



What the Filler? Both ASR Systems and Humans Struggle More With Other Kinds of Disfluencies Than With Filler Particles

Saskia Wepner¹, Lucas Eckert¹, Gernot Kubin¹, Barbara Schuppler¹

¹Signal Processing and Speech Communication Laboratory, Graz University of Technology, Austria

wepner@tugraz.at, eckert@student.tugraz.at, gernot.kubin@tugraz.at,
b.schuppler@tugraz.at

Abstract

Since disfluencies are frequent in conversational speech, they have received notable attention: from speech technologists to make automatic speech recognition (ASR) more robust and from speech scientists to learn more about human speech processing. For ASR, the most established quality measure is the word error rate (WER), while for human recognition, one of the measures is the recall of words or utterance-level semantics. We conduct a transcription experiment in which we present the same disfluent utterances to 54 participants and nine ASR systems. We analyse which factors affect transcription in the context of syntactic disfluencies and filler particles, including well-known factors such as pronunciation variation and articulation rate. We find that, surprisingly, both humans and ASR struggle with largely the same characteristics of conversational speech – despite their mean WERs differing by about 10 % – and that the presence or absence of filler particles does not affect the WER. **Index Terms:** syntactic disfluencies, fillers, human transcription, automatic speech recognition, conversational speech

1. Introduction

Spontaneous conversations are characterised by the frequent occurrence of disfluencies in the form of (long) silent pauses, repetitions, repairs, and other types of utterance fragments that may stem from the joint co-construction of an utterance by more than one interlocutor [1, 2, 3]. One specific form of disfluency is the use of *filler particles* (FPs), a terminology we use here in line with Belz (2023) [4] to refer to semantically empty and syntactically unconstrained particles. FPs’ timing and acoustic characteristics have been analysed from different angles [5], depending on whether the aim was to learn more about speech planning in language processing (e.g., in bilingual speech [6]) or their function in discourse organisation [7]. When taking the listener’s perspective, FPs are a reliable cue for children to infer speaker intention in advance of object labelling [8] and help recall information for adults [9]. Further, words following FPs are recognised more quickly [10]. Given that FPs occur in different syntactic positions [5] and may be accompanied by a slower speech rate in their context (e.g., [11]), the question addressed in this paper is to which degree (word) recognition around FPs is affected by the occurrence of the FP itself and to which degree by disfluent syntax or acoustic characteristics.

Also in the field of speech technology, disfluencies have received attention. Dialogue systems shall become more interactional which has increased the interest in modelling non-lexical tokens for different functions: for backchanneling (e.g., [12]), and in the form of FPs for turn-holding and more naturally sounding longer turns (e.g., [5, 13, 14]). However, the aim of more interactional speech technologies comes with the need

for automatic speech recognition (ASR) systems to be robust to disfluencies produced by the users, making it crucial to improve word error rates (WERs) for spontaneous speech, in which FPs make up a large portion of verbal tokens (e.g., 10 % of all tokens in telephone conversations [15]). Replacing FPs with a unified symbol reduced WERs in spontaneous Japanese [16] and computer-directed English [17]. More fine-grained labelling of non-lexical tokens led to no confusions with lexical tokens and thus did not increase WER in Hungarian and Austrian German conversations [18]. These studies have in common that they report global WERs or separate WERs for lexical tokens and FPs but they do not provide information about the effect of FPs on WER in their context. To the best of our knowledge, [11] provides the only analysis on whether words in the context of FPs and other disfluencies are recognised equally well by HMM-based ASR systems as those in fluent utterances. In this paper, we fill this knowledge gap by analysing WERs from different transformer-based ASR systems in different types of syntactically disfluent structures, with or without FPs.

1.1. Study design

While for humans “there is a system in place that lets speakers deal with errors and disfluencies” [3, p. 68], humans are not used to *transcribing* on a daily basis. We thus do not assume that mistranscriptions made by humans are necessarily equivalent to misrecognitions. However, we use the term human speech recognition (HSR) in line with ASR. ASR systems, in contrast, are built for transcribing but are trained mainly on fluent speech (e.g., [19]). Our aim is to investigate how strongly disfluencies affect transcription for both ASR and HSR, with a specific focus on (1) the kind of syntactic disfluency, (2) the role of which part of the utterance was presented, and (3) the presence or absence of an FP. We therefore define:

(1) *Disfluency types*. All selected utterances were disfluent in one of the following ways: The structure that started before the disfluency was at that point ...

