languages

MDPI

Article

Disfluencies Revisited—Are They Speaker-Specific?

Angelika Braun [1,*], Nathalie Elsässer [1] and Lea Willems [2]

[1] Phonetics, University of Trier, 54296 Trier, Germany
[2] Mathematics, University of Trier, 54296 Trier, Germany
* Correspondence: brauna@uni-trier.de

Abstract: The forensic application of phonetics relies on individuality in speech. In the forensic domain, individual patterns of verbal and paraverbal behavior are of interest which are readily available, measurable, consistent, and robust to disguise and to telephone transmission. This contribution is written from the perspective of the forensic phonetic practitioner and seeks to establish a more comprehensive concept of disfluency than previous studies have. A taxonomy of possible variables forming part of what can be termed disfluency behavior is outlined. It includes the "classical" fillers, but extends well beyond these, covering, among others, additional types of fillers as well as prolongations, but also the way in which fillers are combined with pauses. In the empirical section, the materials collected for an earlier study are re-examined and subjected to two different statistical procedures in an attempt to approach the issue of individuality. Recordings consist of several minutes of spontaneous speech by eight speakers on three different occasions. Beyond the established set of hesitation markers, additional aspects of disfluency behavior which fulfill the criteria outlined above are included in the analysis. The proportion of various types of disfluency markers is determined. Both statistical approaches suggest that these speakers can be distinguished at a level far above chance using the disfluency data. At the same time, the results show that it is difficult to pin down a single measure which characterizes the disfluency behavior of an individual speaker. The forensic implications of these findings are discussed.

Keywords: forensic voice comparison; hesitations; fillers; lengthening; paraverbal behavior

Citation: Braun, Angelika, Nathalie Elsässer, and Lea Willems. 2023. Disfluencies Revisited—Are They Speaker-Specific? *Languages* 8: 155. https://doi.org/10.3390/languages8030155

Academic Editors: Jürgen Trouvain and Bernd Möbius

Received: 12 January 2023
Revised: 2 June 2023
Accepted: 7 June 2023
Published: 26 June 2023

Part A: Disfluencies and Forensic Voice Comparison

1. Introduction

Whenever speakers engage in spontaneous conversation, disfluencies, i.e., disruptions of the speech flow, are bound to occur. These have been studied and explained from two different angles which are by no means mutually exclusive; disfluencies may be regarded as symptoms or signals in the sense of Karl Bühler's organon model (Bühler 1934). According to Bühler, there are three ways in which linguistic signs function: they serve as symbols for concepts and objects from the outside world, as symptoms marking speaker state, and as signals indicating the speaker's intent. Since about the 1960s, fillers in particular have been looked at primarily as indications of verbal planning and self-monitoring on the part of the speaker (cf., e.g., Goldman-Eisler 1961; Shriberg 2001; Maclay and Osgood 1959; Finlayson and Corley 2012) and thus as examples of the symptom function of speech. Levelt (1989, p. 484) summarizes that "[t]he interjection 'er' apparently signals that at the moment when trouble is detected, the source of trouble is still actual or quite recent. But otherwise, 'er' doesn't seem to mean anything. It is a symptom, not a sign." This view is supported by studies such as that by Oviatt (1995), who found that the number of fillers increases with the planning load of utterances, and that there are more disfluencies at the beginning of an utterance, which is when the number of choices to be made is presumably larger than towards the end. Oviatt (1995) also established that long utterances are more likely to be disfluent than shorter ones (see also Eklund 2004).

The lexical status of fillers has been subject to debate.[1] In this context, Clark and Fox Tree (2002) argue vehemently that fillers are normal words (interjections), which would make them signals in the sense of Bühler (1934) because "[t]hey have conventional forms and meanings, conform to the notion of word syntactically and prosodically, and contrast with another signal of delay, the process of prolongation" (p. 105). On the other hand, O'Connell and Kowal (2005) find evidence against *uh* and *um* being interjections, because their functions in speech are fundamentally different. Lickley and Bard (1996) consider fillers to be noise rather than words. Corley and Stewart (2008), based on an in-depth literature search, summarize that "[t]here is no conclusive evidence that fillers are words" (p. 600).

A key issue in this discussion is to what extent speakers have control over the use of fillers. Clark and Fox Tree (2002) present evidence that speakers consciously distinguish between *uh* and *um* in a systematical manner: "[. . .] speakers have selective control over uh and um. They are quite accurate in projecting minor vs. major delays, inserting uh before minor ones and um before major ones. [. . .] speakers can reduce or eliminate their use of fillers when the circumstances require it." (p. 99). Corley and Stewart (2008), on the other hand, find no proof that hesitations are consciously used by speakers to signal delay. They consider alternative explanations for the findings cited by Clark and Fox Tree: "It could be the case that speakers produce fillers quite unintentionally (e.g., as a by-product of delay), but at predictable junctures, and listeners are sensitive to these accidental patterns of occurrence" (p. 598). This is a point which bears relevance to using this parameter in the forensic domain: anything which is subject to active influence on the part of the speaker is prone to voice disguise.

