Who says what *when*? Why timing is mission-critical for conversational speech recognition and dialogue systems

Anonymous ACL submission

Abstract

Speech-to-text systems are a key intermediary in voice-driven human-computer interaction. Although speech recognition does well for pristine monologic audio, real-life use cases in open-ended interactive settings still present many challenges. In this position paper we argue that timing is mission-critical for dialogue systems, and evaluate 5 major commercial ASR systems for their conversational and multilingual support. We find that word error rates for natural conversational data in 6 languages remain abysmal, and that overlap remains a key challenge (study 1). This impacts especially the recognition of conversational words (study 2), and in turn has dire consequences for downstream intent recognition (study 3). Our findings help to evaluate the current state of conversational ASR, contribute towards multi-dimensional error analysis and evaluation, and identify phenomena that need most attention on the way to build robust interactive speech technologies.

1 Introduction

Speech recognition (ASR) is a key technology in voice-drive human-computer interaction. Although by some measures speech-to-text systems approach human transcription performance for pristine audio (Stolcke and Droppo, 2017), real-life use cases of ASR in open-ended interactive settings still present many challenges and opportunities (Addlesee et al., 2020). The most widely used metric for comparison is word error rate, whose main attraction —simplicity— is also its most important pitfall. Here we build on prior work calling for error analysis beyond WER (Mansfield et al., 2021; Zayats et al., 2019) and extend it by looking at multiple languages and considering aspects of timing, confidence, conversational words, and dialog acts.

As voice-based interactive technologies increasingly become part of everyday life, weaknesses in speech-to-text systems are rapidly becoming a key bottleneck (Clark et al., 2019). While speech scientists have long pointed out challenges in diarization and recognition (Shriberg, 2001; Scharenborg, 2007), the current ubiquity of speech technology means new markets of users expecting to be able to rely on speech-to-text systems for conversational AI, and a new crop of commercial offerings claiming to offer exactly this. Here we put some of these systems to the test in a bid to contribute to richer forms of performance evaluation.

Related Work

The struggles of achieving truly conversational speech technologies are well documented. Spontaneous, free-flowing conversations are effortless and efficient for humans but remain challenging for machines (Shriberg, 2005; Baumann et al., 2017). Speech-to-text systems face an interconnected set of challenges including at least voice activity detection, speaker diarization, word recognition, spelling and punctuation, code-switching, intent recognition, and more (Suzuki et al., 2016; Sell et al., 2018; Addlesee et al., 2020; Park et al., 2022). Each of these represents a choice point with downstream consequences that may be hard to predict. Perhaps this is why word error rate, despite its noted defects (Aksënova et al., 2021; Szymański et al., 2020), has gained the upper hand in ASR evaluation: it makes no assumptions and simply delivers a single number to be optimized.

Speech scientists have long worked to supplement word error rate with more informative measures, including error analyses of overlap (Çetin and Shriberg, 2006), disfluencies (Goldwater et al., 2010), and conversational words (Zayats et al., 2019; Mansfield et al., 2021). This work has shown the importance of in-depth error analysis, and also brings home the multi-faceted challenges of truly interactive speech-to-text systems. As speech-to-text systems gain larger user bases, multilingual performance and evaluation becomes more impor-

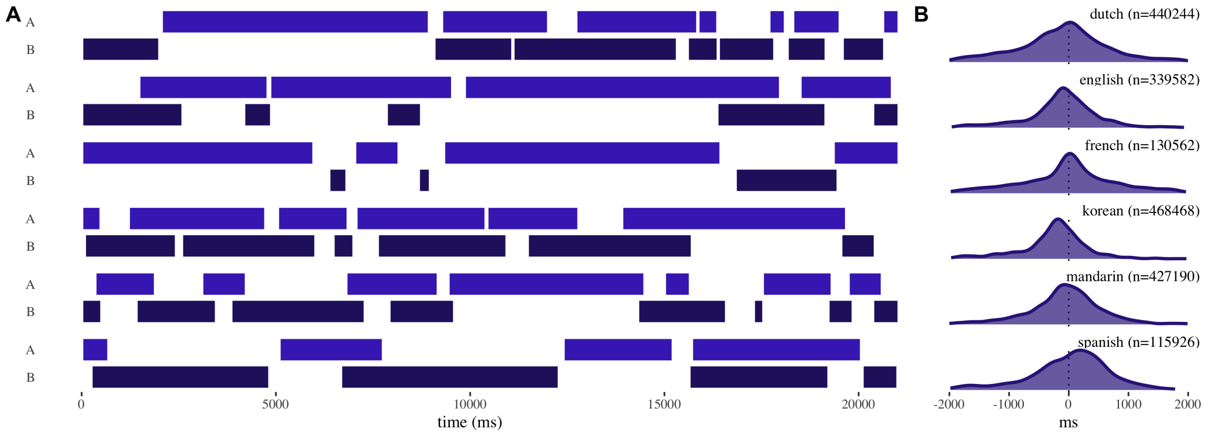


Figure 1: **A** Excerpts of 20 seconds of conversations in six languages, showing the short gaps and overlaps typical of human interaction. **B** Distribution of floor transfer offset times for the human test data across the same six languages, showing that the distributions are broadly normal and tend to peak around 0, with about as many turns occurring in slight overlap (negative values) as coming in after a slight gap (positive values).

tant (Levow et al., 2021; Blasi et al., 2022; Chan et al., 2022; Tadimeti et al., 2022).

