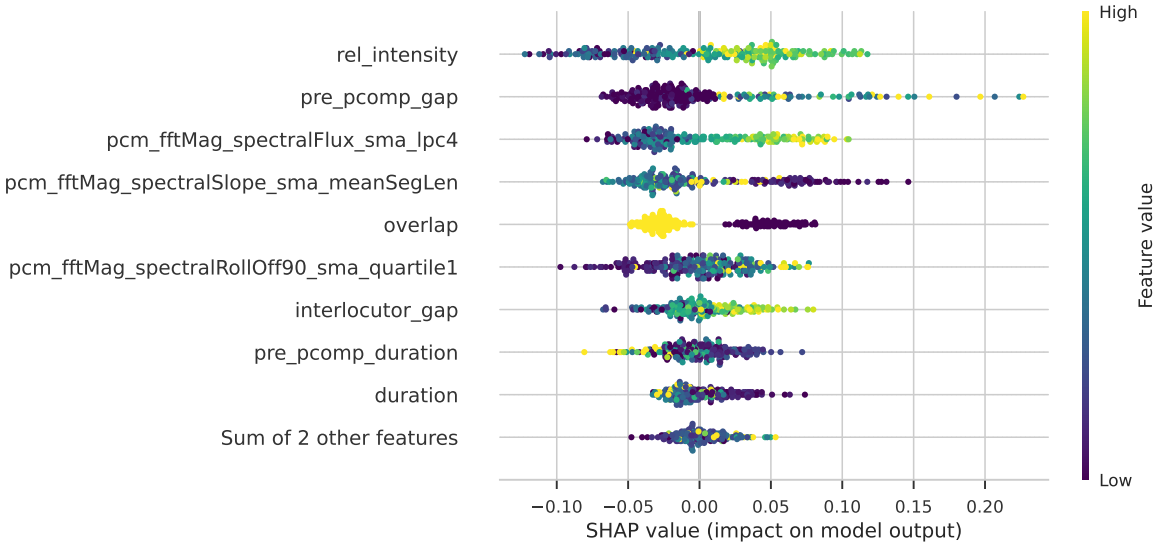
Prediction of `hold` The features and their impact on whether they contribute to the interlocutor's most recent PCOMP group being classified as `hold` is shown in Figure 4.12(a). Several features have a rather direct relationship to the prediction of the model being an `hold`: a low relative intensity, a low value for the first quartile of the spectral roll-off at 90 percent, and present overlap of the breathing with the speech of the interlocutor.

Prediction of `change` The features that contribute to the prediction of the `change`-group are illustrated in Figure 4.12(b). A medium to high relative intensity shifts the models output towards a change, as well as the temporal gap to the preceding PCOMP annotation being relatively long. Moreover, medium to high values for the fourth LPC coefficient of the spectral flux and the breathing not being in overlap with the interlocutor also have an impact on the classification of a change. The behaviour of the mean segment length of the spectral slope is more differentiated. Low values tend to increase the likelihood of a change being present, high values have little impact, and medium values have a slight impact on the target not being a change.

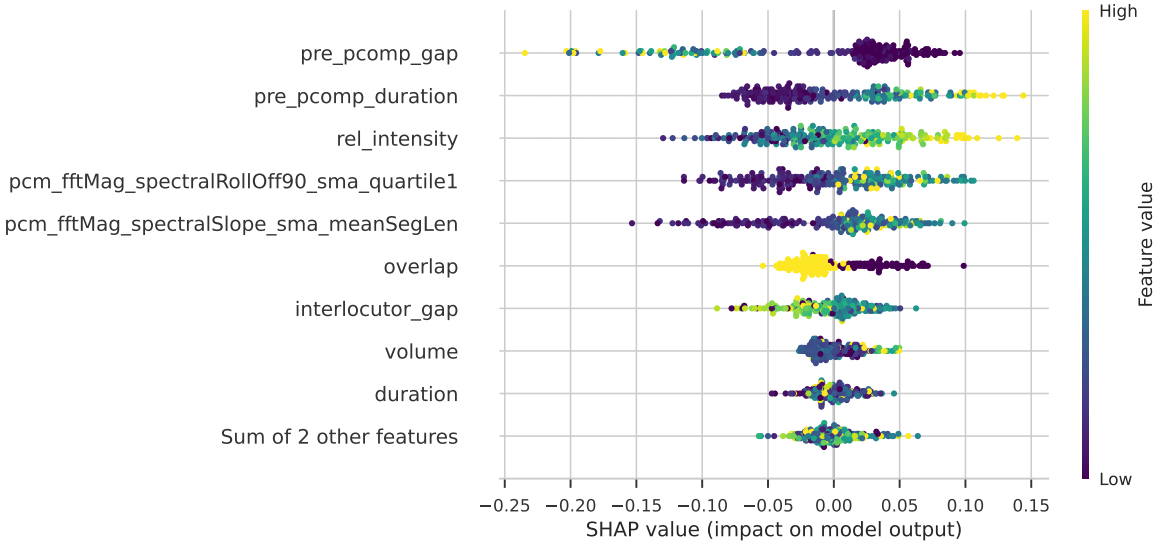
Prediction of `hrt` The features that most impact the prediction of whether an `hrt` is the interlocutor's most recent PCOMP annotation (which is likely in cases where the current speaker maintains the turn) are the temporal gap to the preceding PCOMP annotation and its duration (see Figure 4.12(c)). The shorter the gap and the longer its duration, the more likely it is an `hrt`. Furthermore, the SHAP values indicate that the acoustic features have a smaller capacity to predict an `hrt` of the interlocutor than to predict a yielded or kept turn of the interlocutor. Additionally, a higher relative intensity and a higher value for the first quartile of the spectral roll-off at 90 percent increase the likelihood of an `hrt` being the interlocutor's most recent PCOMP.