... syntactically incomplete¹ and was continued after the disfluency; thus it showed a durational disfluency only (durational):

es wäre dort eine interessante	äh	firma gewesen
(there would have been an interesting	er	company)

... syntactically incomplete and was *not* continued after the disfluency, but instead a new structure was started (syn+dur):

weil das ist	ähm	die fahren dich nämlich ganz rauf
(because that's	uhm	they drive you all the way up)

... syntactically complete (complete) and a new structure was

¹The disfluency occurred at a point of maximum grammatical control [20] and did not reach a transition relevance place (TRP) yet [21].

started after the disfluency, i.e., the speaker reached a transition relevance place (TRP) [21] before the pause:²

das ist eine andere geschichte (that's another story)	ahm umm	und er ist sechsundachtzig and he is sixty-eight
--	------------	---

We expect more transcription errors when a syntactic disfluency occurs in the middle of a syntactic structure than after a TRP.

(2) *Conditions.* To investigate the effect of the part of the utterance, we presented the stimuli in three different conditions: The whole utterance (condition *whole*), the part before the disfluency (*pre-disfluency*, in a dotted frame in the examples above), or the part after the disfluency (*post-disfluency*, dashed above). In *pre-disfluency*, we expect mistranscriptions for stimuli that are interrupted in the middle of a syntactic structure (*durational*, *syn+dur*), than those that are complete.

(3) *With or without FP.* Each utterance was presented with (*w/FP*) or without (*/oFP*) the FP being present in the stimulus (see Sec. 2.1, 2.2) to investigate the influence of the FP on transcription. Given the conflicting findings in the literature, we have no clear expectations concerning FPs.

As known from the literature, also other factors affect word recognition: e.g., word duration, articulation rate, pronunciation variation, lexical frequency, and whether surrounding words are hard to parse. We therefore include these factors in our analysis.

2. Method

2.1. Materials

We used the GRASS corpus [22], which contains one-hour long, unscripted, dyadic conversations in Austrian German.³ We searched the orthographic transcriptions for typical German FPs, such as *äh* [ɛ:], *ähm* [ɛ:m], *ahm* [a:m]. We excluded stimuli that contained foreign language words or additional disfluencies. Based on the available annotations of communicative functions for GRASS [23], we categorised the extracted utterances (with [24]) and categorised them according to disfluency types (*dType*), leaving us with a total of 46 utterances: 17 *durational*, 16 *syn+dur*, and 13 *complete*, consisting of 872 words (excl. FPs, see Tab. 1) which were transcribed by the participants (see Sec. 2.2) and by all ASR systems (see Sec. 2.3).

The stimuli without FP (*/oFP*) were created manually from the selected utterances: We replaced the FP in *whole* stimuli with a silence that resembled the recording's local noise floor. In those cases where removing the FP produced an audible artefact by cutting it off, we softened the edges of the signal (with attack and decay) to obtain naturally sounding stimuli. In conditions *pre-* and *post-disfluency*, we kept the FPs in cases without pause to the neighbouring word, otherwise, we cut it off.

2.2. Transcription experiment

54 native Austrian German speakers (26 female, 1 non-binary, 27 male, aged 19–74 years, mean 28.4 years) participated in the experiment; they were compensated financially. All were normal-hearing, none of them was dyslexic. We also asked them to rate their keyboard skills,⁴ which did not correlate with WER (−0.16). The experiment duration was approx. 60 min per participant (incl. training to familiarise themselves with the

²We did not use the continuation via an increment vs. rephrasing of the rear part as an additional factor as it was not our main focus.

³There has been much research on standard German as spoken in Germany; despite representing another variety of German, GRASS is one of the few unscripted conversational speech corpora for German.

⁴On a four-level scale from very skilled (4) to not skilled at all (1).

