forum acusticum 2026
8-12 September · Graz, Austria

HOW 'AEHM' AND 'ALSO' AFFECT PERCEIVED FLUENCY, CLARITY AND UNDERSTANDABILITY

Johannes Hagmüller[1], Barbara Schuppler[1]

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

This paper is a revised version of the authors' submitted manuscript that was peer-reviewed by at least two anonymous reviewers.

Abstract

This study investigates the perception of disfluencies in speech, focusing on the fillers "aehm" and "also". Using conversational speech recordings from the Austrian German GRASS corpus, we select 132 stimuli similiarly long in number of tokens that contained either "aehm" or "also" and conduct a perception experiment with 30 participants. During the experiment, listeners rated for each stimulus how well they understood the content of the stimulus (i.e., understandability), how clearly pronounced or reduced they perceived the stimulus (i.e., clarity), and also how fluently spoken or disturbingly disfluent they perceived the stimulus. Our statistical analysis shows that speech stimuli containing "aehm" tend to receive significantly higher ratings in clarity, understandability, and fluency than stimuli containing "also". Further, the results show correlations among listeners' perceptions of these three dimensions. Overall, the findings contribute to a better understanding of how specific disfluency markers influence speech perception and highlight the importance of considering specific filler types when studying spoken language processing.

Keywords: *conversational speech, disfluencies, filler particles, perceived fluency, Austrian German*

1. Introduction

One salient characteristic distinguishing read or prepared speech from speech stemming from casual, lively conversational speech is the high amount of disfluencies, either in form of broken and repeated words, prolonged segments, long pauses, and the insertion of filler particles (e.g., [1], [2]). Particles like 'uhm' are referred to in the literature as 'filled pause', 'filler', 'delay marker', or 'discourse particle' [3]. Belz [3] argues that the term 'filler particle' is the most suitable one to refer to semantically empty and syntactically unconstrained particles. Clark and Fox Tree [4] analyzed filler particles but also interjections like 'ugh' and 'damn', which express current emotions of the speaker and 'huh' and 'well' which describe a current state of knowledge of the speaker. They also provide an explanation of the function of 'um' and 'uh': They are used for signalizing an expected minor or major delay before speaking, which are planned and produced by them as a lexical word. This raises the question of whether also listeners perceive 'ahm' similarly to a disfluency marker that has a form shared with a lexical word.

In this paper, we focus on two particles in (Austrian) German frequently occurring around disfluent speech sequences: First, the filler particle 'aehm', the equivalent to 'uhm' in English. Second we focus on disfluent utterances containing 'also'. 'Also' is a lexical item with various meanings and functions: discourse marker ('*So*, what do we do now?'), connective ('thus, like'), pragmatic particle showing speaker stance ('Also bitte!' translates to 'Come on!'), and also as disfluency particle frequently occurring before a repair (e.g., shown in an analysis of different functions of 'like' by [5]). Our study concentrates on disfluent utterances from spontaneous conversations, where the functions of both 'aehm' and 'also' overlap. Specifically, we aim to investigate whether listeners perceive utterances containing 'aehm' to be more or less fluent than utterances containing 'also'.

Studies on the perception of disfluent utterances have shown that filler particles are a reliable cue for children to infer speaker intention in advance of object labelling [6] and for adults, it has been shown that a certain amount of filler particles supports the recall information [7] and to help word recognition just as a silent pause would do too [8]. This is in line with the findings of Wepner et al. [9], who report that participants of a transcription experiment perform worse in utterances with syntactic disfluencies, but equally much benefit from silent or filled pauses. Others reported that filler particles are interpreted by listeners as signals of production difficulty [10].