Only fairly recently has it been shown that the "benefit" provided by hesitations is not confined to the speaker, but that it is also an important aid for speech perception. The time gained on the part of the speaker through hesitating is equaled by the time gained on the part of the listener to process the message (Corley and Hartsuiker 2003; Erard 2007; Tannen 1985, p. 99). Target words were recognized faster with fillers present than when they had been edited out (Fox Tree 2001). Blau (1991) demonstrated that a monologue containing hesitations was considerably easier to comprehend for listeners with English as a second language than one from which the hesitations had been removed. Corley et al. (2007) carried out an EEG study and showed that "[. . .] disfluency clearly affects the processing of language" (p. 666). Beyond this immediate effect, participants were better at recognizing words preceded by a disfluency even after a time delay.

2. The Forensic Angle

This contribution is written from a forensic practitioner's perspective. One of the common tasks in the forensic application of phonetics involves the comparison of one or more speech samples originating from a perpetrator (questioned recording) to one or more reference samples originating from the suspect(s). The aim is to provide the triers of fact with a probabilistic conclusion expressing the likelihood of the two sets originating from one and the same speaker.

Features to be used in forensic voice comparison have to meet a number of criteria. These include, inter alia, being readily available, measurable, consistent, and robust to disguise as well as to mismatch conditions (Wolf 1972). The latter present a major problem, especially to any (semi-)automatic approach to speaker recognition. Channel mismatch is probably the most obvious one. Whether the method of comparison is GMM-UBM (Gaussian Mixture Model—Universal Mackground model) or DNN (Deep Neural Network)—if the questioned recording involves telephone transmission while the reference recording consists of a police interview recorded in a reverberant interview room, there is a danger that the algorithms will identify channels rather than voices. However, situational mismatch is just as problematic. If the questioned recording contains speech in noise, whereas the reference recording was made in a quiet environment, automatic procedures show reduced performance levels. This is one reason why "traditional" auditory and acoustic

phonetic analyses still play a major role in forensic voice comparison, even though various automatic systems have been developed since the turn of the century (Braun 1998, 2021; Meuwly 2001) and have been on the market for many years.[2] The auditory–acoustic approach comprises analyses of voice fundamental frequency and its derivatives; regional or foreign accent, individual sound production including mispronunciations, grammar, lexical habits, etc. (Jessen 2008, 2012; Braun 2021). The search for valid and reliable parameters is ongoing. Research in this area relies heavily on the concept of individuality in speech. Features exhibiting little intra-speaker variability and large between-speaker variability are most suitable for use in forensic casework. Unlike in most "ordinary" research, which is usually carried out with the aim of establishing statistically generalizable results in mind and where individual behavior is considered "noise", precisely these individual patterns of verbal, paraverbal, and even nonverbal behavior are of interest in the forensic domain.

One important feature bundle is what the first author called hesitation behavior earlier on.[3] So far, elements of this behavioral pattern have been considered in isolation, but this has not rendered highly individual results. For instance, the number of disfluencies per 100 words alone has been shown not to distinguish between speakers, because there are within-speaker differences according to the type of task (Bortfeld et al. 2001; Harrington et al. 2021). For instance, speaking on the telephone produces more fillers than face-to-face conversations, dialogical speech is more disfluent than monological (Oviatt 1995), and human-directed speech is more disfluent than machine-directed speech (Shriberg 1996). At the same time, there are considerable similarities between speakers, suggesting that there is a generally applicable mechanism behind hesitating.

A dedicated forensic approach to disfluencies was developed by McDougall and Duckworth (2017, 2018) and McDougall et al. (2019). They list a number of parameters which describe the behavioral profile of a given speaker for use in forensic casework. This parameter set was tested on 20 speakers from the DyViS corpus with promising results. Their TOFFA (Taxonomy Of Fluency Features for Forensic Analysis) framework is more comprehensive than any other, but it still falls short of being exhaustive, and the intra-speaker consistency of the features is assumed but not tested. For example, they do not specifically cover verbal fillers nor the nasal filler, and neither do they consider the connection between filler and surrounding utterance. The study by Hughes et al. (2016) also has a forensic focus. Their approach is confined to studying various aspects of the formant dynamics of *um* in a likelihood-ratio-based format. They find that that provides relevant information about voice identity, which underlines the importance of including the spectral characteristics of fillers in the analysis.

