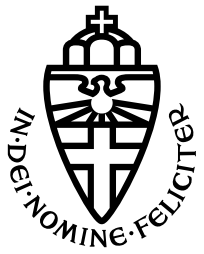
Radboud University Nijmegen



Can I get uhh a better WER

challenges and opportunities in evaluating conversational speech recognition

MA Linguistics and Communication Sciences (Research)

2023 — Nijmegen, The Netherlands

Contents

1 Introduction

In recent years, there has been a surge in interest in conversational speech recognition, driven by its potential to facilitate natural and intuitive interactions between machines and its users. This growing fascination is well-founded, as conversational speech recognition holds great promise across multiple domains. In the medical field, for instance, conversational speech recognition offers a hands-free and more efficient way of documenting clinical information. These systems can also play a vital role in language translation and transcription, encouraging cross-cultural communication and foreign language learning. Furthermore, individuals with speech and hearing difficulties can benefit significantly from adaptive conversational speech technology, which can provide real-time transcriptions – allowing them to follow and even participate in social interactions better. The technology's versatility extends beyond specific applications as well, as it is increasingly incorporated into everyday tools like virtual assistants and voice-controlled devices. As these become more and more ubiquitous, the gap between true conversational ability and the extent to which conversational technology can imitate natural interactions becomes more apparent.

Current conversational technology may be proficient in handling straightforward tasks and providing scripted responses, but they often struggle to navigate the intricacies of real-life, open-ended interactions. Shriberg (2005) describes some of the fundamental challenges of automatic speech recognition (ASR) with conversational speech such as representing overlaps, short conversational words, and rapid turn-taking. Despite the significant technological progress made over the past decades, these challenges remain highly relevant. This is because contemporary ASR systems are typically trained on read, monologic speech (Liesenfeld and Dingemanse, 2022). Although such data is valuable for developing speech recognition systems with high accuracy in controlled environments, spontaneous conversations offer a far more complex and richer context. Conversations are dynamic, with interlocutors engaging in a continuous process of contribution and negotiation to create meaning collaboratively. Understanding and accurately transcribing conversational speech requires ASR systems to be sensitive to these aspects. It goes beyond simply recognizing individual words or phrases and involves capturing the structure and nuances that emerge from interaction. Failing to do so results in frustrating and unnatural exchanges, unmet expectations, and falling short on intended use cases – as evident in various user experience studies on current state-of-the-art technology (Luger and Sellen, 2016; Hoegen et al., 2019; Alač et al., 2020).

Over the years, various attempts have been made to identify the specific elements of interaction that ASR systems tend to overlook. Consistent with the findings of Shriberg (2005), several studies have highlighted the dearth of conversational words in ASR transcripts. These conversational words are crucial for enabling smooth turn-taking and interaction among interlocutors, with recent research underscoring their significance in modern systems (Lopez et al., 2022; Mansfield et al., 2021; Zayats et al., 2019). To address the challenge of conversational speech recognition, it is essential to delve deeper into the intricacies of conversations. This entails moving beyond merely focusing on words and phrases and considering aspects such as timing, structure, and other interactional features.

One critical step in this direction is to tackle the shortcomings of current evaluation methods utilized to assess the performance of conversational speech recognition systems. The Word Error Rate (WER), being the prevailing evaluation metric in the field of ASR, has indeed demonstrated its utility in certain contexts. This simple and easily calculated standard provides a convenient choice for measuring word-level accuracy in controlled speech settings. However, the difficulties arise when we apply WER to conversational speech, which involve interactional infrastructure and resources that play a pivotal role in shaping the conversation's meaning and flow. WER falls short in capturing these crucial features, often leading to an underestimation of the system's true performance in real-world interactive scenarios (Szymański et al., 2020; Aksënova et al., 2021). Neglecting to account for these in the evaluation process hinders the progress of conversational speech recognition systems and limit their effectiveness in practical applications. A more comprehensive evaluation framework is urgently needed, and one that gives due consideration to the intricacies of conversational speech.

In this thesis, I explore in detail the disparities between human and ASR transcriptions of conversations, emphasizing the importance of reevaluating how we assess the performance of conversational speech recognition systems. By examining the unique challenges posed by these interactions, I advocate for a paradigm shift in the evaluation of ASR systems, moving beyond the traditional reliance on Word Error Rate (WER) as the sole metric. I start by briefly discussing conversational speech, describe the

infrastructure that both facilitates and arises from interaction as well as the different resources employed by interlocutors to achieve successful communication. I then provide an overview of the current landscape of conversational speech recognition and its evaluation methods. To further investigate these conversational systems, I run several off-the-shelf ASR engines on spontaneous conversations across a variety of languages. With the results of my analyses, I propose a composite measure that considers multiple aspects of interaction. Ultimately, my research aims to contribute to the advancement of conversational AI technologies, making them more adaptive, responsive, and attuned to the complexities of how humans naturally communicate.

2 Literature Review

Conversational speech interfaces have steadily grown in popularity and are being used more and more in real-world interactive settings. Given the rising demand, speech recognition technology – a critical component of these interfaces – is now faced with the challenge of handling natural and spontaneous conversational speech. Conversational speech, in contrast to other forms of speech, reflects how humans naturally communicate and interact with one another. This makes it more complex than the typically pristine, monologic, and carefully read speech these technologies are usually trained on (Panayotov et al., 2015). In conversations, meaning is established through an ongoing and dynamic collaboration among participants (Levinson, 1983; Clancy et al., 1996), making use of different interactional resources. In this chapter, I describe the necessary elements to take into account for technology to be truly conversational, the current state of conversational speech recognition, and how these systems are evaluated.

2.1 Representing Conversations

The key challenge to conversational speech recognition is adequately representing conversations. In everyday life, people use language in conversations for a myriad of purposes such as to convey information, build relationships, negotiate, and express oneself (Levinson and Torreira, 2015). Naturally, capturing all of its intricacies – especially with the possible variations that different cultures, languages, and communicative situations can bring – is a seemingly impossible task. So instead, I focus on the "bare bones" of conversation in this section. Taking insights from conversation analysis, I describe the fundamental aspects of conversational speech and point out why each one is critical in understanding meaning.

2.1.1 Interactional Infrastructure

To truly understand the underlying mechanisms behind conversations, we must first establish what they are. When we think of the word "communication", we often envision conversations. This is where we use language at its most natural form, continuously working together to establish meaning. The conversational system both supports and reflects this ongoing interaction, accommodating an indeterminate number of interlocutors. Conversations are made up of turns, with participants switching intuitively between comprehension and production and doing this so quickly that the two processes often occur simultaneously. These turns are made up of syntactic units but are not necessarily syntactically complete. And while turn length can indeed vary, they are often short – averaging at around two seconds long. Furthermore, the gap between turns were found to usually be in between 0 and 200 ms (Dingemanse and Liesenfeld, 2022).

With the exchange occurring so rapidly to maximize efficient communication (Levinson and Torreira, 2015), it is also quite common for the turns to overlap. These portions of simultaneous speech are not mere interruptions or disorder in an otherwise tidy structure (Drew, 2009). Instead, overlap is an incredibly common phenomenon that plays a fundamental role in the organization and dynamics of a conversation – allowing for fluid turn-taking (Dingemanse et al., 2015). Overlaps can act as a cooperative device, signaling agreement or completing an anticipated point (Li, 2001; Stivers, 2008). In contrast, they can also be competitive, where a speaker attempts to disrupt ongoing speech in order to express a strong opinion, topic change, or disagreement (Li, 2001; Yang, 2001). Overall, overlaps are used as a resource to navigate and facilitate the turn-taking system as well as to establish meaning.

Pauses within a conversation are also used strategically to convey meaning (Sacks et al., 1974; Heldner and Edlund, 2010; Beach, 1991). They can indicate a temporary hesitation or uncertainty, a shift to a

different topic, or a point for emphasis (Schegloff, 1986). Pauses can also show the speaker's monitoring of the ongoing exchange (Schegloff, 2000). hand, are the silences between turns (Heldner and Edlund, 2010). Furthermore, the length and placement of pauses have different implications in a conversational context. Short pauses may indicate strong agreement, while longer ones may hint at uncertainty or hesitation (Beach, 1991). These temporal features of conversations add a dimension to meaning that cannot be put aside if one wants to fully capture its contents.

2.1.2 Interaction Resources

Conversational speech stands out with a distinct set of lexical features, as it comprises of certain words and phrases that reflect its interactive nature. This includes contracted forms (i.e. *"gonna"* for going to), widely-used expressions (i.e. *"what's up?"*), and idioms – all elements that arise and contribute to natural and successful language use. The familiarity of this vocabulary allows interlocutors to quickly grasp each other's intended meanings, establish rapport, and maintain social appropriateness in a conversation (Bell, 1984). By using these as interaction resources, participants can better navigate the dynamic conversational landscape allowing for a fluid and efficient exchange (Sacks, 1974). Beyond their instrumental role in conversations, these words and phrases also offer subtle insights into a speaker's individual characteristics such as personality and heritage (Eckert, 2000; Buchstaller, 2006). As such, these interaction resources vary across languages and communities but are generally utilized in the same manner.

One cannot discuss interaction resources without giving due attention to interjections. Interjections are typically short utterances that enable smooth and systematic turn-taking – fundamental in establishing and maintaining the whole conversational system (Sacks et al., 1974). These specific resources can be employed to signal the speaker's presence and engagement (*"mhm"*), or to add a layer of expressiveness by indicating surprise (*"wow"*), hesitation (*"uhh..."*), and frustration (*"ugh"*) (Ameka, 1992; Dingemanse, 2021). They are also used to flag communication breakdowns (*"huh?"*) (Dingemanse et al., 2015), allowing for prompt resolution and continuity in the flow of a conversation. Furthermore, interjections convey a speaker's intention to initiate or maintain a turn. Interlocutors therefore use these utterances to coordinate the timing and sequencing of their conversational turns. As an essential feature of conversational speech, interjections exemplify the underlying mechanisms at play in interaction.

Conversations are also distinguished by discourse markers like *"I mean"* and *"you know"*. These words and phrases are instrumental in conversational speech by revealing the speaker's stance, marking turn boundaries, and highlighting important information (Fraser, 1990). They also serve as connectors, indicating relationships like cause-effect (*"[be]cause"*) and exemplification (*"like"*). These ensure that the sequential expression of ideas and information are logical and smooth (Lenk, 1998). Without discourse markers, conversations become incoherent and disorderly.

Conversational speech is not solely reliant on utterances or certain words and phrases. Interlocutors also make use of prosody, encompassing of various aspects such as rate of speech and intonation, as another resource to convey meaning. These features play a crucial role in placing emphasis and punctuation, as well as signaling an individual's intentions and attitudes. Speakers vary their speech rate to highlight relevant portions of their turn, and often produce repair sequences at a relatively faster rate (Auer and Luzio, 1992). Participants in a conversation also use intonation patterns as hidden punctuation, for example to differentiate a statement from a question, and to convey undertones like sarcasm and annoyance (coh, 2003). Research has also shown that speakers often raise their pitch and increase the loudness to signal a topic change (Schegloff, 1979). The lowering and rising is also used as a strategy to indicate continuity between ideas and to mark turn and segment boundaries (Hirschberg and Pierrehumbert, 1986; Lelandais and Thiberge, 2023).