***Corresponding author**: b.schuppler@tugraz.at.*

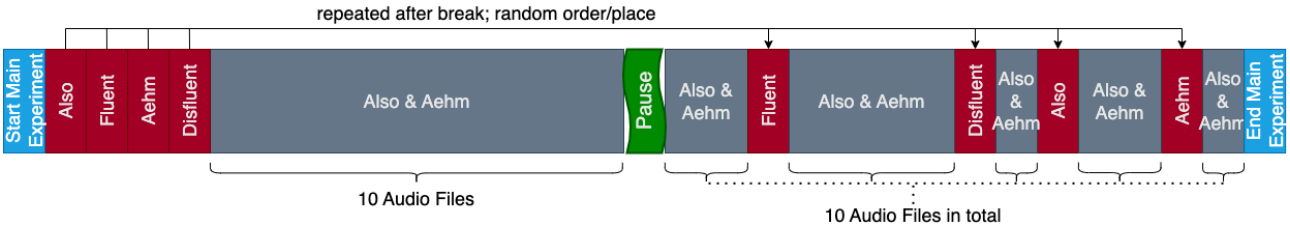

Figure 1: This chart shows the experimental design of the main perception experiment without the training phase.

More lexicalized items, however, are more likely to be interpreted as meaningful contributions to discourse structure rather than as disfluencies per se [11]. Thus there are indicators in the literature that not all fillers need to be equally *disturbing*: their effect on perceived fluency may depend on whether they are interpreted as cues to speaker planning or as part of discourse organization. We further hypothesize that the perception of fluency may further be affected by the degree of reduction and/or regional dialect of the speaker and by the complexity of the semantic content.

On the basis of a perception experiment using stimuli extracted from a corpus of casual, conversational speech [12], this paper aims to answer RQ1) Whether there is a difference in the perception of fluency between the generic filler 'aehm' and the filler 'also', which in other contexts may also appear as a lexical item; and 2) how perceived fluency relates to the perceived *clarity* in pronunciation and *understandability* of the semantic meaning of the stimulus.

2. Methods

2.1. Corpus and stimuli selection

For the perception experiment, we extracted stimuli from the conversational speech component of the GRASS corpus [12], [13], containing a total of 38 one-hour long open topic casual conversations between Austrian German speakers who knew each other well prior to the recording session (family members, couples, friends, colleagues). For GRASS high overlapping rates and disfluency rates have been reported, indicating a very lively speaking style. GRASS comes with manual orthographic and turn-taking annotations [14] and segmentations based on forced alignment [15]. For selection, cutting and noise-level adaptation of the stimuli, we used the python-based tool TNT Slicer [16].

We selected 60 stimuli for each filler particle 'aehm' and 'also' and additional audio examples for the training phase of the perception experiment. For 'aehm' we used all 46 stimuli used in our earlier transcription experiment [9] and selected 14 more from the same criteria (durational and structural) as described in their paper. Next we selected another 60 stimuli containing 'also', where we had to manually check and listen to assure that we include only stimuli where 'also' actually appears together with a disfluency (and not, e.g., in the meaning of 'thus'). In the selected stimuli, neither 'aehm' nor 'also' carried prosodic prominence. Furthermore, we searched for speech stimuli, which were completely fluent and others that were completely disfluent (i.e., containing many pauses, several filler-words, disfluent syntax). These in total 6 fluent and 6 disfluent stimuli were used as extreme reference points in the experiment. The total set of stimuli was approx. balanced with regard to having been produced by female and male speakers and stemmed in total from 34 different speakers from the GRASS corpus.

All stimuli were cut so that they had enough context to understand what was said and were approx. 2-3 phrases long. We only included audio free from background noise. The volume peak of all selected audio stimuli was normalized using SLICER to $-0.8dBFS$ to ensure that the overall volume of all stimuli is similar. A fade-in and a fade-out of 500 samples each was set to avoid clicks. The resulting stimuli had an average duration of $4.51s$ for the subset containing 'aehm' and $5.28s$ for those with 'also'.

2.2. Perception experiment

We recruited 30 participants (13 female, 17 male, 0 diverse) who were on average 24.8 years old and were native speakers of German (29 Austrian German, 1 South Tyrolean resident in Graz). None of the participants were diagnosed with hearing loss. The experiment was conducted individually in a sound-proof room using a PsychoPy-based program [17], which assigned participants with an ID, and guided them through all steps, including a tutorial, a training phase, the main experiment, and metadata collection. Participants could ask questions during scheduled pauses, and the experiment lasted approximately 20–30 minutes in total.