The past decades of work on speech-to-text have led to remarkable improvements in many areas, and shared tasks have played an important role in catalyzing research efforts in diarization and recognition (Ryant et al., 2021; Barker et al., 2018). Still, we see opportunities for new contributions. Most work involves either non-interactive data or widely used meeting corpora, both of them quite distinct from the fluid conversational style people increasingly expect from interactive speech technology. When more conversational data is tested, it tends to be limited to English (Mansfield et al., 2021), raising the question how large the performance gap is in a more diverse array of languages (Besacier et al., 2014). While most benchmarks still rely on word error rates, true progress requires more in-depth forms of error analysis (Szymański et al., 2020) and especially a focus on the role of timing and overlap in speech recognition and intent ascription. Finally, the wide range of speech-to-text systems on offer in a time of need for robust conversational interfaces makes it important to know what current systems can and cannot do.

2 Aims and scope

A central question relevant at every moment of human interaction is *why that now?* (Schegloff and Sacks, 1973), referring to the importance of position and composition in how people ascribe intent to communicative actions. For speech-to-text systems, in order to even approach this question, a key prerequisite is to detect *who says what when*. This means that diarization, content recognition and precise timing are all highly consequential and best considered in tandem.

Here we address this challenge by presenting a multipronged approach that lays some of the empirical groundwork for improving evaluation methods and measures. Using principles of black-box testing (Beizer, 1995), we evaluate major commercial ASR engines for their claimed conversational and multilingual capabilities. We do so by presenting case studies at three levels of analysis. Study 1 considers word error rates and treatment of overlaps. Study 2 looks into what goes missing and why. Study 3 looks into the repercussions for intent recognition and dialog state tracking. We show that across these areas, timing is both a mission-critical challenge and an ingredient for ways forward.

Data and methods

Data preparation. We evaluate using a set of human-transcribed conversational data in multiple languages (Figure 1). We take several steps to ensure the dataset makes for a useful evaluation standard: (1) we pick languages that all or most of the tested systems claim to support (English, Spanish, Dutch, French, Korean, and Mandarin); (2) we source conversational speech data from existing corpora with high quality human-transcribed annotations that were published as peer-reviewed resources; (3) we ensure audio files have comparable audio encoding and channel separation, (4) we curate human transcriptions and timing information of each dataset for completeness and

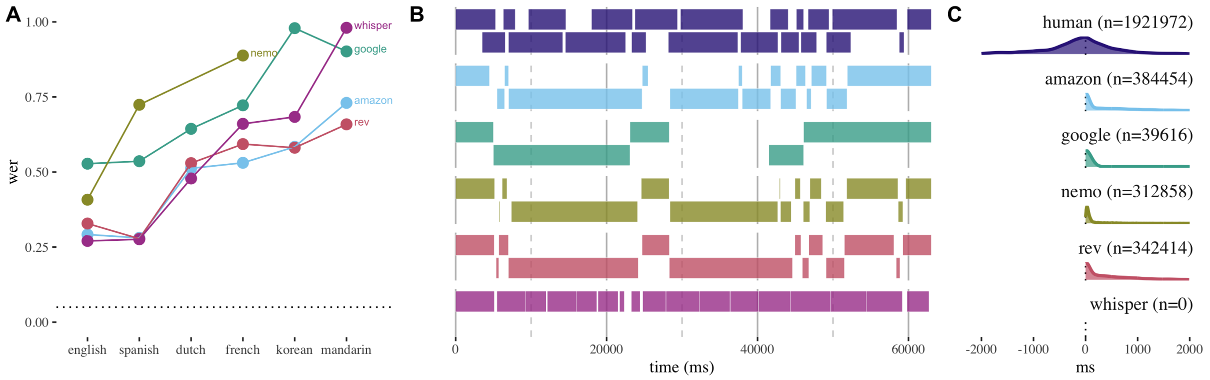


Figure 2: **A** Word error rates (WER) for five speech-to-text systems in six languages. **B** One minute of English conversation as annotated by human transcribers (top) and by five speech-to-text systems, showing that while most do some diarization, all underestimate the number of transitions and none 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.

accuracy, making sure that turn beginnings and ends are marked with at least decisecond precision (0.1ms); (5) we random-select one hour of dyadic conversations per language. More information on data sources and curation is available in this open data respository [anonymized review link]: `https://osf.io/hruva/?view_only=c0a54eba91a74121b644a3e16d4b35d4`.

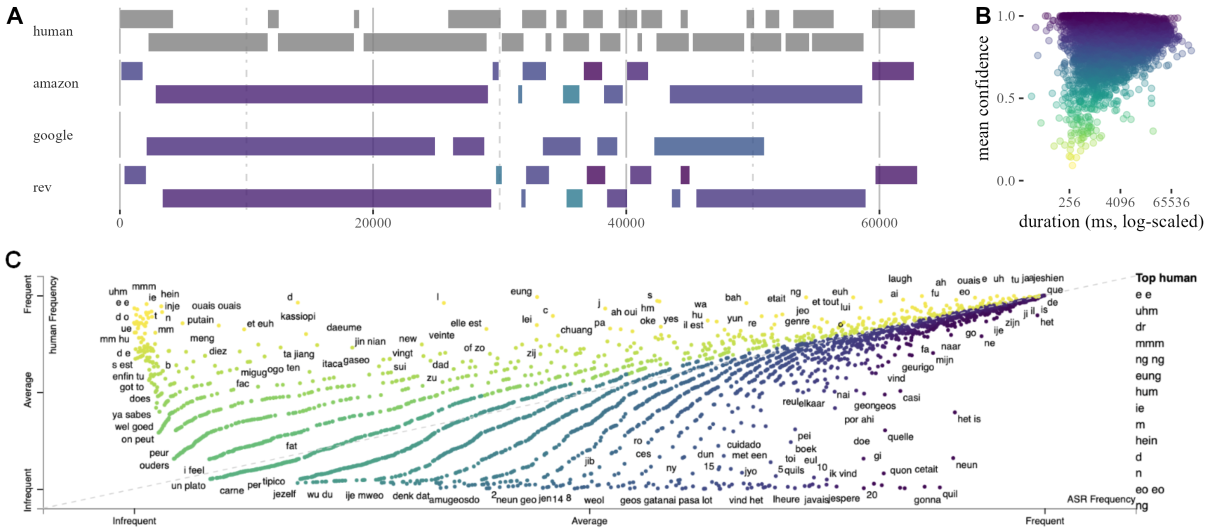
ASR system selection. Following principles of black-box testing (Beizer, 1995), we test five widely used ASR systems, keeping data and testing methods constant to compare them to human transcription baselines. Functional testing does not require access to model code or training data, instead treating models as black boxes tested to specification (Ribeiro et al., 2020). Enabling independent verification and evaluation, it is a key method in the toolbox of NLP evaluation methods.