While this subsection provides a broad overview of the interaction resources utilized in conversations, it is not exhaustive and there may be additional elements and strategies that are not covered here. Moreover, while I give examples in English, it is crucial to acknowledge that these interaction resources exist in the vast majority, if not all, languages used worldwide. Understanding the universality of these resources is vital, as it highlights the fundamental principles of interaction shared across diverse linguistic communities. At the same time, it is equally important to recognize and represent the variation that undoubtedly exists.

2.2 Conversational Speech Recognition

Automatic Speech Recognition (ASR) is a crucial component in conversational technology as it enables the analysis and processing of human speech. However, these ASR systems face unique challenges in conversational settings. This section delves into the uses of conversational speech recognition and examines the specific obstacles that arise.

2.2.1 Uses

Conversational speech recognition has found a series of applications across a range of domains. In the field of Human-Computer Interaction, the recognition of conversational speech is crucial in enabling natural and intuitive interactions with virtual assistants and voice-controlled systems. It also holds great potential in language learning and education by providing automatic feedback and assessment of learners' spoken language proficiency (Neri et al., 2003; Wald, 2005), as well as in healthcare by assisting in medical transcription and communication accessibility (Berez-Kroeker et al., 2022; Healy et al., 2013). The use of automatic transcripts of conver 'sations are also being explored in several other fields such as: marketing (Vajpai and Bora, 2016; Sehgal et al., 2018), journalism (Munteanu et al., 2006; de Lima-Santos and Ceron, 2022), and legal documentation (Lööf et al., 2010). In research, conversational speech recognition can also facilitate large-scale analysis of spoken data, providing unique insights on language and cognition ().

Currently, there are a handful of available ASR systems that claim to be conversational. The state-of-the art engines are developed by big tech companies like Meta, Google, and Amazon; however, there is limited publicly available information on their official performance on conversational speech and on how exactly they process the data. However, there are a few systems that do report on this. For example, Han et al. (2018) developed the CAPIO conversational speech recognition system, reporting word error rates of 5.0% on the Switchboard corpus and 9.1% on CallHome. Microsoft also published documentation on their system a several years ago, with word error rates of 5.1% and 9.8% on the same two corpora respectively (Xiong et al., 2018). Switchboard and CallHome are both collections of spontaneous conversations, while word error rate (WER) is a general percentage of words recognized correctly – all of which will be explained in greater detail in the following sections. Overall, these systems highlight the advancements in the field and work impressively according to the popular benchmarks they are measured against.

2.2.2 Challenges

While state-of-the-art conversational systems report incredibly low error percentages, with some even claiming to be better than human performance (Xiong et al., 2017), these numbers unfortunately do not necessarily translate to real-world applications. Studies have reported that modern conversational agents are far from being truly conversational, failing user expectations majority of the time (Luger and Sellen, 2016; Hoegen et al., 2019). One likely reason for this is that most ASR systems are trained on read, monologic speech (Panayotov et al., 2015) – which, as discussed in prior sections, is quite different from conversational speech. This results into ASR systems missing various aspects crucial to conversations.

For one, performance of ASR systems were found to be affected by overlapping speech – a natural and frequent occurrence in conversations. Çetin and Shriberg (2006b) has shown that ASRs exhibit a higher word error rate in portions with overlapping speech, as well as on portions surrounding the overlap. This tendency occurs even on recordings with speakers closer to the microphone, showing that these errors are likely due to the recognition of background speech; thus introducing confusion in the transcript (Shriberg et al., 2001). Further analysis showed that these portions of overlap, where ASRs are likely to misrecognize words, occur at the following places: potential turn exchanges, relatively more complex utterances, and when speakers are more affectively involved (Çetin and Shriberg, 2006a). These areas are all especially crucial in the context of the data used – meetings – and to all other forms of conversations in general. If machines are likely to make errors during these peaks of interaction, then the whole interaction is also likely to be misinterpreted. This was shown in a user study conducted by Hoegen et al. (2019), wherein they report that when a conversational agent is not equipped to handle overlaps, it results to '*"nonsensical responses"*.

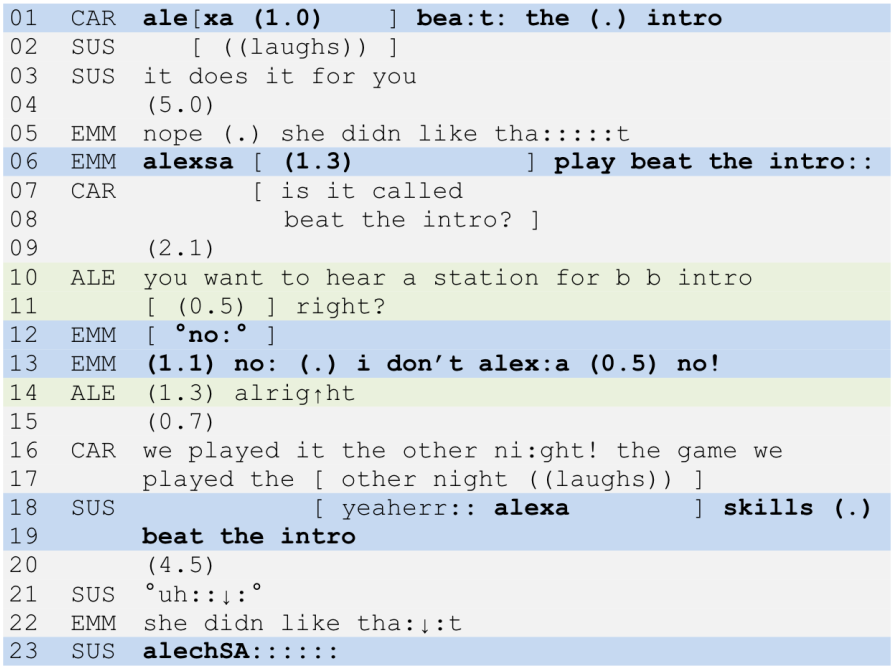


Figure 1: Example illustrating a discrepancy between Echo's error indication and users' response, taken from Porcheron et al. (2018). The users attempted to repair the communication breakdown by reiterating the request, showing that how Echo signals misunderstanding is ineffective.

Another challenge conversational systems face is representing interjections and other conversational words. In the field of speech technology, these are also referred to as "backchannels", "disfluencies", or "fillers" – reflecting how they are often regarded as unimportant or something to be filtered out. This reflects too in the transcripts ASR systems produce, as numerous error analyses have pointed out that conversational words are notably absent in their output (Zayats et al., 2019; Mansfield et al., 2021; Lopez et al., 2022; Shriberg et al., 2001; Goldwater et al., 2008; Xiong et al., 2017). However, misrepresenting these can lead to confusions to whether a user is asking for clarification, holding their turn, transitioning to a different topic, or waiting for acknowledgement. This is exactly what happened in a study about Amazon Echo, where one of the findings is that *"there is a significant mismatch sometimes between the ways in which designed responses from the Echo appear to integrate indicators of the form of trouble, and actually how participants dealt with them"* (Porcheron et al., 2018). This is exemplified in figure 1 where Echo signaled an error in categorizing the user's request. However, the users' response to this miscommunication was to repeat the request, as if the error lied in Echo's failure to accurately hear them the first time.

With ASR systems being mainly trained on written language, prosody is also often overlooked in conversational speech recognition. This results in challenges in accurately detecting turn boundaries and "hidden" punctuation (Shriberg, 2005). Recognizing the significance of prosodic features, Hirschberg et al. (2004) demonstrated that incorporating these features not only helps identify misrecognized turns but also improves the acceptance of correctly recognized ones, leading to smoother and more efficient interactions. They argue that strategically utilizing prosodic features in combination with existing ASR systems significantly reduces the word error rate. This finding aligns with the observations made by Furui et al. (2005), who discovered that words produced in conversational speech exhibit less phonemic distance compared to those in read speech. This means that in conversation, sounds within words are more blended; whereas in read speech, each sound is articulated more distinctly. This phenomenon may explain why users often feel compelled to exaggerate their pronunciation when interacting with conversational agents as a repair mechanism (Pelikan and Broth, 2016; Luger and Sellen, 2016), a behavior less commonly observed in human-human conversations. Aside from identifying turn boundaries and recognizing words, prosodic features are also essential in automatically classifying speech acts and prag-

matic features (Shriberg and Stolcke, 2004). This is further evidenced in a review of state-of-the-art conversational agents by Luger and Sellen (2016), whhere a user described Siri's response to their sarcastic remark as, *"and it just said, I swear in an equally sarcastic tone, 'that's fine, it's my pleasure"'*, negatively affecting user experience.

In summary, conversational speech recognition poses significant challenges that must be addressed to develop robust systems. Two aspects that are often overlooked are the interactive nature of conversations and that meaning is established collaboratively. Quick and overlapping turns, interjections, and conversational words all play crucial roles in accurately representing naturalistic interactions with users. Furthermore, additional dimensions of meaning that are unique to conversations and not commonly found in written text or monologues, should not be disregarded if a system aims to be responsive and efficient. This includes capturing prosodic features such as pauses and intonation. An effective conversational system should adhere to the natural structure of human communication rather than imposing a "tidied" organization for the sake of computational simplicity. Failure to understand and incorporate the intricacies of interaction can hinder the development of truly effective conversational systems.

2.3 Evaluation of ASR for Conversational Speech

In the current technological landscape, there definitely is a demand for conversational agents to be fit for real-world use. But where do we start? I discuss in this section how speech recognition systems are currently evaluated, what existing benchmarks that a system needs to meet to claim human conversational ability, and why there's a need for better methods.

2.3.1 Benchmarks

Benchmarking plays a crucial role in the development of ASR systems for conversational speech. It involves a systematic evaluation with an established standard, aiming to capture both the performance of a system as well as how it compares to others. In ASR, the evaluation method most often used is the Word Error Rate (WER). WER is a simple metric that compares an ASR transcript against a reference, typically human-annotated data, using the following formula:

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

where S is for substitutions, D for deletions, I for insertions, and N is the total number of words in the reference. In the field of automatic speech recognition, the widely referenced word error rate of human parity is set at around 5%. This standard supposedly goes back to a 1996 study entitled "Personal Communication" where the human word error rate for spontaneous speech is calculated at 4%. However, how this was calculated remains unknown as the reference is only cited in another study by Lippmann (1997) where the differences between human and machine-produced transcripts are revisited and further analyzed. In that review, Lippmann concluded that human error rates are at below 5% in spontaneous speech with channel variability and noise. Further work was done by Xiong et al. (2017) where nuances between types of conversational speech were taken into account by using two different corpora – Switchboard (Godfrey et al., 1992), a collection of conversations between strangers, and CallHome (Canavan et al., 1997), a collection of conversations between friends and family.