Table 1: Number of utterances by disfluency type, condition, and FP presence, and number of words excl. FPs (column 2).

Disfluency Type	Condition	w/FP	/oFP
<i>durational</i> (17)	<i>whole</i> (151 words)	17	17
	<i>pre-disfluency</i> (77)	2	15
	<i>post-disfluency</i> (74)	9	8
<i>syn+dur</i> (16)	<i>whole</i> (171 words)	16	16
	<i>pre-disfluency</i> (75)	4	12
	<i>post-disfluency</i> (96)	2	14
<i>complete</i> (13)	<i>whole</i> (149 words)	13	13
	<i>pre-disfluency</i> (74)	2	9
	<i>post-disfluency</i> (75)	2	11

setup and the stimuli). Each participant listened to a total of 46 utterances. Each stimulus was presented twice in immediate succession, allowing for correcting the first transcription.

The stimuli for each participant contained one specific utterance either in condition *whole* or in condition *pre-* and *post-disfluency* (in randomised order with the other stimuli). Further, each participant only heard one specific *whole* utterance either with FP or without FP. We made sure that stimuli from the same GRASS speakers did not occur consecutively in the experiment, to avoid listener adaptation effects. Additionally, the same stimulus was heard by some participants earlier and by some later in the experiment. Since participants were able to correct their first transcription after a second listening, there were two versions from each participant to each stimulus. We only evaluate the second, corrected transcription.

2.3. Automatic speech recognition systems

We decoded the utterances with nine ASR systems, using either Whisper [25] (models *large-v3/v2/v1*, *medium*, *small*, as of 22nd May 2023), and the *medium* model fine-tuned with the corpus data [26]; or *wav2vec2.0* [27], also fine-tuned to GRASS [26]⁵ (models *w2v2-lex-free* (greedy decoding), *w2v2-lex*, *w2v2-LM-2.0*). All utterances were fed to all ASR systems in all three conditions and with or without FP.

2.4. Data post-processing

We normalised the transcriptions of all participants and ASR systems by lower-casing, removing punctuation, duplicate white-spaces, and converting digits to number words. Since we were not interested in the transcription of the FPs themselves, we discarded them in all transcriptions (if any) and in the ground truth. For both HSR and ASR, we treated ambiguous orthographic representations of words as equally valid, for instance, if they originated from different segmentation, such as in separable verbs like ‘zurückzuführen’ vs. ‘zurück zu führen’ (*attributable to*) or in compounds. We accepted transcriptions that represented a typical reduced variant; for instance, (*I*) *have* is often pronounced [h a p] instead of ‘habe’ [h a: b ə], and frequently written as ‘hab’ in (casual) text messaging. For HSR, we also corrected obvious typos, because we were not interested in whether the participants were good at spelling.⁶

⁵Each of the fine-tuned systems did only see the data of the conversations that did *not* contain the speaker of the utterance to be decoded.

⁶Obvious typos were words with, e.g., drops or swaps of characters which did not result in words close to orthographic neighbours (e.g.,

Table 2: *Continuous and categoric variables in the analysis.*

Variable	Description
dType	Disfl. types durational, syn+dur, complete.
cond	Conditions whole, pre-, or post-disfluency.
w/oFP	Stimulus with FP (w/FP) or without (/oFP)?
wFreq	Logarithmic frequency (number of occurrences in the whole corpus) of the word.
pNeigh, sNeigh	Was the transcription of the previous/subsequent neighbour <i>correct</i> , <i>incorrect</i> or was the neighbour not there (<i>none</i>)? ⁷
wDur	Duration of the word in milliseconds.
pDist	Pronunciation distance, representing degree of reduction and strength of dialect.
funVScon	Was the word a function or content word?
artRate	Articulation rate of the stimulus in $\frac{\text{phones}}{\text{sec}}$.
who	Transcribed by ASR or a human (HSR)?