We selected systems that claim to represent and handle conversational speech, and that offer multilingual support: (1) Amazon Transcribe 0.6.1, whose use cases include "transcription of voice-based customer service calls" and "generation of subtitles on audio/video content"; (2) Google Cloud Speech-to-Text API, using the `latest_long` model meant for "any kind of long form content such as media or spontaneous speech and conversations" (for French, Mandarin, and Spanish the long model is not available and we use the default model instead); (3) NVIDIA NeMo Quartznet15x15 for English and Conformer-CTC for French and Spanish, branded as a "Conversational AI Toolkit" that allows humans to "interact naturally"; (4) Rev AI Asynchronous Speech-to-Text API 2.17.1, which claims "accurate speaker separation" and support for "different speakers and conversations"; and (5) Whisper, a multilingual open-source neural net approaching "human-level robustness and accuracy on English speech recognition". We collected the finest-grain data available for each of these systems, using `whisper-timestamped` (Louradour, 2023) to extract word-level timing from Whisper, and `pyannote.metrics` (Bredin, 2017) for speaker diarization with NeMo.

Study 1: WER and overlap in 6 languages

Word error rates vary. We find that word error rates for truly conversational speech vary widely but nowhere approach the oft-cited human baseline of 5% transcription error (Figure 2A), a cross-linguistic replication of prior work on English (Mansfield et al., 2021). Most speech-to-text systems have the lowest rate in English, and even though all systems claimed multilingual support, all fare noticeably worse for typologically more different languages.

Overlap is lost. Human conversation typically features a rapid back-and-forth between participants, with a normal distribution of turn transition times centered around 0-200ms, and around half of all turns occurring in slight overlap (Figure 1; Figure 2B-C, top). Speech-to-text systems record substantially fewer speaker transitions and *no* overlapping annotations. Distributions of speaker transition times show the consequences: current speech-to-text systems miss out on about half of the turns that occur in overlap. Descriptive statistics further



D

	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’était, 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*

Figure 3: **A** Sample minute of Korean conversation comparing human-transcribed and ASR annotations, the latter coloured by mean confidence rating. Shorter utterances and regions with more overlap are associated with lower confidence. **B** Mean confidence for ASR-transcribed utterances (n=17.563) by duration, showing that across all languages, low confidence scores are associated with shorter utterances. **C** Most characteristic elements in human-transcribed (yellow) and ASR transcribed (blue) conversational speech across all languages plotted by Scaled F-score, with top 10 most distinctive items for human annotations on the right. **D** Top elements that are underrepresented or missing in ASR versus human-produced transcripts fall into three categories: short *conversational interjections*, high frequency *function words* (including contractions), and *discourse makers*.

corroborate this: by systematically not representing overlap, speech-to-text systems miss out on up to 15% of all speech (or around 1 in 8 words), which results in an inaccurate picture of conversational content, structure, and flow (Table 1 in Appendix).

Study 2: What goes missing and why

Crosslinguistic replication. Prior work on English has shown that it is especially short utterances and conversational words that go missing (Goldwater et al., 2010; Zayats et al., 2019; Mansfield et al., 2021). Here we replicate this for all six languages in our sample (Figure 3A).

Confidence metrics supplied by three of the speech-to-text systems provide a novel view of this: regions with more overlap and shorter utterances often coincide, and both are associated with dips in word-level and utterance-averaged confidence scores (Figure 3A-B). Across panels A, B and C, lighter coloured regions are associated with higher risk of being missed or misrecognized.

Overlap-vulnerability and reduction. We take the top N items by Scaled F-score, a n-gram scoring metric that adjusts F-score to better treat the extreme ends of the frequency distribution of the data (Kessler, 2017). Then we inductively classify them as standalone interjections, function words, and discourse markers (Figure 3D), following prior work (Zayats et al., 2019). We find that these categories provide good empirical coverage of what goes missing across all six languages in our sample.

Standalone interjections often occur in overlap-vulnerable contexts and are rare in ASR training data, often more formal and monologic (Liesenfeld and Dingemanse, 2022). The category of function words mostly contains highly frequent bits of morphosyntax that may occur in overlap-vulnerable positions (as the Mandarin final particles *la* and *ya*) or that are likely to be phonetically reduced (as in Dutch and French contractions of pronominal forms). Finally, discourse markers are utterance-initial fragments that help direct the flow of a conversation. These too occur in overlap-vulnerable regions and are rare in ASR training data.