At this point, two standards emerge – the use of these two corpora and the reference human rates of 5.9% for Switchboard and 11.3% for CallHome. Saon et al. (2017), on the other hand, argues that the human rates are instead at 5.1% and 6.8% respectively, making "human parity" harder to achieve. The differences between the two calculations are brought about the instructions given to the transcribers, how meticulous they were in transcribing, and how the transcripts were processed.

Moreover, there are also other datasets used to benchmark conversational speech recognition. A notable example is CHiME5, a collection of distant-microphone dinner party conversations, where current systems report WERs ranging from 45% - 73% (Manohar et al., 2019; Szymański et al., 2020). Unlike the dyadic telephone conversations found Switchboard and CallHome, CHiME includes several aspects that are more representative of real-world applications: presence of background noise, unprompted conversations tackling a variety of topics, and multi-party interactions. However, even with such a challenging task, there are still some limitations with CHiME and most benchmark datasets currently employed. One, these datasets only include native speakers of the particular language, where non-native accents

may not be taken into account in the evaluation of ASR systems (Canavan et al., 1997; Godfrey et al., 1992; Manohar et al., 2019; Kawahara et al., 2003). Another limitation is that there is a lack of spontaneous conversational speech data for diverse languages that represent the global population (Liesenfeld and Dingemanse, 2022). With so much to take into account, there is still no standard human word error rate or other set of evaluation methods established as of this writing.

2.3.2 Evaluation Methods

There are also concerns about how well the human word error rate truly represents how humans recognize speech. Word error rate (WER) compares the ASR transcript to a references and categorizes errors into three: substitutions, deletions, and insertions. The rate is then the percentage of the total of number of errors against the total number of words in the reference. A closer look at the human word error rate and ASR systems that have reached this unveils a couple of things. First, Mansfield et al. (2021) shows that there are significant and systematic differences between the types of errors a human and a machine makes in transcribing Switchboard and CallHome. One, they found that humans are more likely to commit deletion, while machines tend to insert and substitute words. Next, they also found that conversational words caused the largest discrepancy between human and ASR WERs. The same finding was also reported in the development of the first conversational recognition system that claimed to reach human parity (Xiong et al., 2017), wherein errors occurred with words characteristic of spontaneous speech (i.e. interjections, repetitions, and corrections). These findings raise the question, if ASR systems that supposedly approach or even reach human parity make errors on the very words that characterize spontaneous conversations, how conversational can they truly be? Especially with the user experience studies discussed previously, there is an apparent and undeniable disparity between how these technologies are benchmarked and how they actually perform in conversational settings.

Starting with the widely used WER, the preference for this metric stems from its simplicity. It provides an easy common ground for comparing different ASR systems as well as their performances on different datasets. Furthermore, it measures accuracy well, serving as an effective evaluation for certain applications such as text dictation. However, its deficits – particularly in conversational speech recognition – are well documented (Szymański et al., 2020; Aksënova et al., 2021; Errattahi et al., 2018). As these studies have found, the different types of words are not recognized equally. ASR systems have been found to systematically miss conversational words even across different languages (Lopez et al., 2022; Stolcke and Droppo, 2017). This error is unique to machines, as humans do not confuse these elements, but rather rely on them to structure a conversation (Stolcke and Droppo, 2017). In addition to that, homophones were also found to be difficult for machines to correctly transcribe (Pasandi and Pasandi, 2023; Wirth and Peinl, 2022), highlighting the need for context – referring nearby words within the utterance – to be considered in evaluation. Additionally, since WER treats all errors equally, it fails to serve as a true percentage in a sense that values above 100 are possible especially in noisy environments. Finally, as what many error analyses have pointed out, a single value fails to pinpoint any strengths or weaknesses of a system. In this regard, it merely serves as a score where developers aim to one-up each other – moving focus away from developing systems catered to representing conversations and are instead catered to producing a low word error rate.

There have been numerous efforts to address these deficits, with several alternatives proposed for conversational speech recognition specifically. Shah et al. (2022) pointed out the limitations of WER for languages that have multiple accepted spellings for each word, and proposed Alternate Word Error Rate (AWER). Unlike WER, AWER takes into consideration ambiguous characters commonly found in certain languages like Hindi. Several approaches have also been proposed to resolve this particular limitation as well. A widely-used alternative to WER is Character Error Rate (CER), where accuracy is measured at the character level instead of word level. This comes in particularly handy when transcribing speech with non-words or on logographic languages. Another approach is using the Phonetically Oriented WER, proposed by (Ruiz and Federico, 2015), to minimize homophonic errors and resolve ambiguously recognized portions. Another key limitation of WER is that it appears to be disconnected to how humans evaluate and perceive speech, make it an ineffective measure for real-life performance. Morris et al. (2004) propose a solution: the Match Error Rate (MER) where the types of errors are optimized and a the rate serves as a true percentage of a system's overall performance.

In conversational speech recognition, it is essential not only to focus on accurately recognizing the

speech but also to allocate it to the appropriate speaker and represent conversational turns. To evaluate this aspect, a commonly used method is known as Diarization Error Rate (DER) (Park et al., 2022). Unlike metrics such as WER and its alternatives discussed earlier, DER specifically addresses the accuracy of speaker diarization, ensuring that each speaker segments are correctly identified and attributed in the conversation. DER is defined as:

$$DER = \frac{FA + MISS + ERROR}{TOTAL}$$

where *FA* is the total hypothesis speaker time not attributed to a reference speaker, *MISS* is the total reference speaker time not attributed to a hypothesis speaker, *ERROR* is the total reference speaker time attributed to the wrong speaker, and *TOTAL* is the total reference speaker time, i.e. the sum of the duration of all reference speaker segments. However, DER also has its own limitations. One particularly relevant to conversational speech is that DER falls short in capturing the significance of short conversational phrases. These utterances, despite their brevity, often play a significant role in structuring a conversation as previously discussed. To address this, (Cheng et al., 2022) proposed the Conversational Short-phrase Diarization Error Rated (CDER), defined as:

$$CDER = \frac{The\ number\ of\ mistakes}{The\ number\ of\ total\ utterances}$$

The CDER is calculated by first merging utterances from the same speaker, resulting into an A-B-A-B format in dyadic conversations. The reference is then compared against the ASR transcript using the formula above computed at the utterance level for each speaker.

As of the current state, the CDER stands as the sole published evaluation metric specifically designed for conversational speech recognition. It addresses certain interactional resources utilized in spontaneous conversations, rightly acknowledging their significance. However, CDER still does not fully capture the interactional infrastructure and temporal dynamics present in conversational speech. These unrepresented aspects highlight the need for a specialized evaluation metric that can comprehensively account for the unique characteristics of conversational speech, including its interactive nature and the precise timing of speaker turns and overlaps. While the CDER is a step in the right direction, there is still room for further refinement and development to comprehensively represent interaction in the evaluation process. Such advancements would enhance our ability to assess and improve the performance of conversational speech recognition systems in real-world settings (Likhomanenko et al., 2021).

3 Methods

This section describes the methodology used to investigate the evaluation of conversational speech recognition. To gain a comprehensive understanding of the differences between how a human and a machine processes conversational speech, a mixed-methods approach combining qualitative and quantitative techniques was employed.

3.1 Data

3.1.1 Human Transcripts

Human transcripts were used as reference in thoroughly evaluating how ASR systems handle conversational speech. The selection of the data was guided by the aim of capturing spontaneous and natural conversations, ensuring that the full range of complexities in conversational speech was accounted for. To ensure the accuracy and reliability of this evaluation, several quality control measures were put in place. First, only transcripts sourced from published conversational corpora that have undergone peer-review were used – increasing the credibility and reproducibility of the evaluation. Next, only transcripts that represent conversations at the level of individual turns, with each turn fully annotated with precise timestamps down to the decisecond (0.1ms), were included. These transcripts also reflect non-speech behaviors that often occur in conversations such as laughter, audible breaths, and unintelligible noises. Such phenomena are tagged and enclosed in brackets. Finally, in order to include a wide range of linguistic features and to compare the performance of ASR systems across languages, transcripts from

multiple widely supported languages were selected. This allows for cross-linguistic comparisons, revealing both language-specific and language-general features that influence ASR performance in conversational settings. Table 1 shows the languages and corpora used in this study.

Language	Corpus	Description
Dutch	IFAA Dialog Video Corpus (van Son et al., 2008)	face-to-face conversations
	Corpus Gesproken Nederlands (CGN) (Taalunie, 2014)	face-to-face and telephone conversations
English	CallHome American English (Canavan et al., 1997)	telephone conversations
French	The Nijmegen Corpus of Casual French (Torreira et al., 2010)	face-to-face conversations
German	Forschungs- und Lehrkorpus Gesprochenes (FOLK) (Reineke and Schmidt, 2022)	face-to-face conversations
Korean	CallFriend Korean (Canavan and Zipperlen, 1996a)	telephone conversations
Mandarin	CallHome Mandarin Chinese (Canavan and Zipperlen, 1996b)	telephone conversations
Spanish	Glissando Corpus (Garrido et al., 2013)	face-to-face conversations

Table 1: Conversational data used in this thesis. Each corpora was carefully selected to represent spontaneous and natural interactions across a diverse set languages, fully annotated with decisecond-precise timestamps at the turn level.

3.1.2 ASR Transcripts

In this research, carefully selected conversational corpora that included sound files were utilized. These sound files were processed using to two types of ASR systems: off-the-shelf systems and systems designed to handle conversational speech to some extent. The inclusion of both types of ASR systems allowed for a comprehensive evaluation of their performance in transcribing conversational speech. The off-the-shelf systems represent the commonly available ASR technologies, while the conversational-specific systems represent technologies that have somewhat addressed the unique challenges posed by conversational speech. By analyzing the output of these different ASR systems against the human transcripts, we can further understand what aspects need to be better represented in both the ASR transcript and how it is evaluated.

The three off-the-shelf ASR systems – one each for Dutch, English, and German – were accessed through the Bavarian Archive for Speech Signals' CLARIN Transcription Portal Draxler et al. (2020)[1]. On the other hand, table 2 shows the commercial systems used, corresponding models employed, and how they claim they can handle conversational speech. Because processing conversational speech entails much more than simply recognizing what was said and in order to match the information found in the human transcripts, several additional features were also used – speaker diarization, which involves assigning a speaker to each speech segment or turn, word-level timestamps, and confidence scores that reflect the certainty of the machine on the transcript it produced. A summary of the features used with each ASR as well as any publicly available information its development can be found in Appendix A.