In the main experiment, each participant listened in total to 28 stimuli (10 'aehm' + 10 'also' + 4 stimuli played twice (incl. 1 'aehm', 1 'also', 1 'fluent', 1 'disfluent')), where for each stimulus they rated three questions on a scale between 1 to 8: 1) How clearly did the person pronounce the sentence? 2) How well did you understand what the person said? and 3) How fluently did the person speak? Participants listened twice to a stimulus before rating.

The session began with a tutorial (introducing the UI and extreme examples of fluent and disfluent speech), followed by a training phase with four audio examples to familiarize participants with the task. In the main experiment (cf. Fig. 1), audio examples included both filler-containing and fluent/disfluent speech, presented

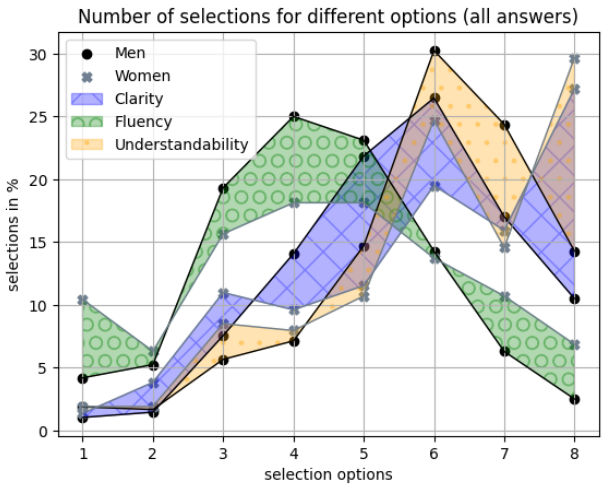

Figure 2: This Plot shows the selections in % per selection option. The results of the different genders are included.

in randomized order. The experiment was split into two halves of 14 examples each, with the first four examples of the first half repeated in the second half to test intra-participant consistency. By using session numbers (1–6), we determined specific audio selections to ensure a balanced stimuli presentation across participants. After completing the experiment, participants provided information on age, gender, hearing ability, and place of residence. This metadata allowed to see potential correlations between participant characteristics and perception results (e.g., for strong outliers).

3. Results and Discussion

3.1. General observations on the ratings

Figure 2 shows the number of selections per selection option. In every single experiment four stimuli get repeated and 20 stimuli are just occurring once in the experiment. In this figure, both the repeated files and the files which occur just once are included. The lines of different colors show the results of the different questions posed in the experiment. The blue line describes the clarity ratings, the orange line describes the understandability ratings and the green line describes the fluency ratings. The y-axis shows the relative number of selections. So, if you sum up the values of the ratings of a specific question posed in the experiment (e.g. clarity ratings) it amounts to 100%. The plot shows that the clarity and understandability ratings are in a similar range. There is a peak at the selection option 6. The clarity ratings are slightly lower than the understandability ratings, indicating that the content of what was said was in general well understood. There is a bigger difference between the clarity and understandability ratings and the fluency ratings. The fluency ratings have a peak at selection option four and the curve for the fluency ratings is shifted to the left. How fluency ratings correlate concretely for each stimulus with the clarity and understandability ratings is analyzed in section 3.2. What we also see from Figure 2 is that men gave overall higher ratings for the stimuli than women, in all three rating-categories.