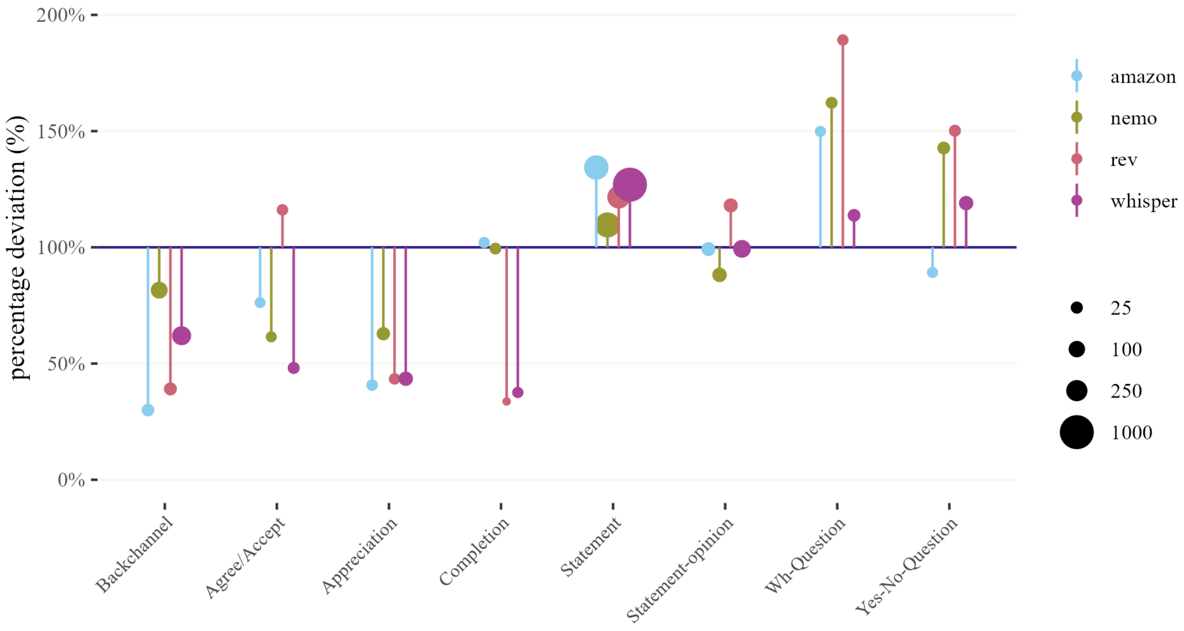


Figure 4: How different speech recognition engines warp dialog act classification in the same dataset of conversational English. For 8 frequent dialog acts, coloured lines show dialog acts based on ASR output deviate from those based on human transcripts of the same data (baseline). Dot size scales to number of times a tag is assigned. Only the most frequently assigned dialog acts (with at least 25 tokens in at least one dataset) are shown here. Mean absolute percentage deviations by ASR: 27.8% (nemo), 31.4% (amazon), 33.8% (whisper), 47.4% (rev).

Study 3: Consequences for dialog flow

So far we have seen that the tested systems struggle with timing and overlap (study 1) and especially underrepresent conversational elements of speech (study 2). But how serious are the consequences for actual dialogue systems? One way of gauging this is to consider intent classification, a downstream task that is key to dialog state tracking and to virtually any task-oriented application (Ye et al., 2022; Gella et al., 2022; Jacqmin et al., 2022).

As a minimal example, we use the Switchboard dialog act tagset (Stolcke et al., 2000) as implemented in the `dialogtag` Python library (Malik, 2021) and apply it to (i) human transcripts and (ii) ASR transcripts of the same English subset of our data. By keeping the dialog tagger and the underlying data constant and manipulating only the transcription method (human versus various ASRs) we make visible how reductions and variations introduced by speech recognition systems impact dialog act classification. We intentionally use the simplest possible dialog act tagger as a proof of concept. While several more sophisticated methods exist, every method is constrained by the data it can work with, and our goal here is to merely to make visible some of these constraints and their consequences for intent ascription and dialog state trackers.

We find that all ASRs warp dialog act classification outcomes in conversational English data (Figure 4). On average across the top 8 most frequently detected dialog act types, dialog act tags based on ASR output deviated between 27.8% (nemo) to 47.4% (rev) from tags based on human transcripts of the same data (this is absolute percentage deviation, i.e. including both overrepresentation and underrepresentation of dialog act tags).

Interactionally consequential dialog acts. Several highly interactionally relevant dialog act types are affected by speech-to-text systems. For instance (as expected based on the findings above), Backchannels and Agree/Accept tags are underrepresented across the board. This can be problematic for applications where it is important to keep track of user understanding and agreement during complex operations. Also, both the Wh-Question and Yes-No-Question dialog act tags tend to be overrepresented relative to the baseline. Since questions differ from other actions in the next moves they invite and expect, getting this wrong is directly consequential for any application in which user input is classified to determine relevant next actions.