3.2 Pre-Processing

In the pre-processing stage of this research, several important steps were taken to ensure the comparability of human and ASR transcripts. First and foremost, all audio files were chosen to show similarity in terms of audio encoding and channel separation. This allows for consistency in the audio data used for transcription. To facilitate the analysis, non-speech tags present in the human transcripts were removed. These tags, indicating phenomena such as laughter or background noises, are not expected to show up

[1] `https://clarin.phonetik.uni-muenchen.de/apps/TranscriptionPortal/`

ASR System	Model	Claim
Amazon Transcribe	default	*"capture meetings and conversations that matter to you"*
Google Cloud Speech-To-Text	latest_long	*"for any kind of long form content such as media or spontaneous speech and conversations"*
	default	*"produces transcription results for any type of audio"*
NVIDIA NeMo ASR	QuartzNet15x15	*"conversational toolkit [...] paving a path to personalized, natural human-machine interactions"*
Rev AI Asynchronous Speech-To-Text	default	*"outperforms other speech-to-text providers in accuracy for virtually every use case", "analyze and monitor conversations with customers"*
Whisper	default	*"human-level robustness and accuracy on English speech recognition"*

Table 2: Conversational ASR systems used in this thesis. These stat-of-the-art commercial engines claim to handle conversational speech to some capacity across multiple languages.

in the ASR transcripts. In order to align the spelling conventions, proper names found in the ASR transcripts were corrected. For example, if the human transcript spelled a name as "Benni," but the ASR transcript had it as "Benny," the ASR transcript was adjusted accordingly to match the intended spelling. Similarly, contractions in the transcripts were unified into a single format for consistency. For instance, if the human transcript had "can't" but the ASR transcript had "ca nt," the ASR transcript was modified to match the standard contraction format of "can't." However, certain elements were preserved in this stage. Word fragments, indicated by hyphens (i.e., "h-") and shortened forms (i.e., "gonna"), were left untouched. These features are common in conversational speech and reflect the natural flow of dialogue. Both types of transcripts were then pre-processed using `cleantext`[2], a python package for systematically removing punctuation, capitalization, and other special characters. Finally, NLTK's `whitespace` was used for tokenization. This was selected based on its approach to tokenizing, which does not heavily rely on linguistic information. This characteristic ensures a consistent and uniform splitting of the transcript into basic units, regardless of the language being analyzed. These steps allow for a more accurate and reliable analysis of the ASR system's performance in transcribing conversational speech.

3.3 Error Analysis

The Word Error Rate (WER) was calculated to show how these systems would be typically evaluated. This was done using `jiwer`[3]. The WER was calculated for each conversational ASR system and for each supported language. This allows for a general but broad analysis of the performance of the ASR systems and provided insights into the challenges posed in evaluating conversational speech recognition in various linguistic contexts.

To gain a more comprehensive view of these challenges, the scaled F-score[4] was used for comparison. The Scaled F-score is a modified version of the traditional F-score designed to address certain limitations. Given a word $w_i \in W$ and a category $c_j \in C$, the precision of word w_i with respect to a category c_j is defined as the following:

$$\text{prec}(i,j) = \frac{\#(w_i, c_j)}{\sum_{c \in C} \#(w_i, c)}$$

The function $\#(w_i, c_j)$ represents either the number of times w_i occurs in an utterance labeled with the category c_j or the number of utterances labeled c_j which contain w_i. The frequency of a word within a category is defined as:

$$\text{freq}(i,j) = \frac{\#(w_i, c_j)}{\sum_{w \in W} \#(w, c_j)}$$

Then, the harmonic mean of these two values is defined as:

$$\mathcal{H}_\beta(i,j) = (1+\beta^2)\frac{\text{prec}(i,j) \cdot \text{freq}(i,j)}{\beta^2 \cdot \text{prec}(i,j) + \text{freq}(i,j)}$$

[2]https://pypi.org/project/cleantext/
[3]https://github.com/jitsi/jiwer
[4]https://github.com/JasonKessler/scattertext#understanding-scaled-f-score

$\beta \in \mathcal{R}^+$ is a scaling factor where frequency is favored if $\beta < 1$, precision if $\beta > 1$, and both are equally weighted if $\beta = 1$. F-score is equivalent to the harmonic mean where $\beta = 1$. To address specific issues, the Scaled F-score introduces two modifications. First, harmonic means tend to be dominated by precision, so the score is adjusted to provide a more balanced consideration of precision and frequency. Second, the score handles "low-frequency brittle terms" by accounting for tokens with extremely high or low frequencies. By making these modifications, the Scaled F-score aims to provide a more accurate assessment of the association between n-grams and a particular class. The score ranges from -1 to 1, with positive scores indicating an association between the n-gram and the class, and negative scores indicating no association.

4 Results

In this chapter, I present a comprehensive analysis of the dissimilarities between human-transcribed conversational speech and ASR-generated transcriptions. The investigation delves into the performance of two types of ASR systems: general-purpose and specialized models. Through this, I aim to contribute to a more nuanced understanding of the complexities involved in representing conversational speech. The results discussed in this chapter are also reported in two recently published papers – Lopez et al. (2022) and Liesenfeld et al. (2023).

4.1 General-Purpose ASRs on Conversational Speech

To see the Word Error Rates (WER) in action, general-purpose ASR systems were ran on Dutch, English, and German conversational speech. Figure 2A shows the word error rates of the three languages along with the "human parity" WER of 0.05 (dotted line) as reference. The results clearly show that, even among high-resource languages, off-the-shelf ASRs struggle with conversational speech. With Dutch exhibiting the lowest WER of 0.45, these systems made an error on roughly two out of three words in a conversation. While the rates do somewhat provide a look into overall performance, they fail to give specific information on the errors made and what needs to be taken into account in conversational speech recognition. An error analysis was conducted to supplement this and highlight missing conversational elements that impact performance.

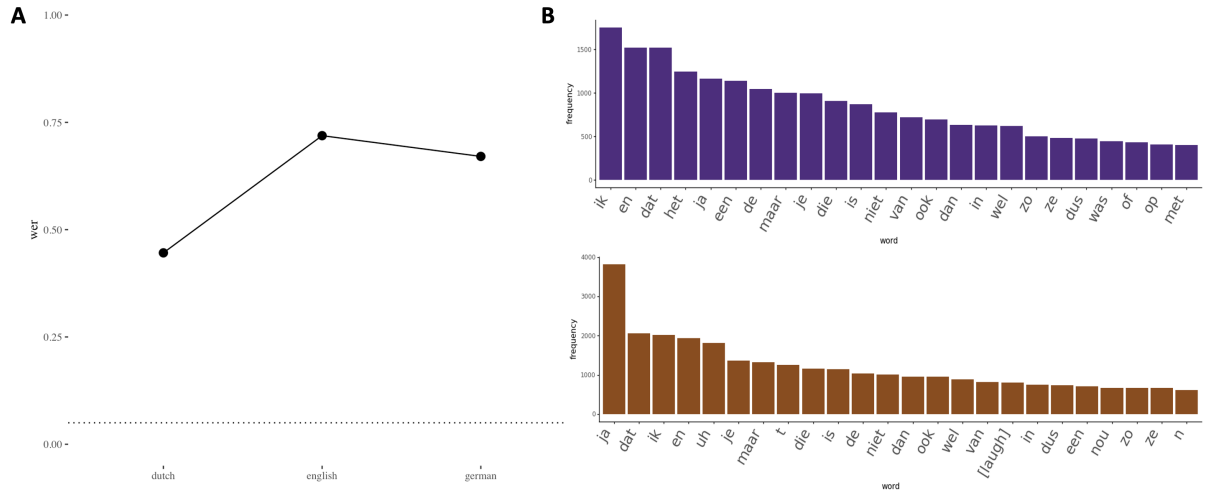


Figure 2: **A** Word error rates (WER) of general-purpose ASR systems on Dutch, English, and German conversational speech. **B** Frequency distributions of top tokens in Dutch conversational speech transcriptions produced by ASR (purple) and humans (orange).

Three key differences were found across the three languages. First, the ASR output contained fewer words than their corresponding human transcripts – with a 33% difference for Dutch, 37% for English, and 57% for German. This finding further shows that there is indeed substantial variation between how

ASRs and humans transcribe conversational speech Scharenborg (2007); Mansfield et al. (2021). Next, the frequency distributions of the ASR Transcripts are skewed differently from that of the human transcripts. This is shown in Figure 2B, where we observe distinct skewness in the frequency distributions of the top Dutch words in the ASR transcript (blue) compared to the human transcript (orange). Notably, the differences extend beyond just the skewness, as the top words themselves are different. This discrepancy suggests that the ASR systems have a tendency to either over- or under-represent certain words or patterns compared to the human transcripts.

To delve deeper into this discrepancy, I conducted an ngram salience score-based analysis Kessler (2017) (visualized in Figure 3). This analysis shed light on the specific elements that off-the-shelf ASR systems tend to miss across the three languages under investigation, bringing us to the final notable difference observed: general-purpose ASRs exhibit tend to miss certain conversational elements (see Table 3). One category of such missed elements is conversational words. Conversational words, which can be likened to interjections in the context of this study, refer to typically short utterances that are critical in managing the flow of a conversation. For instance, in English, an utterance like *uhhuh* shows understanding while a *huh?* shows the need for repair. Remarkably, these types of words were consistently underrepresented in the ASR transcripts across the three languages. Another category is reductions. These refer to shortened forms of words or phrases commonly used in conversational speech. Interestingly, no misrecognized or misrepresented reductions were observed in English. This could be attributed to two possible reasons. First, reductions in Dutch and German may be more prevalent in spoken language compared to English, resulting in their occurrence being less frequent in written texts used for training ASR systems. Second, the English training data used for the ASR system may be more extensive and comprehensive, enabling better recognition of reductions. Finally, self-repairs were also missed in the three languages studied. Self-repairs are abrupt interruptions or modifications in speech made by the speaker while monitoring their own utterance. The errors in recognizing self-repairs may stem from technical challenges, as these utterances are often brief. Additionally, self-repairs are not frequently included in training datasets due to their perceived "incompleteness", leaving ASR systems ill-equipped to handle them.

	Conversational Words	Reductions	Self-repairs
Dutch	*uh, hum, uhm, hum hum, oh, ja*	*'n beetje (een beetje), 't (het), d'r (haar), 'n (een), ie (hij)*	*k-, r -*
English	*uhhuh, mhm, uh, eh, um, hm, mm, ah, huh, okay*		*m-, e-*
German	*hm, mh*	*'n (ein), wa (wir), grade (gerade), det (das)*	*se-*

Table 3: Crucial conversational elements identified to be underrepresented or missing in general-purpose ASR transcripts across three languages. These elements were further categorized into three: conversational words (also referred to as standalone interjections in this thesis), reductions, and self-repairs.

4.2 Specialized ASRs on Conversational Speech

In contrast to general-purpose ASRs, specialized ASR systems are specifically designed to considerably address the nuances of conversational speech. Building upon the previous section's analysis of general-purpose ASRs on conversational speech, I now shift the focus to specialized ASR systems. Five specialized ASR were ran; namely, Amazon Transcribe, Google Cloud Speech-to-Text, NVIDIA NeMo ASR, Rev AI Speech-to-Text, and Whisper. These engines were ran on approximately one hour worth of conversations in six widely supported languages – Dutch, English, French, Korean, Mandarin, and Spanish.