(a) Target class: `hold` — Interlocutors most recent PCOMP group.



(b) Target class: `change` — Interlocutors most recent PCOMP group.



(c) Target class: `hrt` — Interlocutors most recent PCOMP group.

Figure 4.12: Beeswarm plots for the prediction of the interlocutor's most recent PCOMP group for all target classes. The features are presented in a hierarchical order, the most impactful ones are on top.

4.4 Comparison of the experiments

From the three experimental groups with different targets (i.e., *preceding PCOMP group*, *subsequent PCOMP group*, and *most recent interlocutors PCOMP group* — section 4.1, 4.2, and 4.3) our method achieved the most notable improvement by utilizing acoustic functional features for the classification of the most recent interlocutors PCOMP group.

The features that showed to be important across all experiments were the relative intensity, whether the breathing is in overlap with the interlocutor, and the duration of the previous PCOMP annotation. The duration of the breath sound was found to be a relatively important factor in the classification of the self-subsequent, less important in the self-preceding, and least important in the interlocutor's most recent PCOMP group. The temporal gap between the start of the breath sound and the end of the most recent PCOMP of the interlocutor was found to be of importance for the prediction of PCOMP groups of the same speaker. In contrast, the temporal gap between the breath sound and the self-preceding PCOMP was found to be more important for the classification of the interlocutor's most recent PCOMP group. Moreover, across all three experiments, the features indicating whether the audible breathing started or ended with an inbreath or an outbreath was consistently ranked as one of the least important features based on mean absolute SHAP values. However, despite their low importance ranking, these features showed a clear tendency towards specific classes and were decisive in the models' outputs.

5 Discussion

The data employed in this thesis differs substantially from that used in the majority of other studies researching turn-taking in the literature. In this study, we utilized the GRASS corpus and its annotations of Points of Potential Completion (PCOMPs) (Kelterer & Schuppler, subm.). The PCOMP annotations provide a comprehensive representation of all turn-relevance places, offering a more detailed temporal structure than the usual turn-taking annotations, which are typically comprised of longer talk spurts separated by pauses.

5.1 The Impact of audible breathing on turn-taking in conversation

The main aim of this thesis was to examine the impact of breath sounds on turn-taking in conversation. In order to achieve this, the turn-taking behaviour was examined by taking the breath sound's point of view, and identifying relationships between its characteristics and its surrounding PCOMP annotations. Nevertheless, the PCOMP labels were not employed in their original form. Instead, they were grouped to preserve the most relevant turn-taking categories: `hold` (the same speaker continues), `change` (the other speaker most likely continues speaking), and `hrt` (hearer response token). Additionally, `speech` was used as a conjunction of `hold` and `change` (the detailed mapping is shown in Table 3.2). We related the breath sounds to three targets: (1) its self-subsequent, (2) the self-preceding, and (3) the most recent interlocutors PCOMP group. The investigation employed various feature categories to gain insight into the impact of audible breathing. The contextual features considered whether the breath sound overlapped with the interlocutor and whether it started or ended with an inbreath or outbreath. The durational features included, for example, the pause before self and other, and the duration of the breath sound. The acoustic features encompassed the relative intensity as defined by Werner et al. (2023), an estimate of the breathing volume, and a selection of functionals derived from other spectral features.

5.1.1 Initial outcomes

The initial experiments employed a basic feature set that contained solely contextual and durational features to classify the three target PCOMP groups. The results indicated that the durational features of breath sounds and whether it occurs in overlap with the interlocutor are indeed linked to turn-taking. The exclusive utilization of this basic feature set, comprising features that were all known at the time the breath sound ended, resulted in an above-chance prediction for the surrounding PCOMP groups in all cases. The MCC value is observed to be the highest for prediction of the preceding PCOMP group of the same speaker. This is likely a consequence of the target PCOMPs duration being employed directly as a feature in that situation.