Another key aspect to be borne in mind in relation to forensic applications is the language specificity of disfluencies, which is suggested by various authors (see, e.g., Candéa 2000; Candea et al. 2005; Eklund 2000; Eklund and Shriberg 1998). This means that forensic practitioners should refrain from comparing samples differing in language with respect to disfluencies.

The key questions to be answered by the present research are thus as follows:

Are there speaker characteristic features in the disfluency behavior which have so far not been exploited?

Are speakers at all consistent in their disfluency behavior?

Are there features which can reliably serve to distinguish between speakers?

3. Disfluencies and Individuality

Besides the specifically forensic contributions mentioned above, there are observations in various studies pointing to the fact that patterns in the use of disfluency markers may be individual (Maclay and Osgood 1959; Goldman-Eisler 1961; Blankenship and Kay 1964; Henderson et al. 1966; Goldman-Eisler 1968; Butcher 1973; Duez 1982, 2001; Kowal 1991; Shriberg 1994, 2001; Belz 2021; Clark and Fox Tree 2002, pp. 97–98; Eklund 2001, p. 6; Fant et al. 2003). This is consistent with the notion that disfluency behavior reflects the cognitive planning process of a specific individual. If individual patterns were to be

established, this would point to individual cognitive strategies. In the early literature on disfluencies, individuality is stressed much more than in more recent publications. For instance, Maclay and Osgood (1959, p. 38) write: "The relative 'preference' for hesitation phenomena of different types may be considered an aspect of individual style in speakers." Duez (1982, p. 17), in a study on French politicians, observes individual preferences for the type of hesitation. Goldman-Eisler (1961, p. 23) regards hesitation phenomena as "[. . .] a discriminating factor in different individuals." Fant et al. (2003, p. 194) describe a large degree of individuality with respect to pause length: "The individual variations are considerable and greater than expected." Eklund (2001) finds individuals to whom the rule that filled pauses are more common than prolongations does not apply. However, none of these studies report data for individual speakers.

In a previous study, one of the present authors found indications that hesitation behavior is indeed speaker-specific and thus lends itself to being considered in auditory–acoustic voice comparison (Braun 2020). The present contribution seeks to establish a more comprehensive concept of disfluency than has been presented in previous studies. A descriptive framework of disfluency markers is proposed and tested on a small set of recordings. It makes use of parameters which have been studied for decades, such as fillers, repetitions and false starts, repairs, and restarts etc., but it also proposes new elements. Secondly, the data which had been collected for the previous study is analyzed in more detail. In order to shed light on speaker individuality, results are presented separately for each speaker and by session. This unusual format was chosen because it is reminiscent of the forensic setting.

4. A Taxonomy of Disfluencies

The use of terms with regard to disfluencies is far from unanimous (Lickley 2015). It is therefore essential to explain the working definitions used in the present contribution. This section is not intended as a comprehensive theory of disfluencies but focuses instead on topics which are potentially relevant to the forensic setting and have not been studied widely in previous research. *Disfluency* (with an <i> as opposed to a <y>) is used as a cover term for any disruption of the speech flow. Disfluencies may occur as *hesitations* or *repairs*. *Hesitations* are disfluencies allowing the speaker to "gain time". *Repairs* are disfluencies involving a revision of what has already been said. *Disfluency behavior* is used to denote the pattern which an individual speaker exhibits in conjunction with hesitations and repairs. *Hesitation patterns* are the systematic and potentially idiosyncratic constellations of hesitation markers as a result of hesitation behavior. *Hesitation marker* is used as a cover term for any elements which signal hesitation, i.e., pauses, fillers, prolongations, repetitions, voice quality, nonverbal vocalizations, etc. The term *filler* encompasses the "classical" hesitation markers of the *uh*, *um* type, as well as the *verbal fillers*, which, following Stenström (2012), display an "unusual" use of what are otherwise lexical items as fillers. Figures 1 and 2 show the hierarchy of the terms as used in this contribution.