An initial examination of the transcripts reveals notable disparities in specific measures pertinent to conversational speech (see Table 4). Across all six languages examined, conversational ASR systems demonstrated lower coverage, indicating a reduction in the total minutes of talk transcribed compared to human transcriptions. Spanish had the highest coverage, capturing 90.48% of the conversation, while Korean had the lowest coverage at 58.11%. This trend is also reflected in the number of words present in the transcripts, indicating that ASR systems struggle to capture and transcribe the entirety of the conversation. I also look at measures at the turn-level, describing turn length in terms of average length

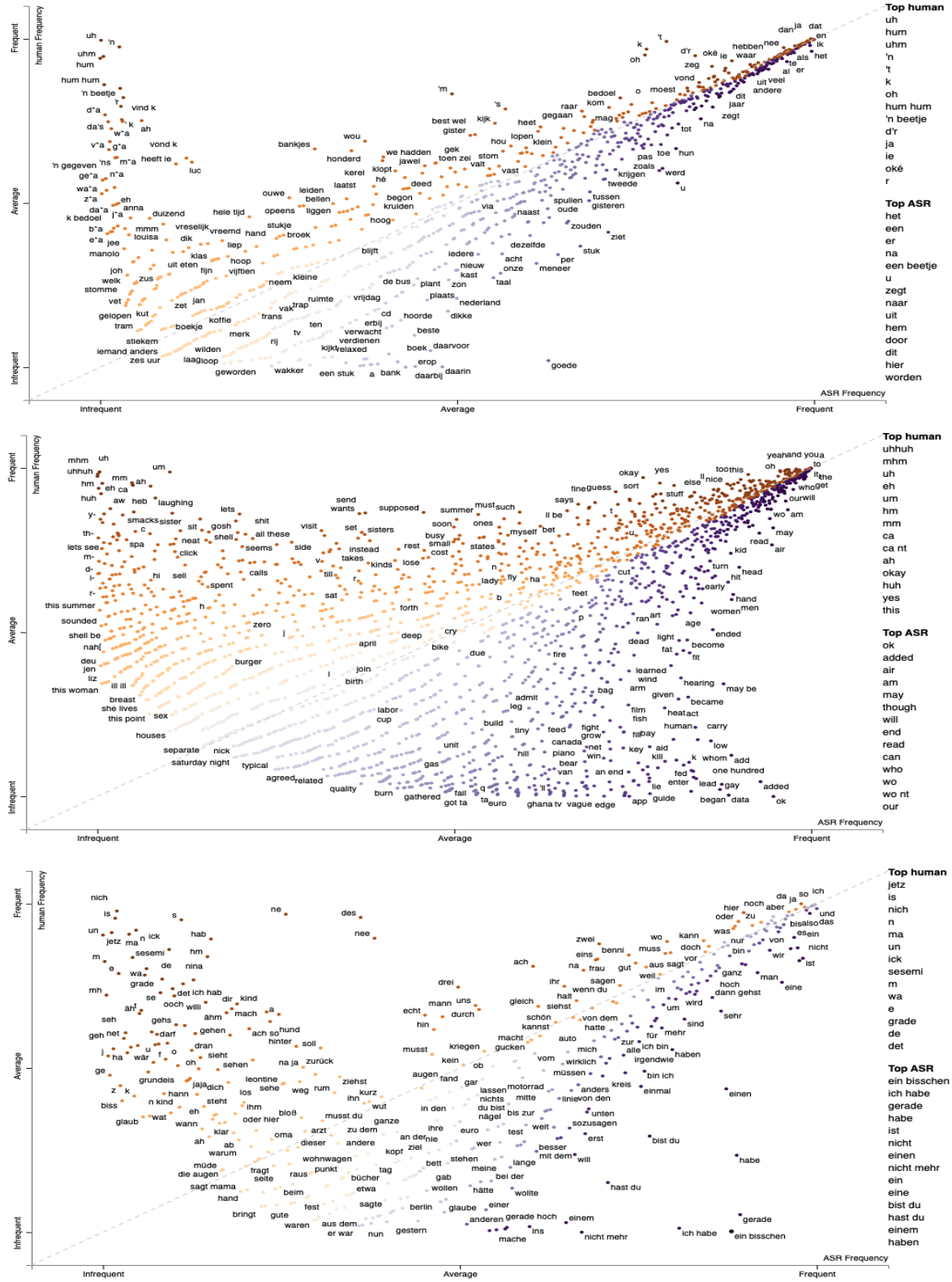


Figure 3: Plots showing the characteristic tokens, inferred from their frequency, present in human-transcribed (orange) and general-purpose ASR-transcribed (purple) conversational speech in Dutch, English, and German illustrated with `scaled F score` metric using `scattertext` Kessler (2017).

and standard deviation. The results revealed that ASR systems transcribed considerably shorter turns on average, but with generally lower variance across most languages – except for Dutch. Unfortunately, this uniformity in turn length exhibited by ASR systems does not accurately reflect the natural turn-taking patterns in real conversations. Figure 4B provides an illustrative example of a snippet of a conversation in English, where conversational turns are typically short, overlapping, and vary in length. Such dynamic and rapid turn-taking is a common characteristic of spontaneous conversations, where participants constantly engage in overlapping exchanges, indicating their collaborative efforts in constructing meaning. However, ASR systems' tendency to produce turns with consistent lengths overlooks this crucial aspect of conversational interactions, leading to a mismatch between the system's output and actual conversations. This discrepancy is further substantiated by the results of the last measure, which examines the percentage of overlapping speech. A striking finding is that none of the conversational ASR systems in any of the languages represent overlap. This highlights the consequence of condensing speech into a single dimension for easier transcription, rather than adequately reflecting the true timing and dynamics of spontaneous exchanges, leads to a measly representation of conversations.

	Coverage (min)	Words (n)	Duration (turn μ, ms)	Duration (turn σ, ms)	Overlap (speech %)
Dutch	63	12023	2840	3468	0.134
	47	9396	5897	7356	0
English	65	13895	2811	12057	0.126
	55	10994	6647	3316	0
French	64	13564	4357	11571	0.144
	49	8359	7042	4597	0
Korean	74	9632	3280	5615	0.208
	43	5923	4186	2824	0
Mandarin	66	15349	2538	12949	0.158
	53	8188	7301	2520	0
Spanish	63	11868	4620	8264	0.105
	57	10177	7534	5113	0

Table 4: Descriptive statistics that compare the transcripts generated by humans (top value) with the average of all conversational ASR systems (bottom value) in six languages.

Continuing the evaluation of conversational ASR systems, I present the word error rates in Figure 4A. This already offers several noteworthy observations. Firstly, most of the ASR systems achieve the lowest WER in English, indicating relatively better performance in transcribing conversational speech in this language. This advantage could be attributed to the availability of larger training datasets for English, allowing the ASR systems to learn and adapt more effectively. Secondly, the WERs vary significantly across different languages, implying that the complexity and linguistic characteristics of each language pose unique challenges for accurate transcription. It highlights the need to account for linguistic diversity in improving the performance of conversational ASR systems. Finally, the WERs for all the ASR systems, even in English, remain far from achieving the so-called human parity rate of 0.05. Despite English's advantageous resources and research, the ASR systems' performance is still significantly below human-level transcription.

As briefly mentioned before, figure 4B provides a visual comparison between the human-transcribed English conversation (top) and the transcriptions generated by the five specialized ASR systems. In this figure, the effect of the discrepancies found in Table 4 are better illustrated. It becomes evident that the ASR systems misrepresent the structure of a conversation almost entirely. The ASR transcripts display notably fewer speaker transitions and conversational turns compared to the human reference. The bottom system (Whisper) fails to indicate any turns at all, indicating a complete breakdown in capturing the basic conversational units. And to reiterate, overlaps – which frequently occur in the human transcript – are conspicuously absent.

For a deeper investigation into how ASR systems handle the interactional infrastructure, we turn our attention to Figure 4C, which presents the number of speaker transitions and the distribution of floor offset times (representing the gaps between turns) in the ASR-generated transcripts compared to the human transcripts. Notably, across all languages examined, the specialized ASR systems consistently

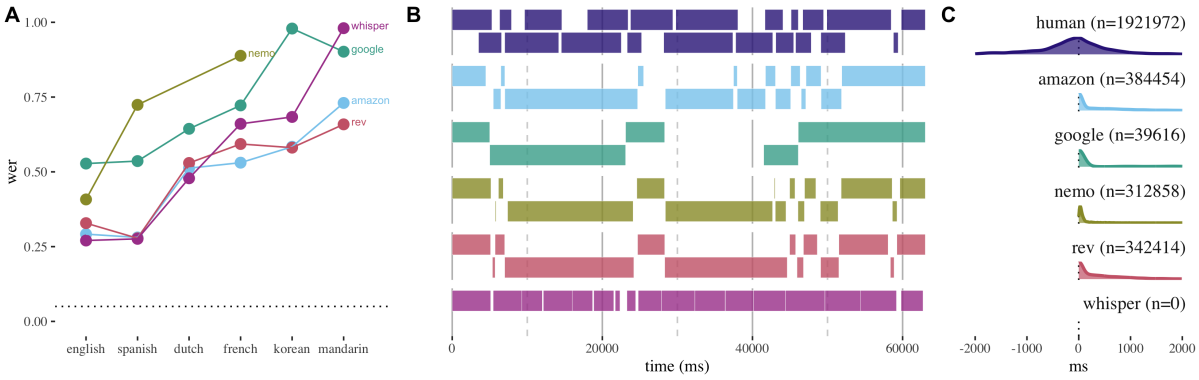


Figure 4: **A** Word error rates (WER) for five speech-to-text systems in six languages. **B** One minute of English conversation as annotated by human transcribers (top) and by five speech-to-text systems, showing that while most do some diarization, all underestimate the number of transitions and none allow overlapping turns (Whisper offers no diarization). **C** Number of speaker transitions and distribution of floor transfer offset times, showing that even speech-to-text systems that support diarization do not allow or represent overlapping annotations.

exhibit a reduced number of speaker transitions in their transcripts when compared to the human-transcribed conversations. This disparity suggests that accurately detecting and representing the changes in speakers during a conversation significantly influences the performance of ASR systems in capturing the conversational flow.