5.1.2 Which features have an impact on which PCOMP group?

Inbreath vs outbreath We investigated both inbreaths and outbreaths, as well as combinations of both, because in our data several subsequent breath sounds (i.e., with no pause in between) were annotated in a single time interval. This phenomenon was captured by two binary features: whether the annotation begins with an inbreath, and whether it ends with an inbreath. Contrary to our initial expectations, this factors did not have an important impact on the performance of our classification models. One potential explanation for this de facto ineffectiveness is that these features are already largely reflected in others. Firstly, the relative intensity of inbreaths is typically stronger than those of outbreaths (Werner et al., 2023). Secondly, as a general rule, the duration of outbreaths is shorter (see Figure 2.4). Thirdly, the gap to the preceding PCOMP annotation is a strong indicator of this too, as outbreaths in our data are mostly annotated directly after speech (see Figure 2.12).

Duration One of the most discussed characteristics of a breath sound is its duration. However, the duration feature had limited interpretive power in this thesis, which may be attributed to the inclusion of both inbreaths and outbreaths, as well as their combinations, in our dataset. In particular, the outbreaths in our dataset were substantially shorter on average than the inbreaths. Consequently, our experiments did not yield a clear picture of the feature's predictive power on turn-taking. Either the duration's feature importance was moderately weak or its relation to the model's output was rather complex (see Figure 4.6). Nevertheless, one of the observations made was that very short breaths were more likely to occur after an `hrt` (backchannel), while longer ones are an indicator for the preceding being the PCOMP group `change`, which means the speaker yielded the turn. The weakest impact of the duration was on the prediction of the interlocutor's most recent PCOMP group. In essence, these outcomes align with the findings of Ishii et al. (2014), although it is important to emphasize that their experimental design is noticeably different, as they analysed respiratory data in multiparty conversations. In scenarios where the speaker changes, they identified the duration of inhalation as a reliable indicator of who would speak next, with a similar tendency to that observed in our experiments. Concerning the duration of inbreaths, in cases when the speaker is continuing to speak, previous studies have shown conflicting results (see section 1.3.3). Our observations, which reveal a nuanced pattern of the impact of the duration of breath sounds, suggest that these apparent contradictions may actually arise from limitations in research methods and the utilized data.

Estimate of the breathing volume and other intensity features The breathing volume estimate, which was introduced by multiplying the duration with the per speaker normalized mean RMS of the breath sound, was found to be a more reliable indicator than the duration in only one instance: when attempting to estimate whether the breath sound was following `speech` or `hrt`. Furthermore, we could not determine any impact of non-relative RMS features on our classification, although such straight-forward intensity features appeared to us to be the most tangible of all.

Durational features The temporal gap between the breath sound and the preceding utterance of the speaker was found to play a substantial role in the classification of the most recent PCOMP group of the interlocutor. Low gaps to the self-preceding speech indicate that the interlocutor's

most recent PCOMP was an `hrt`. This result appears to be a logical consequence of typical patterns in conversational speech. When a speakers last utterance is very recent, it is more probable that the interlocutor has just given a backchannel, rather than interrupting the other speaker.

Overlap Breath sounds in overlap with the interlocutor indicate a tendency for the speaker to provide a hearer response token subsequently. In contrast, the tendency of the overlap status is less clear towards the preceding PCOMP. In summary, a breath sound that does not overlap with the speech of the interlocutor serves as a clear indicator that the speaker will take the next turn. Overall, the impact of the overlap status of breath sounds on turn-taking is moderate, but the tendency is mostly evident. Explaining the results of the classification experiment on the interlocutor's most recent PCOMP group, the outcome is traceable from two perspectives. Firstly, if the breath sound occurs during overlap, the interlocutor may have already started continuing their turn. Secondly, if a speaker starts breathing while the interlocutor has not completely finished an utterance, the breath sound might go unnoticed. This aligns with the hypothesis by Schegloff (1996) that breath sounds can be shaped to be heard more or less, and that strongly audible inhalations might indicate "an extended spate of talk to come", which likely excludes backchannels.

Relative intensity Across all of our target variables, the addition of the relative intensity in the feature set led to a clear increase in prediction scores, particularly in the MCC metric. Overall, a high value for the relative intensity was found to be a clear indicator for non-backchannel utterances to follow and for the speaker to either keep or take the turn. In general, it was found to be the most powerful indicator in our feature set, except for the classification of backchannels, where durational features were slightly more impactful. Note that the relative intensity was calculated via utilizing what is happening after the breath sound. This was achieved by utilizing the mean intensity two seconds after the end of the breath sound, which also captures, to some extent, the intensity, the duration, and the temporal gap of the subsequent utterance. In conclusion, these findings on the relative intensity are consistent with those of Ishii et al. (2014), at least when hypothesizing that the relative intensity can be related to the respiratory slope to some degree. Figure 5.1 shows the distribution of the original values for the relative intensity across different target PCOMP groups.