3.2. Quantiative stastistical analysis

In order to investigate how perceived fluency relates to the perceived clarity in pronunciation and understandability of the semantic meaning of the stimuli, we fitted linear mixed-effects models in R using the function `lmer()`. We created three different models with the dependent variables being the ratings of clarity, understandability and fluency. As fixed effects, we included metadata outputs of the participants, the gender of the speakers of the stimuli, the duration of the stimuli, the speech type ('also' or 'aehm' stimulus) and the ratings, which were not already used as dependent variable. For example, when the fluency ratings were used as dependent variable, the clarity ratings and the understandability ratings were included as fixed effects. Furthermore, we added all possible two-way interaction terms. Higher-order interactions were not included due to the limited number of tokens and concerns about model complexity and interpretability. As random intercept effects we included the participant number and the speaker_id (i.e., the speaker of the stimulus). We then applied a best-fit model selection procedure via manual backward stepwise reduction, using Akaike Information Criterion (AIC) as the criterion for retaining independent variables and interactions (e.g., [18], [19]). Random effects were only kept if they significantly improved model fit, as determined by ANOVA testing. In the remainder of this paper, we report only the significant fixed effects and interactions, where with *highly significant* we refer to p-values < 0.001 (***), with *significant* to p-values < 0.01 (**) and < 0.05 (*), and with *marginally significant* to those between $0.05 - 0.1$ (.).

3.2.1. RQ1: Differences between 'also' and 'aehm'

Figure 3 shows a violin plot of the ratings for 'also' and 'aehm' separately. In this figure just the stimuli which occur once in the perception experiments are included. On the x-axis the clarity ratings, the understandability ratings and the fluency ratings can be seen and on the y-axis the selection options are listed. The orange region describes 'also' stimuli and the blue region describes 'aehm' stimuli. The deflection from zero in the violin plot describes how often a single selection option was chosen. So, for example, if you look at the orange region ('also' stimuli) and at the understandability ratings you can see that the selection option six was chosen the most. The three subplots show a tendency that the 'aehm' stimuli got better ratings than the 'also' stimuli.

To test whether these tendencies are also statistically significant, we fitted three linear mixed-effects regression models with fluency_rating, clarity_rating and understandability_rating as the three dependent variables, all of the following form:

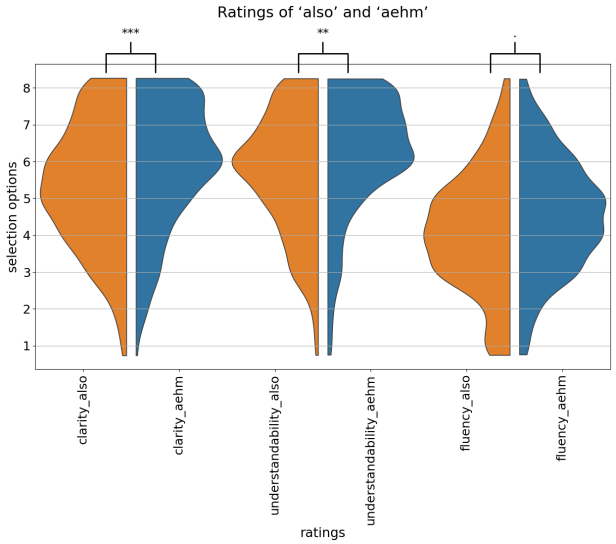

Figure 3: Violin plots showing the ratings for 'also' and 'aehm' stimuli separately ('also' = orange, 'aehm' = blue). The stars above the brackets indicate the significance levels resulting from the lmer models (cf. Section 3.2.1).

```
fluency_rating ~ (gender_speaker +
gender_participant)^2 + speech_type +
duration_file + (1 | id_speaker) + (1 |
participant),
```

where speech_type is a binary variable with the two values 'aehm' and 'also'. We find that fluency_rating is only marginally significantly lower for 'also' than for 'aehm' ($est = -0.23, t = -1.90, p = 0.06$), the duration of the stimulus, however, resulted to be highly significant ($est = -0.18, t = -4.66, p < 0.001$). These results indicate that perceived fluency is strongly correlating with the duration of the stimulus but not with the particle type itself.

In the model with clarity_rating as dependent variable there was only one significant effect - speech_type: Stimuli containing 'also' were highly significanly lower rated in clarity than stimuli containing 'aehm' ($est = -0.59, t = -4.11, p < 0.001$). The same picture was observed for the model with understandability_rating as dependent variable. Only one factor was significant – speech_type ($est = -0.45, t = -3.20, p < 0.01$). These results indicate that whereas perceived fluency tends to be only marginally effected by the particle type, stimuli containing 'aehm' are significantly higher rated in clarity and understandability than stimuli containing 'also'.