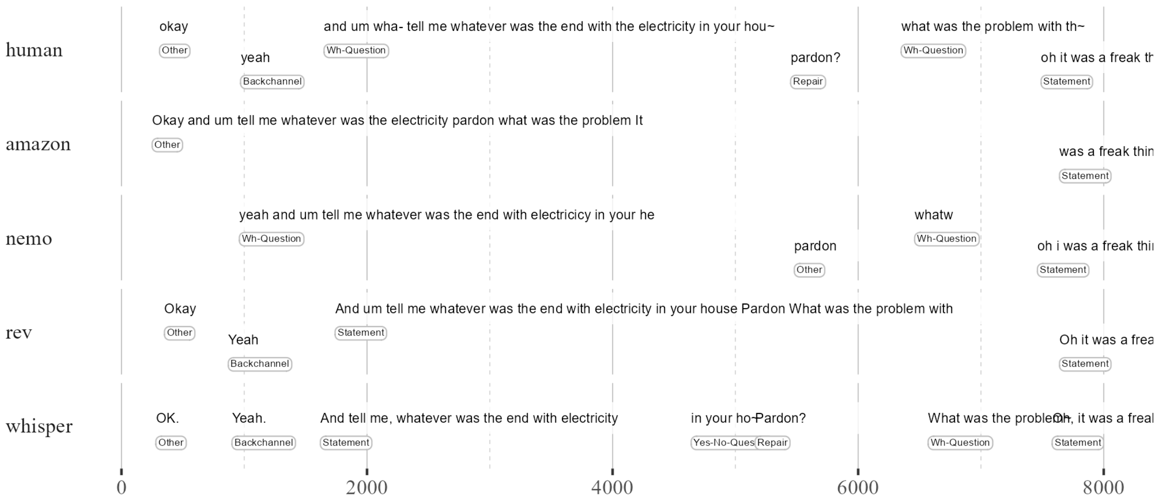


Figure 5: Excerpt of 8 seconds of English conversation showing how differences in how speech-to-text systems carry out segmentation, diarization, and transcription have direct consequences for dialog act classification.

What dialog act deformation looks like. Figure 5 shows an excerpt of English conversation in its human-annotated version (top) and four ASRs, with dialog act annotations. We selected this excerpt because it illustrates many of the larger scale patterns of underrepresentation and overrepresentation evident in Figure 4. Recall that dialog act tags are not supplied by the systems themselves, but applied to their output by `dialogtag`. Note that we speak of intent 'ascription' rather than 'recognition' to stress the fact that intents are often ambiguous and always provisional (Enfield and Sidnell, 2017).

Starting with relatively short conversational elements, we find that *yeah* is sometimes identified as a 'Backchannel' (rev, whisper), sometimes merged with adjacent turns by the other speaker (nemo), and sometimes elided entirely (amazon) — the latter two cases exemplifying the reasons ASR output generally underrepresents this category. Similarly, *pardon?* is variously identified as a 'Repair' signal (whisper), sometimes missed as a separate action because it is merged into adjacent turns by the other speaker (amazon, rev), and sometimes tagged as 'Other' (nemo), possibly because of punctuation.

Moving on to more complex elements, we see that a lumping approach to segmentation can result in interactionally important dialog acts going undetected: Amazon merges two disparate turns, producing *Okay and um tell me whatever was (...)*, which is tagged as Other. Meanwhile, a splitting approach, as Whisper appears to use, can lead to a fragment like *in your house* being tagged as Yes-No-Question in whisper output, showing one likely cause of over-representation of such question tags.

Disfluently produced questions can also pose problems: the utterance *and um wha- tell me whatever was (...)*, which features a self-repaired fragment, is sanitized and identified as a Statement in its rev and whisper versions. In the nemo output, the same turn (though merged, as we saw above, with a preceding "yeah" by the other speaker) is correctly tagged as a Wh-Question.

Even in this simple proof-of-concept, we see that ASR output can affect the recognition and classification of intents in various ways. This means that any real-world implementation relying on the systems tested here is hampered in its abilities to classify interactionally consequential social actions, making fluid interaction that much harder to achieve. Given the magnitude by which all tested ASRs deviate from human annotations in terms of timing, segmentation, diarization, overlap, and content, we expect similar kinds of distortion to appear in any systems for intent recognition or classification.

3 Discussion

The ubiquity of voice interfaces coupled with reports of human parity in speech recognition might make robust voice-driven interaction seem within easy reach. Indeed, all major vendors now advertise speech-to-text pipelines that claim both multilingual ability and conversational utility. Here we put five such systems to the test and find that the results are bleak: word error rates are nowhere near the oft-claimed human parity; performance drops dramatically for languages other than English; precise timing and diarization is hard to come by; overlap is systematically ignored; conversational words go missing; and as a result, intent recognition and state tracking are severely hampered.

Commercial speech-to-text systems are frequently exposed to conversational settings, whether it is in home use, business meetings, or customer service interactions. Our results imply that these systems are likely to fall short of several of their intended applications. Word error rate does not sufficiently reflect the performance of speech-to-text systems in most real-life contexts. The erasure of conversational elements and inability to deal with overlap renders these systems effectively oblivious to important aspects of user feedback. Differences in diarization and turn allocation across systems also have strong effects on dialog act classification, with the implication that switching vendors might have untold consequences for dialog state tracking and intent ascription.

Our results show that current speech recognition systems privilege what is said over when it is said; and that even systems claiming conversational utility appear to treat the problem as fundamentally one of turning a rich tapestry of turns into running text. These text-first design choices become visible when exposed to the rapid turn-taking patterns of natural conversation — not only to analysts in case studies like this, but inevitably also to users, where they cause friction, interactional turbulence, and user dissatisfaction. The results are in line with recent arguments that the current language technology landscape is fundamentally built around monologic text instead of dialogical talk (Dingemanse and Liesenfeld, 2022). The rise of conversational interfaces motivates a course correction if not a refurbishing of the foundations. Here we hope to have shown that data from human interaction can inspire such work.

3.1 Objections

One might object that our test data is unreasonably tough, featuring open-domain informal conversation with rapid turn-taking and lots of overlap. We agree, but would counter that it is at the same time reasonably realistic: this is what typical human interactive behaviour look like. The brute facts of human interaction are something speech-to-text systems will need to reckon with if there is to be a chance of the “natural interaction" and “human-level robustness" promised by current solutions.