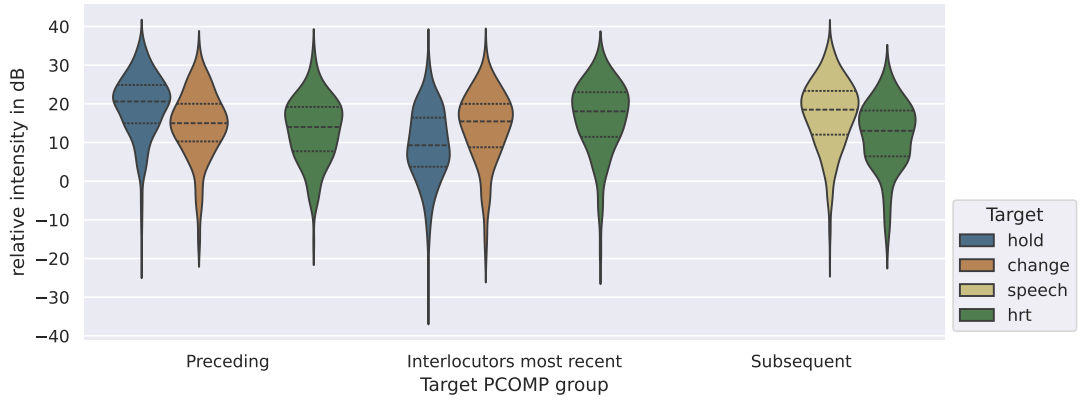


Figure 5.1: Distribution of the relative intensity across different target PCOMP groups.

Further acoustic features By classifying the preceding and subsequent PCOMP group of the current speaker, our method was unable to identify any further relevant and robust acoustic features. Nevertheless, three spectral features indicated whether a speaker is taking the turn after a breath sound or not. These features included: high values for the fourth linear prediction coefficient of the spectral flux, high values for the first quartile of the spectral roll-off at 90%, and very low values for the mean segment length of the spectral slope. However, it cannot be ruled out that these features did not represent the nature of the breath sound itself, but rather the interference of the interlocutor on the breathing's audio signal.

5.2 Breathing in a quesion-and-answer scenario

The results of our analysis do not directly support the findings of Torreira et al. (2015), who investigated the timing of breathing in the context of question-answer pairs. In their study, timing included both the durations of the breaths and the temporal gaps from the start of each breath to the end of the respective interlocutor's question. They found that inhalations preceding a response typically begin promptly after the interlocutor has finished asking a question. Figure 2.14(a) shows that in our data, after a question is finished, in more than 25% of the cases, the gap until the other speaker breathes in is longer than 4 seconds. However, this comparison is likely to be misleading, as the way we define the most recent interlocutors PCOMP and its temporal gap plays an important role (see section 3.4.2, especially Figure 3.4). In instances where the breath sound completely overlaps with the interlocutor's utterance (e.g., speech, laughter, or another vocal activity), this indicates that the questions end after the breath sound has ended. This would then result in the "incorrect" target being depicted in our figure. Furthermore, instances where no (audible) breathing occurs (i.e., the speaker responds with residual breath or the inbreath is either not audible or annotated) also skew this result. Nevertheless, our analyses partly support the conclusions of Torreira et al. (2015), as they demonstrate that inbreaths are less frequently (partially) overlapping with the end of a `question` or a `change` than with any other kind of PCOMPs from the interlocutor (see Figure 2.14(a)).

5.3 Limitations, potential improvements and further research

The methodology employed to examine the impact of audible breathing on turn-taking was novel. As a consequence, several factors that likely have a substantial effect on the interpretability of the results only became apparent towards the end of the study, and thus could not be addressed in our experiments, as this would have exceeded the scope of the Master's thesis.

Breath sounds overlapping with the interlocutor We have not considered the temporal degree of the overlap and its severity in relation to the overlapping elements, namely other breaths or laughter. As the timing of turn-taking has repeatedly been demonstrated to be of great importance, it seems prudent to utilize the overlap status not only as a binary feature, but in a more finely graded manner in future research.