2.5. Feature extraction

We determined the WER for each stimulus and each participant/ASR system, and from this derived for each word whether its transcription was *correct* or *incorrect*. For each word, we determined its duration wDur (via forced alignments done with Kaldi using a pronunciation lexicon [28] covering the realised pronunciations in the data [29]), pronunciation distance pDist (as the Levenshtein distance of the realised pronunciation to its canonical representation), its lexical frequency wFreq, whether it was a function or content word funVScon (based on Part-of-Speech tags gained via TreeTagger [30] with the tag-set for spoken German [31, 32] manually corrected). We determined whether neighbouring words were transcribed correctly, encoded in the variables (pNeigh, previous neighbour and sNeigh, subsequent neighbour). Other factors that might influence transcription may be the number of words in the stimulus: Transformer-based ASR systems may benefit from longer stretches of speech whereas humans’ short-term memory is limited [33]. Since the number of words is highly correlated with the stimulus duration (0.70) and this in turn with the articulation rate (−0.42) and the conditions (as pre and post are naturally shorter than whole), we tested their correlation with the WER.⁸ As the WER had the highest correlation with the articulation rate of the stimulus (artRate), we included the articulation rate into the analysis. The full set of variables is shown in Tab. 2.

2.6. Statistical analysis

To model whether a word was transcribed correctly, we built a Conditional Inference Tree (CIT) [34], which divides the data based on significance while inherently handling interactions and allowing for direct interpretability; we used the R package partykit (version 1.2-10) with (word) correct as outcome and

⁷‘niht’ instead of ‘nicht’, *not*).

⁸Each word had at least one neighbouring word.

⁸Correlation of WER with number of words: 0.04 (ASR), 0.07 (HSR), stimulus duration: −0.1 (ASR), < 0.01 (HSR), and stimulus articulation rate: 0.23 (ASR), 0.11 (HSR).

the predictors dType, cond, w/oFP, wFreq, pDist, wDur, funVScon, artRate, pNeigh, sNeigh, who.

As we were interested in finding overall patterns rather than displaying peculiarities of only a few words, we reduced the alpha level from its default to $\alpha=.01$ ⁹ and set the minimum number of observations in the last split to 5 % of all words.

3. Results

The mean WERs of all utterances were 9.22 % for HSR across all 54 participants and 19.29 % for ASR across all 9 systems. Fig. 1, shows the mean WERs of the individual participants and ASR systems. Previously reported WERs for ASR on the whole GRASS corpus lie in the range of 15.27 % to 63.84 % (e.g., [26]), showing that our WERs lie in an expected range for the data. Five ASR systems achieved lower WERs than several human participants (4 for w2v-LM-2.0).

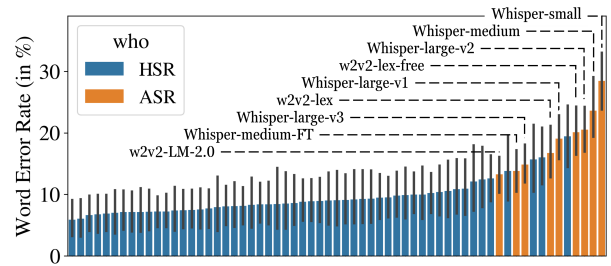


Figure 1: *Mean WER over all utterances separately for each participant and each ASR system (HSR: 9.22 %, ASR 19.29 %).*

In the following, we analyse which factors affected whether individual words were transcribed correctly or not with a CIT (see Fig. 2). Since each node in the tree – except for leaves 13 and 24 – contains a large number of words, we only mention these numbers in the figure and do not repeat them in the text.

The first two levels split the words based on their immediate neighbourhood. The branches on the right (nodes 26, 29, downwards) comprise words with at least one *incorrect* neighbouring word; these words were above average *incorrect* (45.63 %) and represent 15.19 % of the data. Among them, those words in an unsupportive neighbourhood (i.e., with no *correct* neighbour) were more often *incorrect* than *correct* (leaves 27, 31, 32: 69.20 % *incorrect*), while words where the respective other neighbour was *correct* were less often *incorrect* (leaves 28, 33: 24.75 % *incorrect*). High-frequency words ($wFreq > 3.87$) in an unsupportive neighbourhood were more often *incorrect* than words that occurred less frequently (cf. leaves 31 and 32).