One might object that missing 1 in 8 words and having word error rates hovering around 50% may not be fatal, depending on what goes missing. We agree, and point out that what goes missing here is crucial for interactive speech technology. Short recurring utterances like *mmhm*, *oh* and *huh?* are the swiss army knife of conversational competence. These items enable robust communication and fluid coordination; to erase them is to rob users of their agency and to stunt the interactive capabilities of conversational technology.

One might object that dialog acts are an imperfect and language-specific way of looking at intent ascription, and that automated tagging based on form alone does not do justice to the situatedness of action (Rollet and Clavel, 2020; Levinson, 1981). We agree, and have picked dialog acts merely as a proof-of-concept to illustrate the more general problem of garbage in, garbage out: defective diarization, missing words, and neglect of timing will hamper any form-based methods for intent ascription and dialog state tracking.

3.2 Limitations

We are aware of the following limitations.

First, the human reference data is internally quite diverse, differing in recording setting and audio quality. This makes comparisons across datasets harder, so we have refrained from drawing strong comparative conclusions about possible differences across corpora and languages, instead focusing on recurring patterns of what goes missing and why.

Second, we have not collected baseline measures for non-conversational data, making it hard to estimate how large the performance offset really is relative to more typical word error rate studies. Doing this would require a parallel data collection and curation exercise for each of the languages included in our study, which is outside our scope here but represents a good target for future work.

Third, given our choice to evaluate commercial vendor pipelines, we are unable to examine or report details about ASR system architectures, model parameters, and confidence score calculations. This is a necessary consequence of black-box testing. While direct access and manipulability offer important advantages from an engineering perspective, we nonetheless think it is also important to document and evaluate the performance of widely used commercial solutions.

3.3 Recommendations

The interconnectedness of all relevant processes in speech-to-text systems means that any quick fix likely has adverse consequences elsewhere. For instance, it is possible to improve diarization error rates by detecting and removing all overlap (Boakye et al., 2008) — but this means throwing out at least 15% of the data (as we show), putting human parity out of reach. Likewise, one may seek to reduce word error rates and interactional turbulence by excluding interjections (Papadopoulos Korfiatis et al., 2022), but this comes at the cost of losing all opportunity of rapid real-time user feedback. Our recommendations therefore focus on broadening the empirical basis, incrementalizing architectures, and doing more in-depth evaluation.

Improve ecological grounding. The most widely used datasets consist of monologic read speech in well-resourced languages. For speech-to-text systems to gain headway in truly interactive settings, they need to be exposed to more data that is closer to everyday language use both in conversational style and linguistic diversity.

Practice multidimensional evaluation. The downsides of word error rates have led to a flowering of alternative measures (Errattahi et al., 2018; Bredin, 2017). In time, the field will benefit from a degree of consolidation, and holistic evaluation of speech-to-text systems will require taking into account a wide range of measures, including but not limited to timing, diarization, overlap, coverage, phonology, spelling and word error rate.

Value qualitative error analysis. Simple metrics make for attractive optimisation goals, but are always vulnerable to mindless metrics gaming: when a measure becomes a target, it ceases to be a good measure (Strathern, 1996). Qualitative error analysis and thorough human evaluation will remain important to truly get a handle on what goes wrong and how things can be improved. This means incentives must be shifted to reward meaningful forms of evaluation over SOTA-chasing (Rogers, 2021; Church and Kordoni, 2022).

Strengthen incremental approaches. Rapid turn-taking, frequent overlap and disfluencies mean that speech-to-text systems, like human language processing, are probably best reconceived as an incremental architectures (Schlangen and Skantze, 2011). Promising work in this domain exists (Baumann et al., 2017; Addlesee et al., 2020), and this represents an important growth area.

Use timing when available. Current systems at least provide timing for non-overlapping stretches of talk, but even that is rarely used for intent ascription. This despite the fact that we know timing alone can change the interpretation of a turn like "Sure.", with longer delays flipping its polarity from positive to negative (Roberts and Francis, 2013). Timing might be used to improve at least some elements of intent ascription.

4 Conclusion

When you're a voice-driven conversational agent, life comes at you fast, and talk comes at you faster. We have presented evidence and arguments to support our contention that timing is more than a nice-to-have for any truly conversational system: it is mission critical and remains largely unsolved today. But rather than despair we take our findings as an opportunity to identify areas where novel work can make big differences. While diarization may remain a hard problem in ecologically valid settings, optimizing segmentation algorithms is likely to offer meaningful improvements. While overlap-vulnerable elements will always remain acoustically at risk, exposing ASRs to more ecologically valid training data and abandoning text-based transcript sanitizing techniques will likely improve the recognition of short conversational elements. And while intent ascription will always be hampered by missing data, taking timing into account will enable new gains.

Dealing with conversational words computationally is hard: not only are their forms short and prone to overlap, their meanings are cognitively demanding and interactionally subtle. A focus on information and sentence structure over interaction and sequential organization has long enabled us to look away from these elements. As conversational words are backgrounded as 'backchannels' and the artful interweaving of turns is classified as mere 'overlap' if not 'noise', it becomes easy to lose sight of the intricacies of human interaction. One way to see this paper is as contributing to a reframing that is underway in the language sciences at large: a reframing that foregrounds talk over text, that attends to interaction alongside information, and that recognizes the key role of timing. Timing is the secret sauce that can turn text into talk, chat into conversation, and perhaps, one day, clunky bots into fluid interactive tools.

Acknowledgements

[redacted]

References

Angus Addlesee, Yanchao Yu, and Arash Eshghi. 2020. A Comprehensive Evaluation of Incremental Speech Recognition and Diarization for Conversational AI. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3492–3503, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Alëna Aksënova, Daan van Esch, James Flynn, and Pavel Golik. 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.