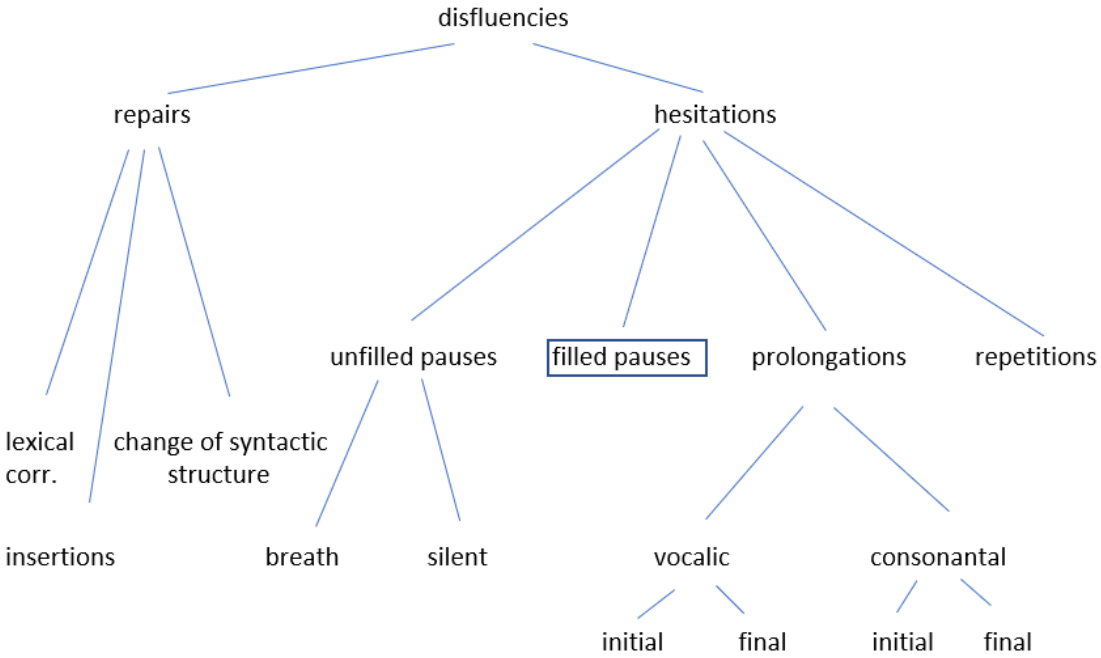

Figure 1. Taxonomy of disfluencies as used in this contribution.

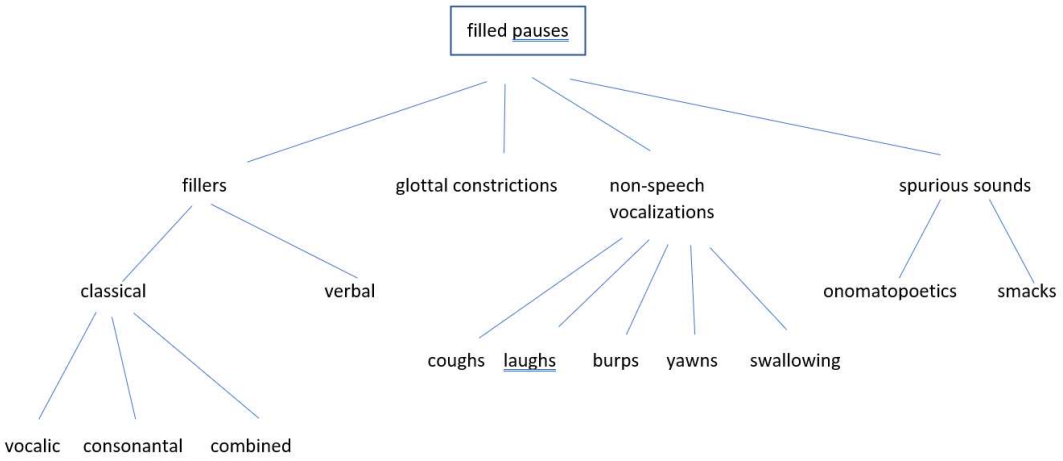

Figure 2. Taxonomy of filled pauses as used in this contribution.

4.1. Hesitations

4.1.1. Pauses

A major distinction between hesitations can be drawn according to whether or not they contain vocalizations. They may be characterized by the absence of audible articulatory activity or by the presence of some kind of vocalization.

4.1.1.1. Unfilled Pauses

Unfilled pauses are pauses without vocalization, i.e., sound generated in the larynx and/or the vocal tract. In other words, they are truly silent or serve the purpose of breathing. The role of breath pauses is somewhat ambiguous. On the one hand, they contain signal, but on the other hand, the origin of this signal is neither the larynx nor the vocal tract. In other words, there is no "phonetic activity", as Trouvain et al. (2016) put it. Therefore, we concur with those researchers and categorize them as unfilled pauses.

For an analysis of hesitation behavior, the frequency of occurrence and duration of unfilled pauses is relevant.