Node 3 splits the words with at least one *correct* neighbour by dType into complete (3.29 % *incorrect*) vs. durational/syn+dur (5.02 % *incorrect*). Among the words in complete, those with a high frequency ($wFreq > 8.08$, leaf 10: 1.07 % *incorrect*) were most often *correct*; among the low-frequency words ($wFreq \leq 8.08$), those with a high pronunciation distance ($pDist > 0.11$) were more often *incorrect* (leaf 9: 7.81 % *incorrect*) than the words that were closer to the canonical pronunciation ($pDist \leq 0.11$). In these latter words, it played a role whether they were transcribed by ASR (leaf 8: 4.69 % *incorrect*) or by HSR (leaf 7: 2.64 % *incorrect*). Stimulus-initial words ($pNeigh=none$) in durational/syn+dur (node 11 downwards) were above average *incorrect* when they were

⁹ $\alpha_{\text{level}}=0.05$ deepened the tree without changing the main splits.

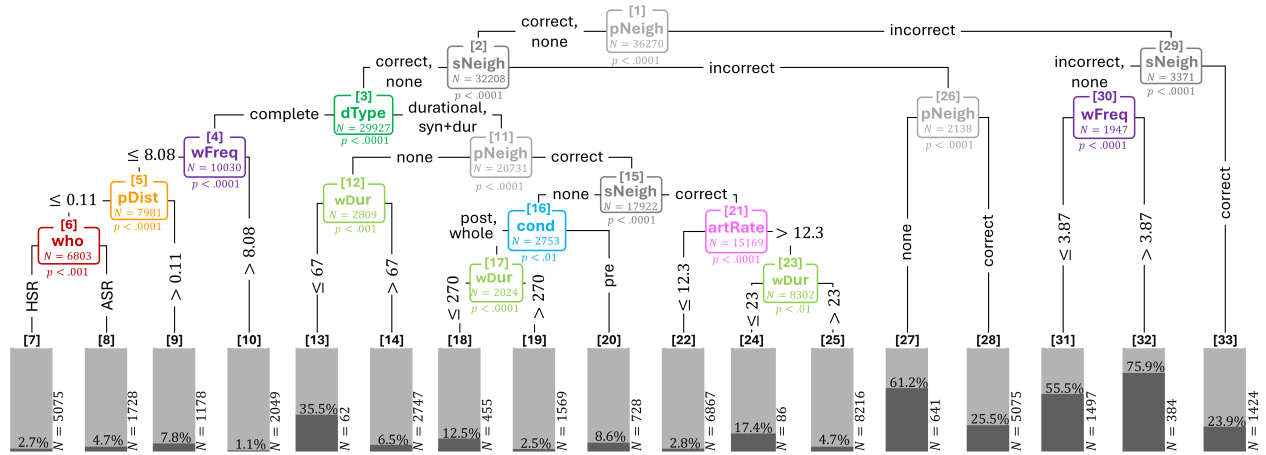


Figure 2: *Conditional Inference Tree of individual words (transcription correct or not). In total, 10.47 % of the words were incorrect. The percentage in the leaves represents the fraction of incorrect words in the respective leaf. Grey nodes represent neighbouring words while coloured nodes represent the characteristics of the word in question. Node numbers are given for reference in square brackets.*

short ($wDur \leq 67$ ms, leaf 13: 35.5 % *incorrect*), and below average *incorrect* when they were longer (leaf 14: 6.5 % *incorrect*). Stimulus-final words ($sNeigh = none$) in durational/syn+dur that were preceded by a *correct* neighbour were more often *incorrect* when they originated from condition whole or post-disfluency and were shorter than 270 ms (leaf 18: 12.5 % *incorrect*) than both longer words (leaf 19: 2.5 % *incorrect*) or those in condition pre-disfluency (leaf 20: 8.6 % *incorrect*). Words in durational/syn+dur in a supportive neighbourhood (i.e., both neighbours being *correct*) were more often *correct* in stimuli with lower articulation rates ($artRate \leq 12.3$, leaf 22: 2.8 % *incorrect*) or when they were longer and in stimuli with higher articulation rates (leaf 24: 17.4 % *incorrect*, $wDur \leq 23$ ms).