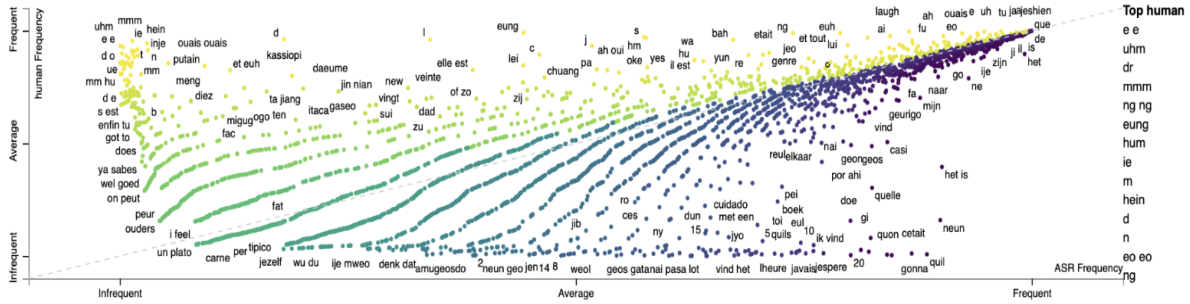


Figure 5: Plot showing the characteristic tokens, inferred from their frequency, present in human- transcribed (yellow) and specialized ASR-transcribed (blue) conversational speech in six languages illustrated with scaled F score metric using `scaled F score` metric using `scattertext` Kessler (2017).

	Standalone Interjections	Function Words	Discourse Markers
Dutch	*uhm, hum, hu, uh ja, mm*	*'n, ie, d'r, da, 's*	*en uh, dat uh*
English	*mhm, uhhuh, hm, oh, wow*	*did, she's, that's, going to, he*	*yeah I, because*
French	*hm hm, hein, ouais ouais, putain*	*c'etait, qu on, l', d', m'*	*euh tu, et euh*
Korean	*ahyu, eung, ye, eo, jeogi*	*hae, gajigo, jeo, geuge, jal*	*geureonigga, geuraegajigo*
Mandarin	*ng ng, ai, a, dui dui, er*	*la, wo wo, re, jiang, ya*	*shi er, gai, shi shuo*
Spanish	*eh, ah, he, claro, vale*	*o, eso, ahi, sea*	*o sea, sabes, verdad es*

Table 5: Crucial conversational elements identified to be underrepresented or missing in specialized ASR transcripts across three languages. These elements were further categorized into three: standalone interjections (similar to conversation words in Table 3), function words, and discourse markers.

In line with the analysis of generalized ASR systems, an ngram salience score-based error analysis (visualized for all languages in Figure 5 was conducted to identify specific conversational elements missed. Table 5 shows that specialized ASR systems struggles with standalone interjections, function words, and

discourse markers across all six languages. Standalone interjections, critical in conversational speech, were underrepresented in all six languages. Not only are these utterances short and often absent in typical ASR training data (Liesenfeld and Dingemanse, 2022), but they also often occur in overlap – which we now know ASR systems particularly struggle with. Meanwhile, the category of function words mostly contain short utterances that play important roles in grammar and turn structure. In conversational speech, these function words can occur in positions that are susceptible to overlaps or may undergo phonetic reduction, where certain sounds or syllables are shortened or omitted. This can pose challenges for ASR systems, particularly when the reduced or altered forms of these words are not accurately reflected orthographically in the training data. Lastly, discourse markers are often used at the start of an utterance to organize the structure of the conversation. Similar to standalone interjections, discourse markers are likely to occur in overlap-vulnerable regions and are not be well-represented in typical ASR training data. Overall, the results of the analyses on conversational speech recognition systems reinforces the critical importance of effectively representing the interactional infrastructure and resources in ASR.

5 Discussion

Conversational speech recognition has been a long-standing challenge in the field of language technology, primarily due to the unique complexities introduced by real-time interaction. Unlike more controlled speech settings, such as monologs and "walkie talkie" mode of exchanges, spontaneous conversations exhibit dynamic turn-taking patterns, overlapping speech, and various interactional resources that require special handling. In this thesis, I have delved precisely into these unique characteristics and examined two main things: (1) how current speech recognition (ASR) systems handle the intricacies of these natural interactions; and (2) how the widely used Word Error Rate (WER) is insufficient in describing the performance of these systems in conversational settings.

The central finding of this thesis is the striking disparity between humans and ASR systems in transcribing conversational speech. A comprehensive evaluation on both general-purpose and specialized ASRs, across multiple languages, resulted to consistently high WERs across all systems investigated – indicating the lack of robustness of these systems when it comes to handling spontaneous and free-flowing conversations. Among the general-purpose ASRs, the WER observed was 0.45 on Dutch, whereas among the specialized ASRs, it was 0.26 on English. These rates imply that, at its best, a general-purpose ASR system makes an error in approximately 1 out of every 2 words, while a specialized ASR system has an error rate of 1 in 4 words. These findings clearly demonstrate that ASR systems, in their current state, are not reliable or accurate enough to effectively handle conversational speech.

The high WERs observed across different languages and ASR systems underscore the limitations of current technology in handling conversational speech. While these rates provided a useful gauge on the overall performance of these systems, they fail to provide insights into how exactly ASR falls short. To address this, I conducted a comprehensive error analysis to shed light on the particular conversational aspects that pose challenges for ASR systems. The analysis revealed several critical elements that were consistently missed or inaccurately transcribed by both general-purpose and specialized ASR systems.

For general-purpose ASR systems, reductions and self-repairs were found to be particularly challenging across Dutch, English, and German. These utterances, which are more prevalent in spoken language, are underrepresented in typical ASR data. This highlights the importance of training ASR systems on data that closely resemble real-life interactions to ensure better performance in conversational settings. Specialized ASR systems, which are designed to handle conversational speech to some extent, were also evaluated in this analysis. Despite their specialized nature, these systems still exhibited errors in transcribing key interactional resources. Function words and discourse markers, which play significant roles in constructing and managing conversational turns, were consistently missed across the six different languages examined. This raises concerns because conversational systems should be capable of accurately representing the fundamental unit of conversation – the conversational turn. Finally, both general-purpose and specialized ASR systems struggled with transcribing conversational words or standalone interjections. These are essential in signaling understanding and miscommunication as well as in managing overall conversational flow. Overall, Interactional features – such as turn-taking patterns, overlaps, repairs, and conversational elements – are fundamental to natural human conversations. Neglecting these crucial features when developing conversational systems forces users to undergo a significant ad-

justment in how they naturally communicate. Imagine engaging with an agent that doesn't effectively handle timing aspects like pauses and overlaps. As a user, you might find yourself having to speak in fragments and waiting for the system to finish an inappropriate response before you can proceed, leading to frustrating and inefficient interactions. This is not only counterproductive but also diminishes the said benefits of conversational AI, which should ideally adapt to human interaction patterns rather than the other way around. The findings of this thesis underline the necessity to reshape how we develop and evaluate conversational AI systems. By emphasizing the integration of interactional features, we can progress towards human-centered, explainable, and truly robust conversational technology.

5.1 Towards a Better WER

The results of this thesis demonstrate significant discrepancies between ASR-generated transcriptions of conversational speech and those produced by humans. These disparities underscore the existing limitations in current conversational systems, as they struggle to capture the intricate details of natural interactions. The analysis further highlights the specific conversational elements that are frequently overlooked or inaccurately transcribed by ASR systems, emphasizing the necessity for improvement in this aspect. In this section, I build on the insights gained from the previous analyses, discuss how these conversational elements can be adequately represented in ASR systems, and how they can be combined in a single composite measure.

5.1.1 Representing what is said

Although it is well-known that Word Error Rate (WER) alone is not sufficient for evaluating conversational speech recognition, it remains a valuable metric for assessing transcription accuracy, which is a fundamental aspect of representing interaction. Rather than discarding WER entirely, I propose retaining it in the evaluation process to measure the fidelity of what is said in a conversation. By doing so, we can continue to capture the accuracy of ASR transcriptions, while also complementing WER with additional measures to address the broader complexities of conversational speech. This hybrid approach allows us to gain a more comprehensive understanding of ASR system performance in real-life interactions, making strides towards more robust conversational AI technologies.

5.1.2 Representing who said what

In addition to measuring accuracy, it is also important to represent the participants involved and the sequence of conversational turns. This entails identifying who said what during the interaction. To assess the performance of an ASR system in this regard, the Conversational Diarization Error Rate (CDER) emerges as a valuable evaluation metric Cheng et al. (2022). The CDER not only considers the identification of speakers but also takes into account the short phrases frequently produced in conversational exchanges. By incorporating the CDER into the evaluation metric, we can better assess the system's ability to recognize interlocutors and their contributions.

5.1.3 Representing conversational words

The findings of this thesis highlight the persistent challenge of ASR systems in accurately capturing crucial conversational words across various languages. These conversational words, although significant in managing the flow of conversation, are frequently missed by ASR systems. While identifying conversational words for every language requires annotation efforts, we can leverage the characteristics of these words to devise a practical operationalization.

To identify conversational words, we can focus on the fact that they occur with high frequency and sometimes even form their own individual turns in the conversation. With this knowledge, a straightforward approach is to extract the top ten most frequent turns from the human transcriptions. These turns are likely to contain the essential conversational words that contribute significantly to the meaning and structure of the conversation.

To operationalize this identification process, we can compute the percentage overlap of the top ten most frequent turns between the ASR-generated transcript and the human transcription. A higher percentage overlap indicates that the ASR system is accurately capturing and transcribing these crucial

conversational words. On the other hand, a lower percentage overlap would indicate that the ASR system is missing or inaccurately transcribing important conversational elements.

By employing this operationalization method, we can effectively measure the ASR system's performance in recognizing conversational words and assess its ability to capture the key elements of conversation. This approach provides a practical and informative way to evaluate the representation of conversational speech in ASR systems, leading to improvements in their performance and better alignment with human-level conversational interactions.

5.1.4 Representing speaker transitions

The findings from the analysis of ASR systems on conversational speech also reveal a notable discrepancy in the representation of speaker transitions, which are vital components of the interactional infrastructure. ASR systems were found to reflect significantly fewer speaker transitions compared to the human transcriptions, indicating a limitation in capturing the dynamic turn-taking patterns in conversational interactions. To ensure that the evaluation of conversational speech recognition systems adequately accounts for this, I propose to represent speaker transitions by calculating the percentage of speaker transitions detected by the system relative to the human data. Incorporating this will provide a more comprehensive assessment of the ASR system's ability to handle the complexities of conversational turn-taking and improve its overall performance in accurately representing the dynamics of spoken interaction.

5.1.5 Representing turn durations

In addition to inaccurately representing speaker transitions, ASR systems also fall short in capturing another crucial aspect of conversational dynamics: the durations of conversational turns. As emphasized previously, conversations predominantly involve swift and concise exchanges between participants. Yet, the findings reveal that ASR systems tend to generate substantially longer turns in their transcriptions compared to those produced by humans. This discrepancy arises due to the conventional training of ASR systems, which often involves scripted and lengthy speech segments. As a result, the systems struggle to adequately encapsulate the interactional structure.

To operationalize the accurate representation of turn durations, I propose to compare the average length of turn durations of ASR and humans. A higher average turn length in the ASR output compared to the human transcriptions would indicate that the ASR system tends to produce longer turns, which might deviate from the natural patterns of conversational speech. Conversely, a lower average turn length in the ASR output would suggest that the system is more successful in generating shorter turns, which aligns better with the rapid exchanges typical of human conversations. While the average turn length metric provides a simplified view of turn length distribution, it remains a valuable tool for evaluating how well an ASR system captures the overall pace and rhythm of conversational interactions.