4.1.1.1.1. Breath Pauses

Breath pauses primarily serve a physiological need, but they may be used for planning purposes at the same time. From a forensic perspective, the issue of individuality in breathing is of interest. There are several studies which demonstrate that breathing patterns are highly individual (Dejours et al. 1961; Shea et al. 1987; Shea and Guz 1992; Benchetrit et al. 1989; Benchetrit 2000; Eisele et al. 1992; Trouvain et al. 2019). Dejours et al. (1961) coined the term "personnalité ventilatoire" to describe the individual dynamics of breathing. Shea and colleagues (Shea et al. 1987; Shea and Guz 1992) confirmed those findings, which were established based on the physiological examination of quiet breathing, and demonstrated their validity for the deepest type of non-REM sleep (S4 sleep) as well. They attempted to define the cause for individuality in breathing and posited that the size and structure of the airways and lungs, as well as the mechanics of breathing, play a key role (Shea and Guz 1992, p. 287). This is a potentially important finding in the context of forensic analysis, because it demonstrates that between-subject variability is likely to exceed within-subject variability. However, the results cited so far were established through physiological measurements, and they refer to the frequency and depth of inhalations alone. The latter are obviously not available in the forensic environment. With the forensic application in mind, one may ask how often the speaker inhales and if the preferred pathway is through the nose, the mouth, or both, and in which sequence. Nasal inhalation followed by oral inhalation is often accompanied by a click sound at the transition, whereas the same is not true for the opposite sequence. Kienast and Glitza (2003) used auditory and acoustic phonetic methods to analyze speech breathing. They studied the frequency of speech breathing as well as the preferred pathway. They concluded that the pathway of air (nose and/or mouth) and the acoustic structure of the breathing noise are the best "candidates" for characterizing the individual speaker. This shows that the physiological findings are mirrored by the acoustics of speech breathing. Lauf (2001) studied the duration, frequency, and spectral composition of breath pauses in read speech and found that speakers fall into different categories (e.g., with respect to the duration of inhalation), but the parameters she studied were not suitable to distinguish individuals.

4.1.1.2. Filled Pauses

Filled pauses are pauses containing vocalized sound which may or may not be preceded and/or followed by a period of silence. They can be subdivided into various categories, some of which are discussed in detail below. The most frequent ones are pauses containing fillers in the traditional sense, but they also include non-speech vocalizations and glottal constrictions. Finally, pauses which are filled by a variety of spurious sounds fall into this category.

4.1.1.2.1. Classical Fillers

The number of fillers which are covered in most of the literature is n = 2, namely *uh* and *um* (Clark and Fox Tree 2002; Kjellmer 2003; Corley and Stewart 2008; Corley et al. 2007; Schegloff 2010; Belz 2021, to name only a few). There are claims that these two are universal (Clark and Fox Tree 2002), although their phonetic form may differ according to language and possibly also dialect. There is also considerable variation in the orthographic representation of these fillers (for an overview with respect to English and German see Belz (2021, pp. 12–14); a more comprehensive list can be found in Clark and Fox Tree (2002, pp. 92–93). For reasons of simplicity, *uh*/*um* and *äh*/*ähm* are used for English and German, respectively, in this contribution.

If one looks at data sets of conversational speech, it is quite clear that the two items discussed so far are by no means the only fillers that may occur. There are many more hesitation markers beyond the "classical" set, which may be much more idiosyncratic than the frequently studied fillers *äh* and *ähm*. Perhaps the filler most frequently encountered beyond those two in German is *mh*. In our data, there are speakers who use *mh* more often than *äh*. Considering this distribution, the present authors find it difficult to understand why the nasal filler is not routinely included in the analysis and consider it on equal

footing with *äh* and *ähm*. Belz (2021) takes it into account but assigns a marginal role to it. Incidentally, *mh* was regularly included in some early studies (e.g., Maclay and Osgood 1959; Blankenship and Kay 1964), but it somehow "got lost" thereafter. Like the other two, *mh* may or may not be preceded by a glottal stop ([ʔmmm] vs. [mmm]).[4] Here is an example, taken from our recordings:

(1) *Als ich von dem Sonnenhof nach Steinbach zurück lief, mh, sah ich einen Hubschrauber über mir*. 'When I returned to Steinbach from the Sonnenhof, *mh*, I saw a helicopter above me'. (S #1)[5]

4.1.1.2.2. Verbal Fillers

There is another group of fillers which is addressed very rarely, and if so, it is addressed in a controversial manner. Clark and Fox Tree (2002) call them "collateral signals" of which the speaker is not aware. Stenström (2012) talks about "verbal fillers". This seems to be a more suitable term and is therefore used in the present contribution.