When allowing the tree to grow deeper, $funVScon$ became relevant for a subset of 455 words, while w/oFP became relevant for 157 words in a separate tree containing only words from pre- and post-disfluency showing that these factors only affected some particular cases but were not overall important.

4. Discussion and conclusion

We compared HSR and ASR to investigate how strongly disfluencies affect transcription quality, with a specific focus on (1) the kind of syntactic disfluency, (2) the role of which part of the utterance was presented, and (3) the presence or absence of an FP, while taking into account other factors from the literature known to affect WER. The mean WERs for ASR across all systems were approx. 10 % higher than for HSR. For both HSR and ASR, we found that words with mistranscribed neighbours were most often mistranscribed themselves. This result is intuitive and indicates that disfluencies affect speech processing of not just a single word. As reported for ASR [11], we showed that also HSR was affected by lexical frequency and pronunciation variation: Words of high lexical frequency resulted in worse transcriptions than low-frequency words when surrounded by mistranscribed neighbours, but not when surrounded by correctly transcribed neighbours within structures that were syntactically complete at the disfluency (i.e., the speaker reached a TRP). A larger distance from the canonical pronunciation led to worse transcriptions for infrequent words occurring in struc-

tures that reached a (syntactic if not necessarily prosodic) TRP at the disfluency. Finally, in each branch where the word duration played a role, shorter words were transcribed worse by HSR and ASR than longer ones (regardless of the duration's range in the respective branch). These findings are in line with those earlier reported for ASR [35, 36], while we here additionally show that they are also relevant to HSR.

What role do disfluencies play? Regarding (1), we found that for words having at least one correct neighbour, the disfluency type was the most important factor: Words from utterances that were syntactically incomplete at the disfluency were mistranscribed more often than those from utterances in which the disfluency consisted of a (filled) pause at a TRP. Regarding (2), we found that words immediately before a (filled) pause were transcribed worse in utterances without a TRP at the disfluency than in those with a TRP there. Hence, preceding correct neighbours compensated for utterances that were interrupted in the middle of a syntactic structure. Words in utterances that did not reach a TRP at the disfluency and were surrounded by correctly transcribed neighbours were transcribed worse when the articulation rate in the stimulus was high than when it was low. Regarding (3), our analysis showed that FPs played no significant role which is in line with [11] who reported a non-significant effect of FPs. Later studies found the FPs themselves often to be missed [37] or mistranscribed [38] by ASR, but they did not analyse the FPs' effect on surrounding words.

Limitations and Conclusions. ASR and HSR will never be fully comparable. We did our best to reduce the factors that cause differences, e.g., by allowing participants to listen twice and controlling for the number of words in the stimuli to decrease the effect of their limited short-term memory [33]. However, transformer-based ASR systems cannot fully exploit their potential in shorter utterances [39]. Further, we are aware that the number of selected utterances is small, which was necessary to keep an on-site experiment with participants feasible. We therefore focused on analysing the common patterns of mistranscribed words rather than on comparing absolute WERs of utterances. This allowed us to find that the factors affecting whether a word was mistranscribed or not were largely the same for HSR and ASR, showing that the same characteristics of spontaneous speech are difficult for both humans and ASR.

5. Acknowledgements

This research was funded in part by the Austrian Science Fund (FWF) [10.55776/P32700].