Jon Barker, Shinji Watanabe, Emmanuel Vincent, and Jan Trmal. 2018. The Fifth 'CHiME' Speech Separation and Recognition Challenge: Dataset, Task and Baselines. In *Interspeech 2018*, pages 1561–1565. ISCA.

Timo Baumann, Casey Kennington, Julian Hough, and David Schlangen. 2017. Recognising Conversational Speech: What an Incremental ASR Should Do for a Dialogue System and How to Get There. In Kristiina Jokinen and Graham Wilcock, editors, *Dialogues with Social Robots: Enablements, Analyses, and Evaluation*, Lecture Notes in Electrical Engineering, pages 421–432. Springer, Singapore.

Boris Beizer. 1995. *Black-box testing: techniques for functional testing of software and systems*. Wiley, New York.

Laurent Besacier, Etienne Barnard, Alexey Karpov, and Tanja Schultz. 2014. Automatic speech recognition for under-resourced languages: A survey. *Speech Communication*, 56:85–100.

Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. Systematic Inequalities in Language Technology Performance across the World's Languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505, Dublin, Ireland. Association for Computational Linguistics.

Kofi Boakye, Oriol Vinyals, and Gerald Friedland. 2008. Two's a crowd: improving speaker diarization by automatically identifying and excluding overlapped speech. In *Interspeech 2008*, pages 32–35. ISCA.

Hervé Bredin. 2017. pyannote.metrics: A Toolkit for Reproducible Evaluation, Diagnostic, and Error Analysis of Speaker Diarization Systems. In *Interspeech 2017*, pages 3587–3591. ISCA.

May Pik Yu Chan, June Choe, Aini Li, Yiran Chen, Xin Gao, and Nicole Holliday. 2022. Training and typological bias in ASR performance for world Englishes. In *Interspeech 2022*, pages 1273–1277. ISCA.

Kenneth Ward Church and Valia Kordoni. 2022. Emerging Trends: SOTA-Chasing. *Natural Language Engineering*, 28(2):249–269. Publisher: Cambridge University Press.

Leigh Clark, Philip Doyle, Diego Garaialde, Emer Gilmartin, Stephan Schlögl, Jens Edlund, Matthew Aylett, João Cabral, Cosmin Munteanu, Justin Edwards, and Benjamin R Cowan. 2019. The State of Speech in HCI: Trends, Themes and Challenges. *Interacting with Computers*, 31(4):349–371.

Mark Dingemanse and Andreas Liesenfeld. 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.

N. J. Enfield and Jack Sidnell. 2017. *The Concept of Action*. Cambridge University Press, Cambridge.

Rahhal Errattahi, Asmaa El Hannani, and Hassan Ouahmane. 2018. Automatic Speech Recognition Errors Detection and Correction: A Review. *Procedia Computer Science*, 128:32–37.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.

Spandana Gella, Aishwarya Padmakumar, Patrick Lange, and Dilek Hakkani-Tur. 2022. Dialog Acts for Task Driven Embodied Agents. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 111–123, Edinburgh, UK. Association for Computational Linguistics.

Sharon Goldwater, Dan Jurafsky, and Christopher D. Manning. 2010. Which words are hard to recognize? Prosodic, lexical, and disfluency factors that increase speech recognition error rates. *Speech Communication*, 52(3):181–200.

Léo Jacqmin, Lina M. Rojas Barahona, and Benoit Favre. 2022. "Do you follow me?": A Survey of Recent Approaches in Dialogue State Tracking. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 336–350, Edinburgh, UK. Association for Computational Linguistics.

Jason Kessler. 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.

Stephen C. Levinson. 1981. The essential inadequacies of speech act models of dialogue. In Herman Parret, Marina Sbisà, and Jef Verschueren, editors, *Possibilities and Limitations of Pragmatics: Proceedings of the Conference on Pragmatics, Urbino, July 8-14, 1979*, pages 473–492. Benjamins, Amsterdam.

Gina-Anne Levow, Emily P. Ahn, and Emily M. Bender. 2021. Developing a Shared Task for Speech Processing on Endangered Languages. *Proceedings of the Workshop on Computational Methods for Endangered Languages*, 1(2).

Andreas Liesenfeld and Mark Dingemanse. 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.

Jerome Louradour. 2023. whisper-timestamped. Original-date: 2023-01-13T11:30:19Z.

Bhavitvya Malik. 2021. DialogTag: A python library to classify dialogue tag.

Courtney Mansfield, Sara Ng, Gina-Anne Levow, Richard A. Wright, and Mari Ostendorf. 2021. Revisiting Parity of Human vs. Machine Conversational Speech Transcription. In *Interspeech 2021*, pages 1997–2001. ISCA.

Alex Papadopoulos Korfiatis, Francesco Moramarco, Radmila Sarac, and Aleksandar Savkov. 2022. PriMock57: A Dataset Of Primary Care Mock Consultations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 588–598, Dublin, Ireland. Association for Computational Linguistics.

Tae Jin Park, Naoyuki Kanda, Dimitrios Dimitriadis, Kyu J. Han, Shinji Watanabe, and Shrikanth Narayanan. 2022. A review of speaker diarization: Recent advances with deep learning. *Computer Speech & Language*, 72:101317.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond Accuracy: Behavioral Testing of NLP Models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Felicia Roberts and Alexander L. Francis. 2013. Identifying a temporal threshold of tolerance for silent gaps after requests. *The Journal of the Acoustical Society of America*, 133(6):EL471–EL477.

Anna Rogers. 2021. Changing the World by Changing the Data. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2182–2194, Online. Association for Computational Linguistics.