Verbal fillers are multifunctional lexical items which "can also have various discourse, pragmatic and interactional functions" (Stenström 2012, p. 540). When used as fillers, "[. . .] they add nothing to the propositional content of an utterance, only to the pragmatic content [. . .]" (Stenström 2012, p. 540).

We discuss the most frequent—and therefore, the most relevant—verbal fillers in more detail. In German, *und* and *ja* are prime examples. Both may adopt different roles, and the categorization of *ja* in particular as a part of speech is controversial. *Und* is a connective in the first place, of course, but it may be used completely devoid of its lexical meaning, as is demonstrated by an example from our data:

(2) [. . .] *und es es riecht nicht, ähm ja und und äh es es soll ja auch ich mein ich weiß es ja selber nich* [. . .] '[. . .] and it it doesn't smell, *um* well, and and *uh* it it should also I mean I don't know myself [. . .]'. (S #7)

Discussing the full range of use of *ja* in German is well beyond the scope of this contribution,[6] but the aspects which are most relevant to the classification made here are mentioned. In the first place, *ja* signals affirmation in response to a question. In this context, it is stressed.

(3) *Kommst Du mit?—Ja*. 'Are you coming along?—Yes.'

Beyond that, *ja* is a modal particle serving different purposes depending on the lexical stress. When stressed, it is used as an intensifier and conveys a strong urge, possibly even implying a threat on the part of the speaker, as in the following example:

(4) *Pass auf*! 'Be careful!' vs.
(5) *Pass **ja** auf*! 'You had better be careful' or 'Do be careful'.

When unstressed, *ja* may be used to indicate that the speaker is stating the obvious and expects the listener to be privy to that information. Note the difference between:

(6) *Bayern München hat das Spiel gewonnen.* 'Bayern Munich won the game' [I am telling you] and
(7) *Bayern München hat ja das Spiel gewonnen.* 'Bayern Munich won the game' [as you know].

Another use of *ja* signals speaker attitude, specifically verbal irony, as is evident from the following example.

(8) *Das kann ja heiter werden!* 'This is really going to be fun, ha'.

Yet another variant of the German *ja* is its use as an interjection. In this capacity, *ja* may be replaced by *tja*.

(9) *(T)ja, was soll man dazu sagen?* 'What can you say to this?'

But *ja* may also be used as a question tag, replacing *nicht*, *nicht wahr*, etc.:[7]

(10) *Ich war von dem langen Flug völlig übermüdet, ja, und habe mein Auto im Parkhaus nicht mehr gefunden*. 'I was very exhausted from the long flight, you know, I could not find my car in the parking garage.'

(11) [. . .] *und dann kommt noch etwas Butter dazu, ja* [. . .] '[. . .] and then you add a little butter, okay?' (S #1)

Finally, *ja* may serve as a filler, as in (12) and (13) below:

(12) [. . .] *wenn man, ja, die Grenze erreicht* [. . .] '[. . .] once you, *uh*, get to the border [. . .]'. (S #1)

In our material, it was occasionally replaced by *pja*. This demonstrates once again that there is no propositional content left in this usage of *ja*.

(13) [. . .] *wenn da geraucht wird, ähm muss man ja nicht hingehen, wenn man das nicht äh dulden möchte, und, ja, wie gesagt ähm, ich bin halt gegen das Rauchen* [. . .] '[. . .] if people are smoking there, *uh*, you don't have to go there if you are not willing to *uh* tolerate that, and, *uh*, as I said, *um*, I am opposed to smoking [. . .]' (S #4)

This last example has two instances of *ja* side by side: The first is of the same kind as (7) above, whereas the second is a filler.

The use of *ja* as a filler differs from its use as a modal particle in several respects. First of all, this kind of *ja* is usually preceded or followed by a pause. Just like in most "common" fillers, it has a lower F0 than the immediate vicinity. Furthermore, the intonational phrase, but not necessarily the syntactic phrase, is interrupted. In its capacity as a hesitation marker, *ja* cannot be replaced by other expressions signaling consent, such *as freilich*, *jawohl*, *okay*, or *japp*.

In our materials, *ja* as a filler was often accompanied by a second filler.

(14) *Ich würde dann [. . .] erst mal anfangen äh, ja, erst mal mit der Grundfarbe*. 'I would start out with *uh*, well, the foundation first of all.' (S #2)

(15) *Das Schneewittchen hat natürlich aufgemacht und ähm, ja, ähm die hat gefragt* [. . .] 'Snowwhite opened the door, of course, and *um um um* she asked [. . .]'. (S #4)

But *ja* is not the only additional option for a verbal filler. There are numerous lexical items, such as *halt stopp*, *Moment* (which outrightly expresses the speaker's need for more time), or even *wie sagt man gleich*, etc., which lend themselves to be used as hesitation markers. In English, *well*, *okay*, or *how do I put it* as a conscious marker of the speaker's search for words might adopt this role, as in (16).