6. References

- [1] J. Ginzburg, *The Interactive Stance: Meaning for Conversation*. Oxford University Press, 2012.
- [2] E. Shriberg, “Disfluencies in Switchboard,” in *Proc. ICSPLP*, vol. 96, 1996, pp. 11–14.
- [3] M. Wiltchko, *The Grammar of Interactional Language*. Cambridge University Press, 2021.
- [4] M. Belz, “Defining Filler Particles: A Phonetic Account of the Terminology, Form, and Grammatical Classification of “Filled Pauses”,” *Languages*, vol. 8, no. 1, pp. 57–63, 2023.
- [5] S. Betz and M. S. Lopez Gambino, “Are We All Disfluent in Our Own Special Way and Should Dialogue Systems Also Be?” *Elektronische Sprachsignalverarbeitung*, vol. 81, 2016.
- [6] M. Böttcher and M. Zellers, “Do You Say uh or uhm? A Cross-Linguistic Approach to Filler Particle Use in Heritage and Majority Speakers Across Three Languages,” *Frontiers in Psychology*, vol. 15, pp. 1–19, 2024.
- [7] E. A. Schegloff, G. Jefferson, and H. Sacks, “The Preference for Self-Correction in the Organization of Repair in Conversation,” *Language*, vol. 53, no. 2, pp. 361–382, 1977.
- [8] C. Kidd, K. S. White, and R. N. Aslin, “Toddlers Use Speech Disfluencies to Predict Speakers’ Referential Intentions,” *Developmental Science*, vol. 14, no. 4, pp. 925–934, 2011.
- [9] S. H. Fraundorf and D. G. Watson, “The Disfluent Discourse: Effects of Filled Pauses on Recall,” *Journal of Memory and Language*, vol. 65, no. 2, pp. 161–175, 2011.
- [10] M. Corley and R. J. Hartsuiker, “Why um helps auditory word recognition: The temporal delay hypothesis,” *PLOS ONE*, vol. 6, no. 5, pp. 1–6, 05 2011. [Online]. Available: <https://doi.org/10.1371/journal.pone.0019792>
- [11] S. Goldwater, D. Jurafsky, and C. D. Manning, “Which Words Are Hard to Recognize? Prosodic, Lexical, and Disfluency Factors that Increase Speech Recognition Error Rates,” *Speech Communication*, vol. 52, no. 3, pp. 181–200, 2010.
- [12] O. Engwall, R. Cumbal, and A. R. Majlesi, “Socio-cultural perception of robot backchannels,” *Frontiers in Robotics and AI*, vol. 10, Jan. 2023.
- [13] R. Nakanishi, K. Inoue, S. Nakamura, K. Takanashi, and T. Kawahara, “Generating fillers based on dialog act pairs for smooth turn-taking by humanoid robot,” in *9th International Workshop on Spoken Dialogue System Technology*, L. F. D’Haro, R. E. Banchs, and H. Li, Eds. Singapore: Springer Singapore, 2019, pp. 91–101.
- [14] K. Yamamoto, K. Inoue, and T. Kawahara, “Character expression for spoken dialogue systems with semi-supervised learning using variational auto-encoder,” *Computer Speech & Language*, vol. 79, p. 101469, 2023. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0885230822000924>
- [15] E. Shriberg, “To ‘errrr’ is Human: Ecology and Acoustics of Speech Disfluencies,” *Journal of the International Phonetic Association*, vol. 31, pp. 153–169, 2001.
- [16] K. Horii, M. Fukuda, K. Ohta, R. Nishimura, A. Ogawa, and N. Kitaoka, “End-to-End Spontaneous Speech Recognition Using Disfluency Labeling,” in *Interspeech*, 2022, pp. 4108–4112.
- [17] V. Mendeleev, T. Raissi, G. Camporese, and M. Giollo, “Improved robustness to disfluencies in RNN-transducer based speech recognition,” in *ICASSP 2021*. IEEE, 2021, pp. 6878–6882.
- [18] P. Mihajlik, Y. Meng, M. Kádár, J. Linke, B. Schuppler, and K. Mády, “On Disfluency and Non-lexical Sound Labeling for End-to-end Automatic Speech Recognition,” in *Proc. of Interspeech 2024*, 2024, pp. 1270–1274.
- [19] M. Riviere, J. Copet, and G. Synnaeve, “ASR4REAL: An Extended Benchmark for Speech Models,” in *arXiv:2110.08583*, 2021.