3.2.2. RQ2: Effects of clarity and understandability on perceived fluency

In order to analyze which factors affected the perceived fluency, we fitted a linear mixed-effects model predicting the fluency ratings. After backward stepwise model selection, the final linear mixed-effects model included:

```
fluency_rating ~ (gender_speaker +
gender_participant)^2 + (understandability_rating
+ clarity_rating)^2 + duration_file + (1 |
id_speaker) + (1 | participant)
```

Table 1 shows the full model output. It can be seen that the duration of the stimulus highly significantly affects the fluency rating: stimuli tend to be perceived more fluent in shorter audio files. Next, the interaction term between the gender of the speaker (of the stimulus) and the gender of the participant is marginally significant, whereas their individual fixed effects were not significant. Looking into the data, we see that ratings for female speakers rated by female listening participants were higher than other gender groupings of speaker and listening participant.

Finally, also the interaction term between the understandability rating and the clarity rating is highly significant. To better understand this interaction term, we plotted the understandability rating and the clarity rating in relation to the fluency rating (cf. 4). The figure shows that for $clarity\_rating = 1$ the fluency rating stays relatively constant, independent of how well the understandability of the stimulus was. With increasing clarity of the utterance, also the understandability affects the perceived fluency more. Also interesting, even when perceived pronunciation of stimuli was very clear (clarity_rating = 8), they also contained disfluent and badly understandable content.

4. Conclusion

The analysis of this perception experiment with 30 participants provides insights into the relationship between perceived clarity, understandability, and fluency of casual, conversational speech. Regarding RQ1, the results show that there is a perceptual difference when the filler word 'aehm' or the particle 'also' occurs in disfluent stimuli. The ratings show a tendency for speech stimuli containing 'aehm' to receive significantly higher evaluations in terms of clarity and understandability than stimuli containing 'also', for perceived fluency the effect was only marginally significant and also the descriptive statistics via violin plots show that when it comes to perceived fluency, the stimuli containing 'aehm' and 'also' show similar patterns. Interestingly, and in line with intuition, the duration of the stimuli was only a significant factor for perceived fluency, but not for perceived clarity and understandability.

Previous research has shown that filled pauses such as *um* and *uh* are not merely errors but can function as communicative signals that support speech processing and help listeners anticipate upcoming information. For example, [4] showed in a corpus study that speakers systematically use *uh* and *um* to signal upcoming delays in speech production, with *uh* typically preceding shorter delays and *um* preceding longer delays. This corpus-based evidence suggests that filled pauses function as conventionalized signals in spoken interaction. The relatively positive evaluation of stimuli containing 'aehm' in terms of clarity and understandability the present perception experiment is compatible with this interpretation, as listeners may treat such fillers as expected signals of speech planning rather than as disruptive elements.

Based on perception experiments, [20] demonstrated that hearing a filled pause such as *uh* can facilitate listeners' recognition of upcoming words, indicating that listeners use these markers to adjust their expectations during real-time speech processing. Another aspect to consider in the present study is that the item "also" differs from "aehm" in that it can function both as a lexical item with propositional meaning and as a discourse marker in spontaneous speech. Research on discourse markers with lexical origins, such as *you know* or *I mean*, shows that listeners may interpret them differently from purely non-lexical filled pauses [21]. This may influence how listeners evaluate

Table 1: This table shows the model estimates of the model with fluency_rating as dependent variable.