Interferences in the breath sound's audio signal The hypothesis put forth by Schegloff (1996), which suggests that audible breathing is a practice that can be shaped, and therefore information is carried within, leads one to postulate that it is likely to find multiple acoustic features, especially functional ones that are related to turn-taking. As we have demonstrated, timing-related features, including whether the feature is overlapping, already indicate the consequences of a breathing. Nevertheless, in the instances where the breathing is overlapped, the interlocutor distorts the audio, rendering the acoustic features unreliable. Potential improvements for further research include: (1) using data where the breathing audio is less affected by the interlocutor; (2) imputing the acoustic features with mean or median values in such instances; or (3) conducting an analysis with more exclusive, more narrowly defined scenarios, which exclude those cases. Furthermore, it is not only the interlocutor who distorts the audio signal; sounds such as smacks, laughter, and other noises (e.g., the speaker playing with the headset microphone cable) may also affect the acoustic features, particularly some functional features. For example, the functional feature which depicts the maximum position of the derivation of the spectral roll-off is particularly susceptible to such interfering factors. Note that the frequencies of smacks, laughter, and noises within audible breathing are heavily speaker-dependent in our data (i.e., some individuals produce almost none of these sounds). Excluding them was also not a viable option given the already limited size of the dataset.

Speaker-dependence We observed substantial differences in the performance of the classifiers across different speakers. However, due to the time constraints and workload associated with this Master's thesis, a detailed exploration of this aspect was not possible. One potential reason for speaker-dependence is that, as previously mentioned, the stability and quality of certain features are likely influenced by a speaker's breathing characteristics (e.g., frequency of smacks). The overall predictive power of these features could be improved by making them more robust against such influences. One potential solution to this issue is to remove or reduce the impact of smacks by implementing a filter. Alternatively, the skewing effect of smacks could be mitigated by increasing the segment length and adjusting the smoothing algorithm. Another explanation for the speaker-dependent performance is that it may have been suboptimal that we did not normalize certain features per speaker, specifically the duration of the breath sound. Another observation we made while temporarily using SMOTE to address the imbalance of the subsequent PCOMP group targets (which was ultimately not employed, see section 3.7.2), was a resulting higher balanced accuracy. However, when examining the change in the metrics per speaker, we observed that approximately two-thirds of the speakers exhibited a substantial improvement in their predictions, while the remaining speakers exhibited a corresponding decrease.

Improving the classification targets In order to analyse the impact of breath sounds on turn-taking via classifying several surrounding target PCOMP labels, we grouped PCOMP labels to more general turn-taking classes. However, the assumption that all labels represent turn-taking is not completely true, because the label `question` does directly represent whether the question has been answered. In addition, it is possible that audible breathing may only have a significant impact within a subset of the labels within our chosen groups. Consequently, focusing on smaller, more precisely defined situations by directly utilizing PCOMP labels could

potentially yield further insights. Furthermore, with regard to our concrete methodology, in particular the definition of the target for the most recent PCOMP of the interlocutors, there is scope for improvement. For instance, it could be beneficial to include targets where the annotation has commenced a certain amount of time before the breathing started and is still ongoing when the breathing ended. Or another example would be to improve the criteria for dropping targets on the maximum distance of one second: a specific scenario that we might misinterpret is when the interlocutor asks a question and the speaker gives a short answer on residual breath, followed by an audible breath within one second after the end of the question. In such cases, the breathing may not be directly related to the action of the interlocutor, but rather to the preceding utterance. Finally, the interpretability of a similar study could be improved by either limiting it to inhalation or exhalation noises, or by omitting annotations in which only an outbreath is present.

The point-of-view is crucial Our methodology, which adopts the breathing point-of-view — where each breath is treated as an entry in our dataset — is novel. The predominant approach in turn-taking research typically involves analysing situations defined by inter-pausal units. In contrast, the PCOMP annotations we utilized provide a more nuanced framework, necessitating later simplification, at least for the experiments in this thesis. For instance, Ishii et al. (2014) used the moment 350ms before a new utterance as the point of view. Conducting a similar experiment focusing on audible breathing rather than respiration could serve as a well-defined starting point, potentially offering clearer insights with fewer contingencies. Such a methodology would also make it possible to investigate the impact of the sheer presence of an audible breathing. Furthermore, this would facilitate the in-depth analysis of the overlap status in such cases, after which the acoustic features, particularly those of non-overlapping breath sounds, could be examined in detail.

5.4 Conclusion

The main motivation and aim of this thesis was to gain more insight into whether audible breathing is more than a mere preparation for speech. Does it provide information on turn-taking in spontaneous speech, and do the ways breath sounds are shaped by the speaker also have a relation to what is happening in the conversation?