- [20] E. A. Schegloff, “Turn Organization: One Intersection of Grammar and Interaction,” in *Interaction and Grammar*, E. Ochs, E. A. Schegloff, and S. Thompson, Eds. Cambridge: Cambridge University Press, 1996, pp. 52–133.
- [21] M. Selting, “On the Interplay of Syntax and Prosody in the Constitution of Turn-Constructional Units and Turns in Conversation,” *Pragmatics*, vol. 6, no. 3, pp. 371–388, 1996.
- [22] B. Schuppler, M. Hagmüller, and A. Zahrer, “A Corpus of Read and Conversational Austrian German,” *Speech Communication*, vol. 94, pp. 62–74, 2017.
- [23] A. Kelterer and B. Schuppler, “Turn-taking annotation for quantitative and qualitative analyses of conversation,” 2025. [Online]. Available: <https://arxiv.org/abs/2504.09980>
- [24] L. Eckert, S. Wepner, and B. Schuppler, “SLICER – A Tool for Efficient Stimuli Extraction from Large Speech Corpora,” in *Accepted for Forum Acusticum Euronoise*, 2025.
- [25] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, “Robust Speech Recognition via Large-Scale Weak Supervision,” in *Proc. ICML*. PMLR, 2023, pp. 28 492–28 518.
- [26] J. Linke, B. C. Geiger, G. Kubin, and B. Schuppler, “What’s so complex about conversational asr? A comparison of HMM-based and transformer-based ASR architectures,” *Computer Speech & Language*, vol. 90, 2025.
- [27] Facebook Research, “Fairseq Model (XLSR),” in *GitHub Repository*, 2022. [Online]. Available: <https://github.com/facebookresearch/fairseq/tree/main/examples/wav2vec>
- [28] S. Wepner, “pvlex – Lexicon with Pronunciation Variants for (Austrian) German Conversational Speech,” Jan 2025. [Online]. Available: <https://github.com/SPSC-TUGraz/pvlex>
- [29] J. Linke, S. Wepner, G. Kubin, and B. Schuppler, “Using Kaldi for Automatic Speech Recognition of Conversational Austrian German,” in *arXiv:2301.06475*, 2023.
- [30] H. Schmid, “Improvements in Part-of-Speech Tagging with an Application to German,” in *Proc. SIGDAT*, 1995, pp. 1–9.
- [31] S. Westpfahl, “STTS 2.0? Improving the Tagset for the Part-of-Speech-Tagging of German Spoken Data,” in *Proc. 8th Linguistic Annotation Workshop*, L. Levin and M. Stede, Eds. Dublin: Association for Computational Linguistics and Dublin City University, 2014, pp. 1–10. [Online]. Available: <https://nbn-resolving.org/urn:nbn:de:bsz:mh39-31897>
- [32] S. Westpfahl and T. Schmidt, “FOLK-Gold—A Gold Standard for Part-of-Speech-Tagging of Spoken German,” in *Proc. LREC*, 2016, pp. 1493–1499.
- [33] A. D. Baddeley and G. J. Hitch, “The Phonological Loop as a Buffer Store: An Update,” *Cortex*, vol. 112, pp. 91–106, 2019.
- [34] N. Levshina, “Conditional Inference Trees and Random Forests,” in *A Practical Handbook of Corpus Linguistics*, M. Paquot and S. T. Gries, Eds. Springer, 2021, pp. 611–643.
- [35] J. Hirschberg, D. Litman, and M. Swerts, “Prosodic and other cues to speech recognition failures,” *Speech Communication*, vol. 43, no. 1, pp. 155–175, 2004.
- [36] S. Wepner, B. Schuppler, and G. Kubin, “How Prosody Affects ASR Performance in Conversational Austrian German,” in *Proc. Speech Prosody*, 2022, pp. 195–199.
- [37] A. Lopez, A. Liesenfeld, and M. Dingemans, “Evaluation of Automatic Speech Recognition for Conversational Speech in Dutch, English and German: What Goes Missing?” in *Proc. 18th ICNLP*, 2022, pp. 135–143.
- [38] S. O. Russell, I. Gessinger, A. Krason, G. Vigliocco, and N. Harte, “What automatic speech recognition can and cannot do for conversational speech transcription?” *Research Methods in Applied Linguistics*, vol. 3, no. 3, p. 100163, 2024.
- [39] J. Linke, J. Winkler, and B. Schuppler, “Context Is All You Need? Low-Resource Conversational ASR Profits From Context, Coming From the Same or From the Other Speaker,” *Accepted for Interspeech*, 2025.