(16) *I believe there were, well, maybe 200 guests at the wedding*.

4.1.1.2.3. Repetitions

Hesitation markers do not necessarily occur as singular events. They may be repeated, and the same is true for other lexical items, as in (17), which is a quotation from Katharina Thalbach, a well-known German actress and director. These repetitions also serve the purpose of gaining time and can therefore be considered hesitations. They may be combined with fillers, and fillers may be repeated, as in (18):

(17) *Wo is sein sein Fundus, mit dem er arbeitet, und wo is sein sein Reser- sein sein sein sein Becken, aus dem er die Dinge holt?* 'Where is his his fund that he is working with and where is his his reser- his his his his pool that he is getting things from?'

(18) [. . .] *und das ähm mh mh naja, ja im Augenblick ist das ein bisschen ein zäher Roman.* '[. . .] and this *um mh mh*, well, right now this is a bit of a boring novel'. (S #1)

Stenström (2012, p. 541) mentions that repetitions mostly affect function words. Without having addressed this question systematically, it can be said that this appears to be the case for our materials as well.

4.1.1.2.4. Glottal Constrictions

One further type of filler, which has so far been largely neglected in the literature, can be described as increased glottal constriction with low subglottal pressure. In fact, Belz (2021) is the only author who describes precisely this case. He considers it as one of the "glottal fillers".

The result is a very short creaky sound which is distinct from a creaky *uh* or *um* by a shorter duration and a lesser degree of periodicity. It does not appear to possess any specific

vowel quality but a transitional narrowing of the vocal folds with insufficient subglottal pressure to achieve regular vocal fold vibration. It could be interpreted as a false start in the sense that the speaker adducts the vocal folds in preparation for speaking, then realizes that they are not ready to start and abducts the vocal folds again. Here is an example (19), which is also shown in Figure 3:

(19) [. . .] *irgendwie vvv <constr.> vonn* [„,] '[. . .] somehow fff <constr.> frommm [. . .]' (S #6)