Particularly the acoustic domain of breathing has not been widely researched yet, although it is a process that accompanies us at all times. Most would intuitively agree that there is a relation, that's why we have set out to investigate this further. To address this, we employed the spontaneous speech part of the GRASS corpus. We adopted a broad approach, utilizing each annotation of audible breathing and investigating the relations to its surrounding annotations of Points of Potential Completion (PCOMPs). The PCOMP annotations offer a deep and systematic insight into turn-taking at all turn-relevance places in a conversation. Our findings largely support existing research and intuitive expectations, although the prediction scores of our classification models fell short of initial expectations. This discrepancy may be attributed to the inherent complexities of analysing highly spontaneous speech: we extracted

data from conversations between people who had known each other for years and had very natural conversations, including lots of laughter and overlap, which led to lots of variation in the data. Especially contextual and durational features emerged as decisive factors in turn-taking. It was observed that the audibility of the inbreath, which accounts factors like whether it is in overlap with the interlocutor, or its relative intensity, are (indeed) promising predictors of the likelihood of the speaker taking the subsequent turn. In contrast, backchannels were more frequently provided on residual breaths (where no audible inbreath was preceding), or tended to be accompanied by longer audible breathing. The interpretation of acoustic features also posed challenges due to methodological limitations. Nevertheless, by outlining these limitations and potential workarounds and by providing a comprehensive discussion, this thesis establishes a foundation for future quantitative research on the relationship between audible breathing and turn-taking in conversation. Overall, I hope this thesis will serve as a motivation for further investigation of this still highly underexplored research domain.

A Appendix

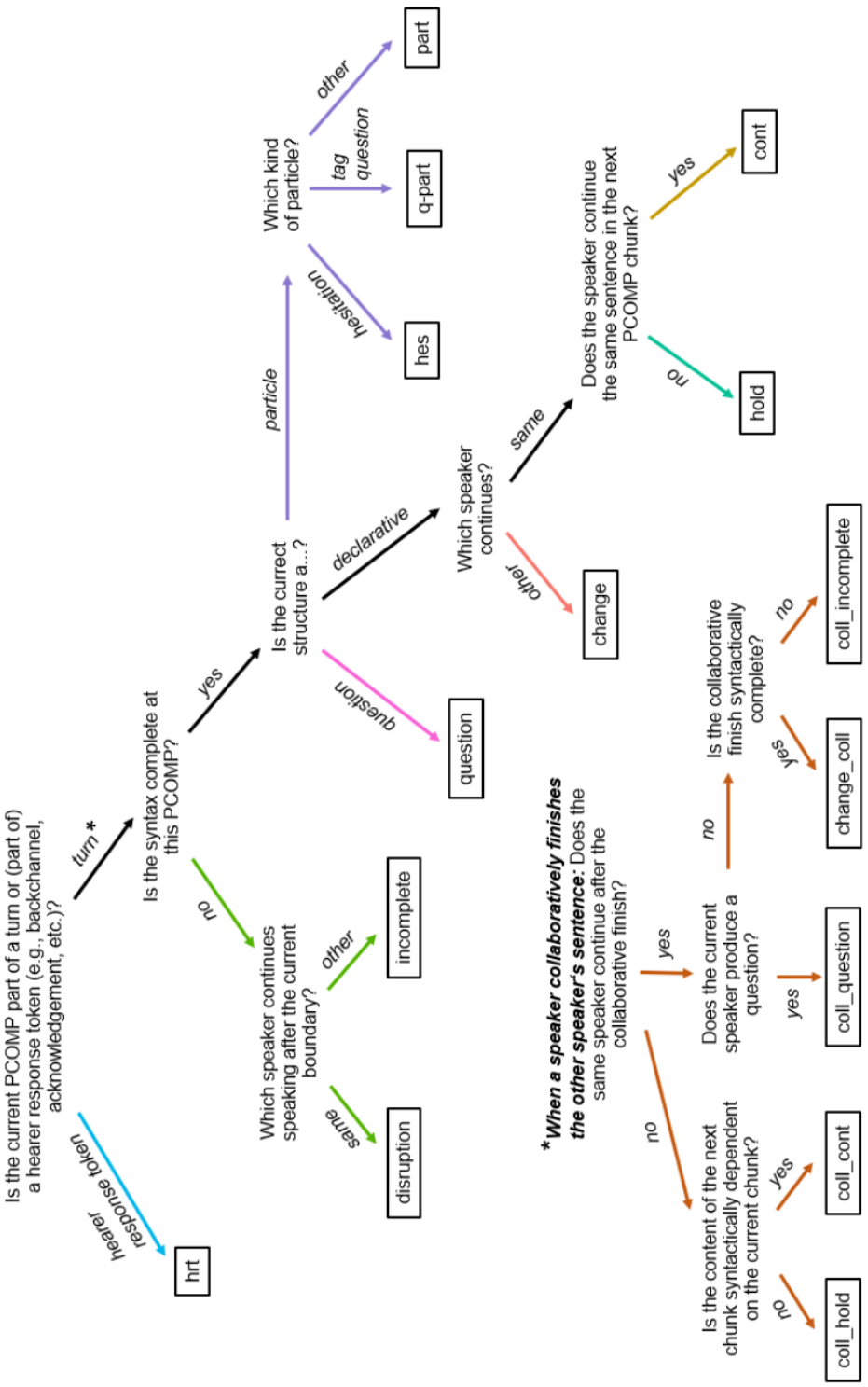


Figure A.1: Decicion tree for assigning PCOMP labels. Taken from Kelterer and Schuppler (subm.).