Nicolas Rollet and Chloé Clavel. 2020. "Talk to you later": Doing social robotics with conversation analysis. Towards the development of an automatic system for the prediction of disengagement. *Interaction Studies. Social Behaviour and Communication in Biological and Artificial Systems*, 21(2):268–292.

Neville Ryant, Prachi Singh, Venkat Krishnamohan, Rajat Varma, Kenneth Church, Christopher Cieri, Jun Du, Sriram Ganapathy, and Mark Liberman. 2021. The Third DIHARD Diarization Challenge. ArXiv:2012.01477 [cs, eess].

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

Emanuel A. Schegloff and Harvey Sacks. 1973. Opening up closings. *Semiotica*, 8(4):289–327.

David Schlangen and Gabriel Skantze. 2011. A General, Abstract Model of Incremental Dialogue Processing. *Dialogue & Discourse*, 2(1):83–111. Number: 1.

Gregory Sell, David Snyder, Alan McCree, Daniel Garcia-Romero, Jesús Villalba, Matthew Maciejewski, Vimal Manohar, Najim Dehak, Daniel Povey, Shinji Watanabe, and Sanjeev Khudanpur. 2018. Diarization is Hard: Some Experiences and Lessons Learned for the JHU Team in the Inaugural DIHARD Challenge. In *Interspeech 2018*, pages 2808–2812. ISCA.

Elizabeth Shriberg. 2001. To 'errrr' is human: ecology and acoustics of speech disfluencies. *Journal of the International Phonetic Association*, 31(01):153–169.

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

Andreas Stolcke and Jasha Droppo. 2017. Comparing Human and Machine Errors in Conversational Speech Transcription. In *Interspeech 2017*, pages 137–141. ISCA.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech. *Computational Linguistics*, 26(3):339–373.

Marilyn Strathern. 1996. From Improvement to Enhancement: An Anthropological Comment on the Audit Culture. *Cambridge Anthropology*, 19(3):1–21.

Masayuki Suzuki, Gakuto Kurata, Tohru Nagano, and Ryuki Tachibana. 2016. Speech recognition robust against speech overlapping in monaural recordings of telephone conversations. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5685–5689. ISSN: 2379-190X.

Piotr Szymański, Piotr Żelasko, Mikolaj Morzy, Adrian Szymczak, Marzena Ży\la Hoppe, Joanna Banaszczak, Lukasz Augustyniak, Jan Mizgajski, and Yishay Carmiel. 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.

Divya Tadimeti, Kallirroi Georgila, and David Traum. 2022. Evaluation of Off-the-shelf Speech Recognizers on Different Accents in a Dialogue Domain. In *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*, pages 6001–6008, Marseille.

Chenchen Ye, Lizi Liao, Fuli Feng, Wei Ji, and Tat-Seng Chua. 2022. Structured and Natural Responses Co-generation for Conversational Search. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, pages 155–164, New York, NY, USA. Association for Computing Machinery.

Vicky Zayats, Trang Tran, Richard Wright, Courtney Mansfield, and Mari Ostendorf. 2019. Disfluencies and Human Speech Transcription Errors. In *Proceedings of Interspeech 2019*, pages 3088–3092. ISCA.

Özgür Çetin and Elizabeth Shriberg. 2006. Overlap in Meetings: ASR Effects and Analysis by Dialog Factors, Speakers, and Collection Site. In *Machine Learning for Multimodal Interaction*, Lecture Notes in Computer Science, pages 212–224, Berlin, Heidelberg. Springer.

A Appendix

A.1 Datasheets

Each processing step is reflected in the processing pipeline avaliable in the repository, which also includes a datasheet (Gebru et al., 2021) and instructions on how to replicate the study given access to the data. For Dutch and Spanish, the evaluation datasets are freely available for academic research purposes. For English, French, Korean and Mandarin, the study repository provides information how to obtain the datasets used: https://osf.io/hruva/?view_only=c0a54eba91a74121b644a3e16d4b35d4.

A.2 Study 1 methods

For both the human and ASR-transcribed data we calculate turn transition times in ms, number of speaker transitions, and the presence and duration of overlaps. For error analysis at the content level, we removed punctuation and excluded tags for non-speech behavior such as [laugh] and [breath] to bring all transcripts to a more comparable format. We used cleantext for pre-processing and whitespace for tokenizing. We then calculated word error rate (WER) using jiwer, and for a more in-depth investigation on the differences between human and speech-to-text annotations, we adopt Scaled F-score (Kessler, 2017) as a metric of n-gram salience scoring.

A.3 Study 1 detailed results

Table 1 provides a more detailed look at key differences between human transcriptions and ASR output across the six languages in our sample. For every language, it lists the mean amount of speech covered by the transcriptions (coverage); the mean total number of words in the transcripts (words); the mean turn duration in milliseconds; and the mean percentage of overlapping annotations.

Human vs ASR	Coverage (min)	Words (n)	Turn duration (ms)	Overlap (speech %)
Dutch	63	12023	2840	13.4
	47	9396	5897	0
English	65	13895	2811	12.6
	55	10994	6647	0
French	64	13564	4357	14.4
	49	8359	7042	0
Korean	74	9632	3280	20.8
	43	5923	4186	0
Mandarin	66	15349	2538	15.8
	53	8188	7301	0
Spanish	63	11868	4620	10.5
	57	10177	7534	0

Table 1: Comparison of human (top) and ASR transcripts (bottom) in each language in terms of coverage (amount of speech transcribed (in minutes), number of words, mean duration of each conversational turn (ms), and percentage of overlapped annotations relative to the length of the whole conversation. Human annotations add up to 395 minutes of transcribed speech; ASR-produced annotations for the same data on average add up to only 304, or 77% of the observed speech.