	Est.	Std. Error	df	t value	$Pr(>\|t\|)$	Signif.
(Intercept)	4.03	0.54	540.66	7.43	<0.001	***
gender_speakerweiblich	0.17	0.18	46.60	0.97	0.34	
gender_participantweiblich	-0.21	0.26	35.14	-0.81	0.42	
understandability_rating	-0.03	0.09	568.79	-0.33	0.74	
clarity_rating	-0.16	0.10	564.85	-1.68	0.09	.
duration_file	-0.17	0.03	539.07	-5.10	<0.001	***
gender_speakerweiblich:gender_participantweiblich	0.35	0.18	536.87	1.93	0.05	.
understandability_rating:clarity_rating	0.06	0.02	569.51	3.91	<0.001	***

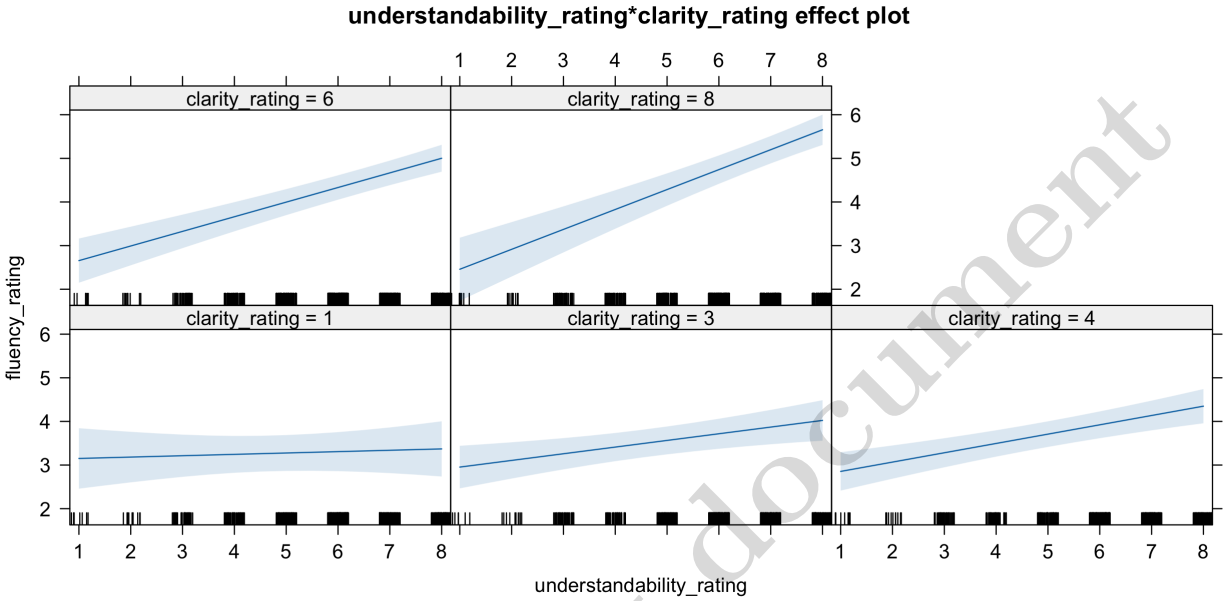

Figure 4: This figure shows the effect plot of the fluency_rating as a function of the interaction between understandability_rating and clarity_rating.

such items in perception tasks, and that its' occurrence may appear less expected in the given speech context than a purely hesitation-based filler such as "aehm".

With respect to RQ2, we observed that the ratings for perceived *fluency* are significantly affected by the ratings for the semantic *understandability* of the utterance and the *clarity* of pronunciation, and that this effect is independent of whether stimuli contained 'aehm' or 'also': fluency ratings are stable at low clarity levels regardless of understandability, but as clarity increases, understandability has a stronger positive correlation with fluency, with some highly clear utterances still being perceived as disfluent and poorly understandable. Previous research has shown that speech disfluencies can influence both comprehension and the processing of spoken language. For instance, [22] conducted experimental studies showing that the presence of hesitations such as *uh* or *er* can guide the listeners' expectations during sentence processing and may even facilitate comprehension by signaling upcoming difficulty in speech production. This is in line with observations that listeners are able to compensate for disfluencies in spontaneous speech and incorporate them into their interpretation of the speaker's message [23].

Our findings do not only increase our knowledge about human speech processing, but may also inform speech processing models: for instance, tools that automatically detect or modify utterances to improve comprehensibility, that edit conversational recordings or speech-to-text systems. They may benefit from considering that different types of filler words are perceived differently depending on their context.