Bibliography

Ali, M. (2020, April). *Pycaret: An open source, low-code machine learning library in python* [PyCaret version 1.0.0]. https://www.pycaret.org

Boersma, P., & Weenink, D. (2001). Praat, a system for doing phonetics by computer. *Glot international*, *5*, 341–345.

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5–32. https://doi.org/10.1023/A:1010933404324

Buschmeier, H., & Włodarczak, M. (2013). TextGridTools: A TextGrid processing and analysis toolkit for Python. *Proceedings der 24. Konferenz zur elektronischen Sprachsignalverarbeitung*, 152–157.

Butzberger, J., Murveit, H., Shriberg, E., & Price, P. (1992). Spontaneous speech effects in large vocabulary speech recognition applications. in darpa speech and natural language workshop, pages 339-343, 339–343. https://doi.org/10.3115/1075527.1075607

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. https://doi.org/10.1145/2939672.2939785

Chicco, D., & Jurman, G. (2023). The matthews correlation coefficient (mcc) should replace the roc auc as the standard metric for assessing binary classification. *BioData Mining*, *16*(1), 4. https://doi.org/10.1186/s13040-023-00322-4

Ćwiek, A., Neueder, S., & Wagner, P. (2016). Investigating the communicative function of breathing and non-breathing "silent" pauses. In F. Draxler Christoph; Kleber (Ed.), *Tagungsband der 12. Tagung Phonetik und Phonologie im deutschsprachigen Raum* (pp. 27–29).

Ćwiek, A., Włodarczak, M., Heldner, M., & Wagner, P. (2017). Acoustics and discourse function of two types of breathing signals. *Nordic Prosody: Proceedings of the XIIth Conference, Trondheim 2016*, 83–91.

Dijk, O., oegesam, Bell, R., Lily, Simon-Free, Serna, B., rajgupt, yanhong-zhao-ef, Gädke, A., Todor, A., Evgeniy, Hugo, Haizad, M., Okumus, T., & woochan-jang. (2023, February). *oegedijk/explainerdashboard: explainerdashboard 0.4.2: dtreeviz v2 compatiblity* (Version v0.4.2). Zenodo. https://doi.org/10.5281/zenodo.7633294

Eyben, F., Wöllmer, M., & Schuller, B. (2010). Opensmile: The munich versatile and fast open-source audio feature extractor. *Proceedings of the 18th ACM International Conference on Multimedia*, 1459–1462. https://doi.org/10.1145/1873951.1874246

Fukuda, T., Ichikawa, O., & Nishimura, M. (2018). Detecting breathing sounds in realistic japanese telephone conversations and its application to automatic speech recognition. *Speech Communication*, *98*. https://doi.org/10.1016/j.specom.2018.01.008

Hara, K., Inoue, K., Takanashi, K., & Kawahara, T. (2018). Prediction of turn-taking using multitask learning with prediction of backchannels and fillers. *Listener*, *162*, 364.

Heldner, M., & Włodarczak, M. (2016). Is breathing silence? *Proceedings of Fonetik 2016 :* (57 (1)), 35–38.

Heldner, M., Włodarczak, M., Branderud, P., & Stark, J. (2019). The resptrack system. *1st International Seminar on the Foundations of Speech: BREATHING, PAUSING, AND THE VOICE, Sønderborg, Denmark, 1-3 December, 2019*, 16–18.

Ishii, R., Otsuka, K., Kumano, S., Matsuda, M., & Yamato, J. (2013). Predicting next speaker and timing from gaze transition patterns in multi-party meetings. *Proceedings of the 15th ACM on International Conference on Multimodal Interaction*. https://doi.org/10.1145/2522848.2522856

Ishii, R., Otsuka, K., Kumano, S., & Yamato, J. (2014). Analysis of respiration for prediction of "who will be next speaker and when?" in multi-party meetings. *Proceedings of the 16th International Conference on Multimodal Interaction*, 18–25. https://doi.org/10.1145/2663204.2663271

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 30). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf

Kelterer, A., & Schuppler, B. (subm.). Turn-taking annotation for quantitative and qualitative analyses of conversation [Submitted to Dialoge & Discourse in Jan. 2024].