5.1.6 Representing overlap

Finally, another notable discrepancy seen is that ASR systems do not represent overlap at all. Overlaps, as previously discussed, are essential in the interactional infrastructure and provide another dimension of meaning. A simple operationalization of this ability is calculating the percentage of turns percentage of turns occurring in overlap in ASR output, relative to same percentage in human. This can be done by first identifying turns that are overlapped.

5.1.7 Putting it all together

To summarize, I advocate to supplement the prevalent ASR evaluation metric, the Word Error Rate (WER), with an array of five additional measures tailored to address the vulnerabilities identified in conversational speech recognition. My recommendation is to assign equal weights to these measures, signifying the collective significance they hold in faithfully capturing spontaneous conversations. This balanced approach is a reflection of the findings of this thesis – that each aspect was found to uniquely contribute to the comprehensive and accurate representation of conversational speech. This composite metric is designed to be a seminal framework for future refinements guided by evolving empirical evidence.

The BWER (Better WER, pronounced as *beaver*) is defined as the following:

$$BWER = (0.2 \cdot WER) + (0.2 \cdot CDER) + (0.2 \cdot CW) + (0.2 \cdot ST) + (0.2 \cdot TD) + (0.2 \cdot O)$$

where *WER* is Word Error Rate, *CDER* is Conversational Diarization Error Rate Cheng et al. (2022), CW is measure of Conversational Words defined as:

$$CW = \frac{N \textit{ overlapping turns of the hypothesis and reference}}{10} \cdot 100$$

ST is measure of Speaker Transitions defined as:

$$ST = \frac{\textit{number speaker transitions in the hypothesis}}{\textit{number speaker transitions in the reference}} \cdot 100$$

TD is measure of Turn Durations defined as:

$$TD = \frac{\textit{average turn duration in the hypothesis}}{\textit{average turn duration in the reference}} \cdot 100$$

and finally, O is measure of Overlap defined as:

$$O = \frac{\textit{overlapped turns in the hypothesis}}{\textit{overlapped turns in the reference}} \cdot 100$$

5.2 Limitations

While this research has shed light on the limitations and opportunities in representing interaction in conversational speech recognition, it is important to acknowledge several limitations. Firstly, while the study has laid the groundwork for measuring various aspects of interaction and proposing a composite metric, the specific weights assigned to each element are preliminary placeholders at this stage. Further research and experimentation are needed to fine-tune and validate the proposed weighting scheme to ensure its effectiveness in capturing the nuances of conversational interactions accurately. Secondly, the datasets used in this research indeed represent a diverse set of languages, however all of them are considered high-resource. High-resource languages refer to languages for which a substantial amount of linguistic resources are available. The actual users of these technologies include communities that use underrepresented languages, and including these in the analysis may introduce additional challenges not identified in this research. Furthermore, while the proposal towards a better Word Error Rate (WER) offers promising avenues for enhancing the evaluation of conversational speech recognition systems, it is essential to acknowledge and address the remaining limitations of the WER metric. One notable limitation is the challenge of normalization. Normalization refers to the process of converting the ASR output and the human transcription into a standardized form to facilitate accurate comparison and evaluation. This poses as a complex task for conversations, as conversational words and other expressions characteristic of spontaneous speech can be transcribed in different ways (i.e. *"um"* vs *"uhm"*).

6 Conclusion

In this thesis, I have embarked on a comprehensive exploration of the challenges and opportunities in representing conversational speech using automatic speech recognition (ASR) systems. By examining the differences between human and ASR-generated transcriptions, I have highlighted the limitations of current technology in comprehensively capturing the interactional infrastructure and resources present in conversations. The findings underscore the need for a paradigm shift in the evaluation of conversational ASR systems, going beyond the traditional Word Error Rate (WER) to incorporate additional measures that address the identified weak spots. I propose the *BWER*, a composite measure that represents certain relevant measures of interaction – namely conversational words, turn durations, speaker transitions, and overlap.

Additionally, the different analyses conducted has shed light on the specific challenges faced by ASR systems across different languages, highlighting the importance of employing an approach informed by

linguistic diversity in improving speech technology. Overall, this thesis contributes to advancing the field of conversational AI by offering a comprehensive evaluation framework and highlighting the need for further research and development in this domain. It is my hope that this work will catalyze future studies that continue to push the boundaries of conversational speech recognition and ultimately lead to more natural, accurate, and inclusive interactions between humans and machines. As we strive for conversational AI that reflects the richness and complexity of human interactions as well as developments grounded in theory, this thesis serves as a stepping stone towards realizing that goal.

References

(2003). The Meaning of Intonational Contours in the Interpretation of Discourse. In Cohen, P. R., Morgan, J., and Pollack, M. E., editors, *Intentions in Communication*. The MIT Press.

Aksënova, A., van Esch, D., Flynn, J., and Golik, P. (2021). How Might We Create Better Benchmarks for Speech Recognition? In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, pages 22–34, Online. Association for Computational Linguistics.

Alač, M., Gluzman, Y., Aflatoun, T., Bari, A., Jing, B., and Mozqueda, G. (2020). Talking to a Toaster: How Everyday Interactions with Digital Voice Assistants Resist a Return to the Individual. *Evental Aesthetics*, 9(1):3–53.

Ameka, F. K. (1992). Interjections: The Universal Yet Neglected Part of Speech. *Journal of Pragmatics*, 18(2-3):101–118.

Auer, P. and Luzio, A. D. (1992). *The Contextualization of Language*. John Benjamins Publishing. Google-Books-ID: KU5IAAAAQBAJ.

Beach, C. M. (1991). The interpretation of prosodic patterns at points of syntactic structure ambiguity: Evidence for cue trading relations. *Journal of Memory and Language*, 30(6):644–663.

Bell, A. (1984). Language style as audience design. *Language in Society*, 13(2):145–204. Publisher: Cambridge University Press.

Berez-Kroeker, A. L., McDonnell, B., Koller, E., and Collister, L. B., editors (2022). *The Open Handbook of Linguistic Data Management*. The MIT Press.

Buchstaller, I. (2006). Social stereotypes, personality traits and regional perception displaced: Attitudes towards the 'new' quotatives in the U.K.1. *Journal of Sociolinguistics*, 10(3):362–381. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1360-6441.2006.00332.x.

Canavan, A., Graff, D., and Zipperlen, G. (1997). CALLHOME American English Speech. Artwork Size: 1830160 KB Pages: 1830160 KB.

Canavan, A. and Zipperlen, G. (1996a). CALLFRIEND Korean.

Canavan, A. and Zipperlen, G. (1996b). CALLHOME Mandarin Chinese Speech. Artwork Size: 1080128 KB Pages: 1080128 KB.

Cheng, G., Chen, Y., Yang, R., Li, Q., Yang, Z., Ye, L., Zhang, P., Zhang, Q., Xie, L., Qian, Y., Lee, K. A., and Yan, Y. (2022). The Conversational Short-phrase Speaker Diarization (CSSD) Task: Dataset, Evaluation Metric and Baselines. arXiv:2208.08042 [cs, eess].

Clancy, P. M., Thompson, S. A., Suzuki, R., and Tao, H. (1996). The conversational use of reactive tokens in English, Japanese, and Mandarin. *Journal of Pragmatics*, 26(3):355–387.

de Lima-Santos, M.-F. and Ceron, W. (2022). Artificial Intelligence in News Media: Current Perceptions and Future Outlook. *Journalism and Media*, 3(1):13–26. Number: 1 Publisher: Multidisciplinary Digital Publishing Institute.

Dingemanse, M. (2021). Interjections. In Lier, E. v., editor, *The Oxford Handbook of Word Classes*. Oxford University Press. type: article.

Dingemanse, M. and Liesenfeld, A. (2022). From text to talk: Harnessing conversational corpora for humane and diversity-aware language technology. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5614–5633, Dublin. Association for Computational Linguistics.

Dingemanse, M., Roberts, S. G., Baranova, J., Blythe, J., Drew, P., Floyd, S., Gisladottir, R. S., Kendrick, K. H., Levinson, S. C., Manrique, E., Rossi, G., and Enfield, N. J. (2015). Universal Principles in the Repair of Communication Problems. *PLOS ONE*, 10(9):e0136100.

Draxler, C., van den Heuvel, H., van Hessen, A., Calamai, S., and Corti, L. (2020). A CLARIN Transcription Portal for Interview Data. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3353–3359, Marseille, France. European Language Resources Association.

Drew, P. (2009). Quit talking while I'm interrupting: a comparison between positions of overlap onset in conversation. *Talk in Interaction: Comparative Dimensions*, pages 70–93.

Eckert, P. (2000). *Linguistic Variation as Social Practice: The Linguistic Construction of Identity in Belten High*. Wiley-Blackwell, Malden, Mass.

Errattahi, R., El Hannani, A., and Ouahmane, H. (2018). Automatic Speech Recognition Errors Detection and Correction: A Review. *Procedia Computer Science*, 128:32–37.

Fraser, B. (1990). An approach to discourse markers. *Journal of Pragmatics - J PRAGMATICS*, 14:383–398.

Furui, S., Nakamura, M., Ichiba, T., and Iwano, K. (2005). Why Is the Recognition of Spontaneous Speech so Hard? In Matoušek, V., Mautner, P., and Pavelka, T., editors, *Text, Speech and Dialogue*, Lecture Notes in Computer Science, pages 9–22, Berlin, Heidelberg. Springer.

Garrido, J. M., Escudero, D., Aguilar, L., Cardeñoso, V., Rodero, E., De-La-Mota, C., González, C., Vivaracho, C., Rustullet, S., Larrea, O., and others (2013). Glissando: a corpus for multidisciplinary prosodic studies in Spanish and Catalan. *Language resources and evaluation*, 47(4):945–971. Publisher: Springer.

Godfrey, J., Holliman, E., and McDaniel, J. (1992). SWITCHBOARD: telephone speech corpus for research and development. In *[Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 517–520 vol.1. ISSN: 1520-6149.

Goldwater, S., Jurafsky, D., and Manning, C. D. (2008). Which Words Are Hard to Recognize? Prosodic, Lexical, and Disfluency Factors that Increase ASR Error Rates. In *Proceedings of ACL-08: HLT*, pages 380–388, Columbus, Ohio. Association for Computational Linguistics.

Han, K. J., Chandrashekaran, A., Kim, J., and Lane, I. (2018). The CAPIO 2017 Conversational Speech Recognition System. arXiv:1801.00059 [cs].

Healy, E. W., Yoho, S. E., Wang, Y., and Wang, D. (2013). An algorithm to improve speech recognition in noise for hearing-impaired listeners. *The Journal of the Acoustical Society of America*, 134(4):3029–3038.

Heldner, M. and Edlund, J. (2010). Pauses, gaps and overlaps in conversations. *Journal of Phonetics*, 38(4):555–568.