5. Acknowledgements

This research was funded in part by the Austrian Science Fund (FWF) [10.55776/P32700]. We also thank Lucas Eckert for support with SLICER and stimuli selection.

References

[1] J. Ginzburg, *The Interactive Stance: Meaning for Conversation*. Oxford University Press, 2012.

[2] M. Wiltschko, *The Grammar of Interactional Language*. Cambridge University Press, 2021.

[3] M. Belz, "Die Phonetik von äh und ähm," Ph.D. dissertation, Humboldt-Universität zu Berlin, 2021.

[4] H. H. Clark and J. E. Fox Tree, "Using uh and um in spontaneous speaking," *Cognition*, vol. 84, pp. 73–111, 2002.

[5] C. Diskin, "The use of the discourse-pragmatic marker 'like' by native and non-native speakers of English in Ireland," *Journal of Pragmatics*, vol. 120, pp. 144–157, 2017.

[6] 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.

[7] 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.

[8] 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, May 2011. DOI: 10.1371/journal.pone.0019792.

[9] S. Wepner, L. Eckert, G. Kubin, and B. Schuppler, "What the Filler? Both ASR Systems and Humans Struggle More With Other Kinds of Disfluencies Than With Filler Particles," in *Proceedings of Interspeech 2025*, 2025, pp. 2325–2329.

[10] J. E. Arnold, M. K. Tanenhaus, R. J. Altmann, and M. Fagnano, "The old and thee, uh, new: Disfluency and reference resolution," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 33, no. 5, pp. 914–930, 2007. DOI: 10.1037/0278-7393.33.5.914.

[11] M. Corley and O. W. Stewart, "Hesitation disfluencies in spontaneous speech: The meaning of um," *Language and Linguistics Compass*, vol. 2, no. 4, pp. 589–602, 2008.

[12] B. Schuppler, M. Hagmüller, and A. Zahrer, "A corpus of read and conversational Austrian German," *Speech Communication*, vol. 94, pp. 62–74, 2017.

[13] B. Schuppler, M. Hagmüller, J. A. Morales-Cordovilla, and H. Pessentheiner, "GRASS: The Graz corpus of Read And Spontaneous Speech," in *Proc. LREC*, 2014, pp. 1465–1470.

[14] A. Kelterer and B. Schuppler, *Turn-taking annotation for quantitative and qualitative analyses of conversation*, 2025. DOI: https://doi.org/10.48550/arXiv.2504.09980. arXiv: 2504.09980 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2504.09980.

[15] J. Linke, S. Wepner, G. Kubin, and B. Schuppler, "Using Kaldi for Automatic Speech Recognition of Conversational Austrian German," in *arXiv:2301.06475*, 2023.

[16] L. Eckert, S. Wepner, and B. Schuppler, "Slicer – A Tool for Efficient Stimuli Extraction from Large Speech Corpora," in *Proceedings of Forum Acusticum Euronoise 2025*, 2025.

[17] J. Peirce et al., "PsychoPy2: Experiments in behavior made easy," *Behavior Research Methods*, vol. 51, pp. 195–203, 2019.

[18] R. H. Baayen, D. J. Davidson, and D. M. Bates, "Mixed-effects modeling with crossed random effects for subjects and items," *Journal of Memory and Language*, vol. 59, no. 4, pp. 390–412, 2008.

[19] 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.

[20] J. E. Fox Tree, "Listeners' uses of um and uh in speech comprehension," *Memory & Cognition*, vol. 29, no. 2, pp. 320–326, 2001.

[21] J. E. Fox Tree and J. C. Schrock, "Basic meanings of you know and I mean," *Journal of Pragmatics*, vol. 34, no. 6, pp. 727–747, 2002.

[22] M. Corley, L. J. MacGregor, and D. I. Donaldson, "It's the way that you, er, say it: Hesitations in speech affect language comprehension," *Cognition*, vol. 105, no. 3, pp. 658–668, 2007.

[23] S. E. Brennan and M. F. Schober, "How listeners compensate for disfluencies in spontaneous speech," *Journal of Memory and Language*, vol. 44, no. 2, pp. 274–296, 2001.