Lemaître, G., Nogueira, F., & Aridas, C. K. (2017). Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, *18*(17), 1–5. http://jmlr.org/papers/v18/16-365

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable ai for trees. *Nature Machine Intelligence*, *2*(1), 2522–5839.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 30). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

McFee, B., McVicar, M., Faronbi, D., Roman, I., Gover, M., Balke, S., Seyfarth, S., Malek, A., Raffel, C., Lostanlen, V., van Niekirk, B., Lee, D., Cwitkowitz, F., Zalkow, F., Nieto, O., Ellis, D., Mason, J., Lee, K., Steers, B., … Pimenta, W. (2023, August). *Librosa/librosa: 0.10.1* (Version 0.10.1). Zenodo. https://doi.org/10.5281/zenodo.8252662

McKinney, W. (2010). Data Structures for Statistical Computing in Python. In S. van der Walt & J. Millman (Eds.), *Proceedings of the 9th Python in Science Conference* (pp. 56–61). https://doi.org/10.25080/Majora-92bf1922-00a

Navarretta, C. (2020). Speech pauses and dialogue acts. *2020 IEEE International Conference on Human-Machine Systems (ICHMS)*, 1–6. https://doi.org/10.1109/ICHMS49158.2020.9209502

Ogden, R. (2024, September). Listening to talk-in-interaction: Ways of observing speech [© 2024 Cambridge University Press. This is an author-produced version of the published paper. Uploaded in accordance with the publisher's self-archiving policy. Further copying may not be permitted; contact the publisher for details.]. In J. Robertson, R. Clift, K.

Kendrick, & C. Raymond (Eds.), *The cambridge handbook of methods in conversation analysis*. Cambridge University Press.

Pohjalainen, J., Räsänen, O., & Kadioglu, S. (2015). Feature selection methods and their combinations in high-dimensional classification of speaker likability, intelligibility and personality traits. *Computer Speech Language*, *29*(1), 145–171. https://doi.org/https://doi.org/10.1016/j.csl.2013.11.004

Rochet-Capellan, A., & Fuchs, S. (2014). Take a breath and take the turn: How breathing meets turns in spontaneous dialogue. *Philosophical Transactions of the Royal Society B*, *20130399*. https://doi.org/10.1098/rstb.2013.0399

Sainburg, T. (2021). *Noisereduce* (Version 2.0.0). https://github.com/timsainb/noisereduce

Schegloff, E. A. (1996). Turn organization: One intersection of grammar and interaction. In E. Ochs, E. A. Schegloff, & S. A. Thompson (Eds.), *Interaction and grammar* (pp. 52–133). Cambridge University Press.

Schuller, B., Steidl, S., Batliner, A., Hirschberg, J., Burgoon, J. K., Baird, A., Elkins, A., Zhang, Y., Coutinho, E., & Evanini, K. (2016). The interspeech 2016 computational paralinguistics challenge: Deception, sincerity native language. *Proc. Interspeech 2016*, 2001–2005. https://doi.org/10.21437/Interspeech.2016-129

Schuppler, B., Hagmüller, M., & Zahrer, A. (2017). A corpus of read and conversational austrian german. *Speech Communication*, *94*. https://doi.org/10.1016/j.specom.2017.09.003

Schuppler, B., & Kelterer, A. (2021). Developing an annotation system for communicative functions for a cross-layer ASR system. *Proceedings of the First Workshop on Integrating Perspectives on Discourse Annotation*, 14–18. https://aclanthology.org/2021.discann-1.3

Torreira, F., Bögels, S., & Levinson, S. (2015). Breathing for answering: The time course of response planning in conversation. *Frontiers in Psychology*, *1*. https://doi.org/10.3389/fpsyg.2015.00284

Trouvain, J., Werner, R., & Möbius, B. (2020). An Acoustic Analysis of Inbreath Noises in Read and Spontaneous Speech. *Proc. 10th International Conference on Speech Prosody 2020*, 789–793. https://doi.org/10.21437/SpeechProsody.2020-161

Werner, R., Fuchs, S., Trouvain, J., Kürbis, S., Möbius, B., & Birkholz, P. (2023). Acoustics of breath noises in human speech: Descriptive and three-dimensional modeling approaches. *Journal of Speech Language and Hearing Research*, 1–15. https://doi.org/10.1044/2023_JSLHR-23-00112

Whalen, D., Hoequist, C., & Sheffert, S. (1995). The effects of breath sounds on the perception of synthetic speech. *The Journal of the Acoustical Society of America*, *97*, 3147–53. https://doi.org/10.1121/1.411875

Winter, B., & Grawunder, S. (2012). The phonetic profile of korean formal and informal speech registers. *Journal of Phonetics*, *40*, 808–815. https://doi.org/10.1016/j.wocn.2012.08.006

Włodarczak, M., & Heldner, M. (2020). Breathing in conversation. *Frontiers in Psychology*, *11*, 2574. https://doi.org/10.3389/fpsyg.2020.575566

Yeo, I.-K., & Johnson, R. A. (2000). A new family of power transformations to improve normality or symmetry. *Biometrika*, *87*(4), 954–959. https://doi.org/10.1093/biomet/87.4.954

List of Figures

List of Tables