Hirschberg, J., Litman, D., and Swerts, M. (2004). Prosodic and other cues to speech recognition failures. *Speech Communication*, 43:155–175.

Hirschberg, J. and Pierrehumbert, J. (1986). The Intonational Structuring of Discourse. In *24th Annual Meeting of the Association for Computational Linguistics*, pages 136–144, New York, New York, USA. Association for Computational Linguistics.

Hoegen, R., Aneja, D., McDuff, D., and Czerwinski, M. (2019). An End-to-End Conversational Style Matching Agent. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, IVA '19, pages 111–118, New York, NY, USA. Association for Computing Machinery.

Kawahara, T., Nanjo, H., Shinozaki, T., and Furui, S. (2003). Benchmark Test For Speech Recognition Using The Corpus . . .

Kessler, J. (2017). Scattertext: a Browser-Based Tool for Visualizing how Corpora Differ. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (System Demonstrations)*, pages 85–90.

Lelandais, M. and Thiberge, G. (2023). The role of prosody and hand gestures in the perception of boundaries in speech. *Speech Communication*, 150:41–65.

Lenk, U. (1998). Discourse markers and global coherence in conversation. *Journal of Pragmatics*, 30(2):245–257.

Levinson, S. C. (1983). *Pragmatics*. Cambridge Textbooks in Linguistics. Cambridge University Press, Cambridge.

Levinson, S. C. and Torreira, F. (2015). Timing in turn-taking and its implications for processing models of language. *Frontiers in Psychology: Language Sciences*, page 731.

Li, H. Z. (2001). Cooperative and Intrusive Interruptions in Inter- and Intracultural Dyadic Discourse. *Journal of Language and Social Psychology*, 20(3):259–284. Publisher: SAGE Publications Inc.

Liesenfeld, A. and Dingemanse, M. (2022). Building and curating conversational corpora for diversity-aware language science and technology. In *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*, pages 1178–1192, Marseille. arXiv: 2203.03399.

Liesenfeld, A., Lopez, A., and Dingemanse, M. (2023). Who says what when? Why timing is mission-critical for conversational speech recognition and dialogue systems.

Likhomanenko, T., Xu, Q., Pratap, V., Tomasello, P., Kahn, J., Avidov, G., Collobert, R., and Synnaeve, G. (2021). Rethinking Evaluation in ASR: Are Our Models Robust Enough? arXiv:2010.11745 [cs, eess].

Lippmann, R. P. (1997). Speech recognition by machines and humans. *Speech Communication*, 22(1):1–15.

Lopez, A., Liesenfeld, A., and Dingemanse, M. (2022). Evaluation of Automatic Speech Recognition for Conversational Speech in Dutch, English and German: What Goes Missing? In *Proceedings of the 18th Conference on Natural Language Processing (KONVENS 2022)*, pages 135–143, Potsdam, Germany. KONVENS 2022 Organizers.

Luger, E. and Sellen, A. (2016). "Like Having a Really Bad PA": The Gulf between User Expectation and Experience of Conversational Agents. pages 5286–5297.

Lööf, J., Falavigna, D., Schlüter, R., Giuliani, D., Gretter, R., and Ney, H. (2010). Evaluation of automatic transcription systems for the judicial domain. In *2010 IEEE Spoken Language Technology Workshop*, pages 206–211.

Manohar, V., Chen, S.-J., Wang, Z., Fujita, Y., Watanabe, S., and Khudanpur, S. (2019). Acoustic Modeling for Overlapping Speech Recognition: Jhu Chime-5 Challenge System. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6665–6669. ISSN: 2379-190X.

Mansfield, C., Ng, S., Levow, G.-A., Wright, R. A., and Ostendorf, M. (2021). Revisiting Parity of Human vs. Machine Conversational Speech Transcription. In *Interspeech 2021*, pages 1997–2001. ISCA.

Morris, A. C., Maier, V., and Green, P. (2004). From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition. In *Interspeech 2004*, pages 2765–2768. ISCA.

Munteanu, C., Penn, G., Baecker, R., and Zhang, Y. (2006). Automatic speech recognition for webcasts: how good is good enough and what to do when it isn't. In *Proceedings of the 8th international conference on Multimodal interfaces*, ICMI '06, pages 39–42, New York, NY, USA. Association for Computing Machinery.

Neri, A., Cucchiarini, C., and Strik, H. (2003). Automatic speech recognition for second language learning: How and why it actually works. *Speech Communication*.

Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. (2015). Librispeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210, South Brisbane, Queensland, Australia. IEEE.

Park, T. J., Kanda, N., Dimitriadis, D., Han, K. J., Watanabe, S., and Narayanan, S. (2022). A review of speaker diarization: Recent advances with deep learning. *Computer Speech & Language*, 72:101317.

Pasandi, H. B. and Pasandi, H. B. (2023). Evaluation of ASR Systems for Conversational Speech: A Linguistic Perspective. In *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*, SenSys '22, pages 962–965, New York, NY, USA. Association for Computing Machinery.

Pelikan, H. R. and Broth, M. (2016). Why That Nao? How Humans Adapt to a Conventional Humanoid Robot in Taking Turns-at-Talk. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pages 4921–4932, New York, NY, USA. Association for Computing Machinery.

Porcheron, M., Fischer, J. E., Reeves, S., and Sharples, S. (2018). Voice Interfaces in Everyday Life. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–12, Montreal QC Canada. ACM.

Reineke, S. and Schmidt, T. (2022). Das Archiv für Gesprochenes Deutsch und das Forschungs-und Lehrkorpus für Gesprochenes Deutsch. In *Sprache in Politik und Gesellschaft*, pages 323–330. de Gruyter.

Ruiz, N. and Federico, M. (2015). Phonetically-Oriented Word Error Alignment for Speech Recognition Error Analysis in Speech Translation.

Sacks, H. (1974). An analysis of the course of a joke's telling in conversation. In Bauman, R. and Sherzer, J., editors, *Explorations in the ethnography of speaking*, pages 337–353. Cambridge University Press, Cambridge.

Sacks, H., Schelgoff, E. A., and Jefferson, G. (1974). A Simplest Systematics for the Organization of Turn-Taking for Conversation. 50(4).

Saon, G., Kurata, G., Sercu, T., Audhkhasi, K., Thomas, S., Dimitriadis, D., Cui, X., Ramabhadran, B., Picheny, M., Lim, L.-L., Roomi, B., and Hall, P. (2017). English Conversational Telephone Speech Recognition by Humans and Machines. arXiv:1703.02136 [cs].

Scharenborg, O. (2007). Reaching over the gap: A review of efforts to link human and automatic speech recognition research. *Speech Communication*, 49(5):336–347.

Schegloff, E. A. (1979). The relevance of repair to syntax-for-conversation. In Givón, T., editor, *Syntax and Semantics*, volume 12, pages 261–286.

Schegloff, E. A. (1986). The Routine as Achievement. *Human Studies*, 9(2/3):111–151. ArticleType: primary_article / Issue Title: Interaction and Language Use / Full publication date: 1986 / Copyright © 1986 Springer.

Schegloff, E. A. (2000). Overlapping Talk and the Organization of Turn-Taking for Conversation. *Language in Society*, 29(1):1–63. ArticleType: primary_article / Full publication date: Mar., 2000 / Copyright © 2000 Cambridge University Press.

Sehgal, R. R., Agarwal, S., and Raj, G. (2018). Interactive Voice Response using Sentiment Analysis in Automatic Speech Recognition Systems. In *2018 International Conference on Advances in Computing and Communication Engineering (ICACCE)*, pages 213–218.

Shah, P., Chadha, H., Gupta, A., Dhuriya, A., Chhimwal, N., Gaur, R., and Raghavan, V. (2022). *Is Word Error Rate a good evaluation metric for Speech Recognition in Indic Languages?*

Shriberg, E. (2005). Spontaneous speech: how people really talk and why engineers should care. In *Interspeech 2005*, pages 1781–1784. ISCA.

Shriberg, E. and Stolcke, A. (2004). Direct Modeling of Prosody: An Overview of Applications in Automatic Speech Processing.

Shriberg, E., Stolcke, A., and Baron, D. (2001). Observations on overlap: findings and implications for automatic processing of multi-party conversation. In *7th European Conference on Speech Communication and Technology (Eurospeech 2001)*, pages 1359–1362. ISCA.

Stivers, T. (2008). Stance, alignment, and affiliation during storytelling: When nodding is a token of affiliation. *Research on Language and Social Interaction*, 41(1):31–57.

Stolcke, A. and Droppo, J. (2017). Comparing Human and Machine Errors in Conversational Speech Transcription. In *Interspeech 2017*, pages 137–141. ISCA.

Szymański, P., Żelasko, P., Morzy, M., Szymczak, A., Ży\la Hoppe, M., Banaszczak, J., Augustyniak, L., Mizgajski, J., and Carmiel, Y. (2020). WER we are and WER we think we are. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3290–3295, Online. Association for Computational Linguistics.

Taalunie (2014). Corpus Gesproken Nederlands - CGN (Version 2.0.3).

Torreira, F., Adda-Decker, M., and Ernestus, M. (2010). The Nijmegen Corpus of Casual French. *Speech Communication*, 52(3):201–212.

Vajpai, J. and Bora, A. (2016). Industrial Applications of Automatic Speech Recognition Systems.

van Son, R., Wesseling, W., Sanders, E., and van den Heuvel, H. (2008). The IFADV corpus: A free dialog video corpus. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*.

Wald, M. (2005). Using Automatic Speech Recognition to Enhance Education for All Students: Turning a Vision into Reality. In *Proceedings Frontiers in Education 35th Annual Conference*, pages S3G–S3G. ISSN: 2377-634X.

Wirth, J. and Peinl, R. (2022). ASR in German: A Detailed Error Analysis. arXiv:2204.05617 [cs].

Xiong, W., Droppo, J., Huang, X., Seide, F., Seltzer, M., Stolcke, A., Yu, D., and Zweig, G. (2017). Achieving Human Parity in Conversational Speech Recognition. arXiv:1610.05256 [cs, eess].

Xiong, W., Wu, L., Alleva, F., Droppo, J., Huang, X., and Stolcke, A. (2018). The Microsoft 2017 Conversational Speech Recognition System. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5934–5938. ISSN: 2379-190X.

Yang, L.-c. (2001). Visualizing Spoken Discourse: Prosodic Form and Discourse Functions of Interruptions. In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue*.

Zayats, V., Tran, T., Wright, R., Mansfield, C., and Ostendorf, M. (2019). Disfluencies and Human Speech Transcription Errors. In *Proceedings of Interspeech 2019*, pages 3088–3092. ISCA.

Çetin, and Shriberg, E. (2006a). Analysis of overlaps in meetings by dialog factors, hot spots, speakers, and collection site: insights for automatic speech recognition.

Çetin, and Shriberg, E. (2006b). Speaker Overlaps and ASR Errors in Meetings: Effects Before, During, and After the Overlap. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, volume 1, pages I–I. ISSN: 2379-190X.