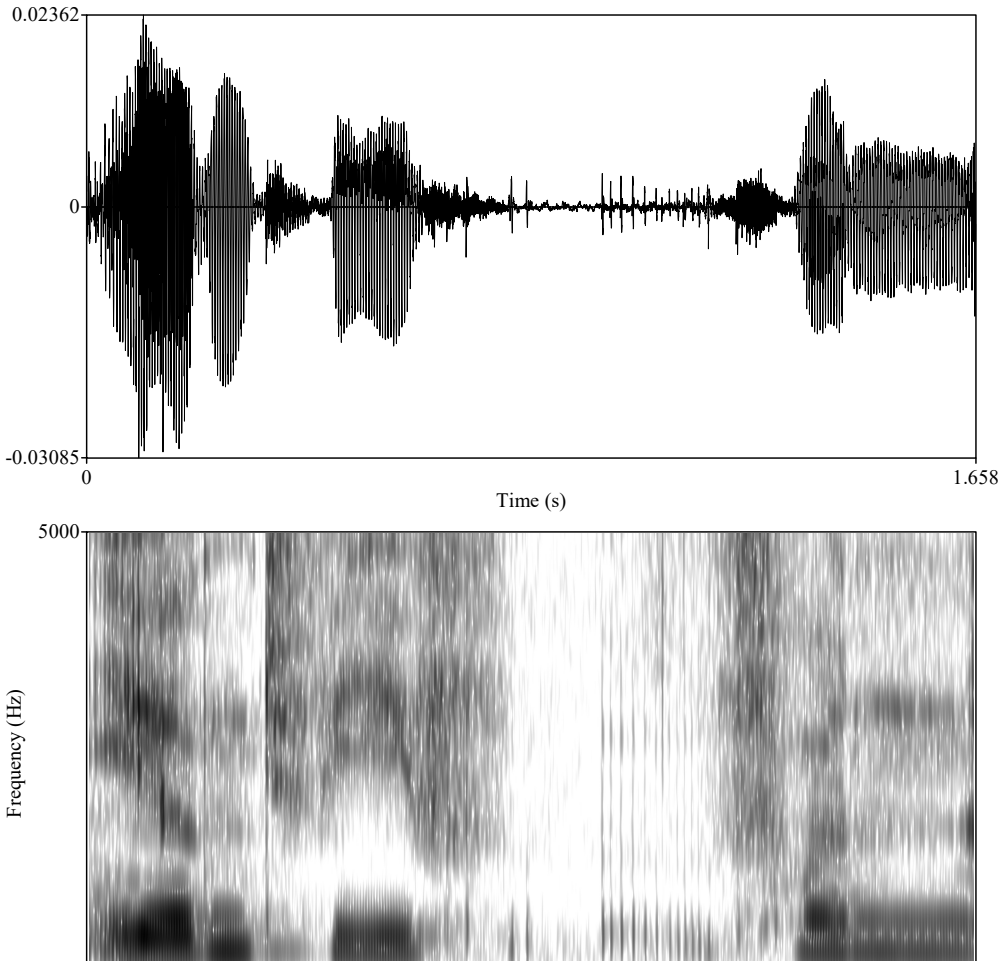

Figure 3. Typical example of a glottal constriction. The example is the one cited in (19). The constriction is visible as a very brief irregularity in the signal.

4.1.1.2.5. Non-Speech Phenomena

Naturally, there is the full repertoire of nonverbal vocalizations, such as clicking, laughing, coughing, throat clearing, swallowing, yawning, and possibly even burping, which can all be used to fill pauses. (cf. Trouvain 2014). There is a difference between clicking and laughing on the one hand and the rest on the other hand, though. The last five listed above usually fulfill a physical need and can therefore not be regarded as hesitation markers in a strict sense. They may also contribute to the idiosyncratic behavior of a speaker in rare cases.[8]

4.1.1.2.6. Onomatopoetics

There is still another way of (repetitious) hesitating, and this is by producing spurious sounds such as

(20) [B̥B̥B̥] in: [. . .] *ach so, das war ja letzte Woche, mh, j-* [B̥B̥B̥] *ja ja an einem Tag essen wir wahnsinnig gerne immer in der Woche mh einen Salat*. '[. . .] oh yes, that was last week, *mh*, y- [B̥B̥B̥], well, yes, we really like to eat a salad *mh* once a week [. . .]' (S #1)

This is a type of hesitation marker which the present authors have not seen mentioned in any previous publication, even though it is by no means a hapax legomenon in our materials.

4.1.1.3. Prolongations

If hesitations serve the purpose of gaining time in order to plan the upcoming utterance or search for the adequate lexical item, it is quite clear that non-linguistic sounds or fillers are not the only means to achieve this goal. Instead of inserting an element, existing elements may be lengthened. And yet, Eklund seems to have been the first to draw detailed attention to this (Eklund 2000, 2001). He points out that prolongations are more common than most other types of hesitations, outnumbered only by filled pauses and unfilled pauses (Eklund 2000; Eklund and Shriberg 1998). Betz (2020) observes that prolongations, such as verbal fillers, may serve various purposes. He distinguishes disfluent lengthenings from accentual ones and forced-alignment errors. The present contribution addresses the first type only.

Stenström (2012) states that prolongation typically affects conjunctions. While this is certainly true for *und*, a superficial look at the data for the present study shows a large number of counterexamples. This certainly merits looking into in future research.

According to Duez (1993), prolongations work like pauses or fillers do. Clark and Fox Tree (2002) consider them as a phonological alternative to fillers. This constitutes one of their arguments that fillers are regular words. Betz et al. (2017), on the other hand, follow Eklund (2001) and argue that durations of fillers and lengthened segments are fundamentally different. As a consequence, they consider them to be different in function.

Prolongations, however, are not random. It seems reasonable to establish the sound class that is lengthened (vowels or consonants), as well as the position of the lengthened sound within the syllable (initial vs final). The following examples from the recordings analyzed in this contribution illustrate the lengthening of a word-initial or word-final vowel or consonant:

(21) Initial vowel prolongation: [. . .] *uuuund damit er auch nicht abhauen konnte aus diesem Käfig*. '[. . .] aaaand so that he couldn't escape from this cage'. (S #4)

(22) Final vowel prolongation: *Xanten waaa ja eine alte Römerstadt*. 'Xanten issss a town going back to Roman times'. (S #1)

(23) Initial consonant prolongation: *Das ist halt fffürn Rücken und für die Knie und für alles Mögliche gut.* 'This is good ffffor your back and for your knees and for all sorts of things'. (S #2)

(24) Final consonant prolongation: [. . .] *weil ich ja Nichtraucher bin und eigentlichchchch ähm dieses äh diese Rauchschwaden nicht leiden kann* [. . .] '[. . .] because I am a nonsmoker and I really can't stannnnd[9] this uh these clouds of smoke [. . .]'. (S #1)

Prolongations may be combined with fillers. The precise way in which this is happening would merit looking into. It could well be that there is some sort of implicit signaling of trouble first (by the prolongation), and if that turns out to be too short, it is complemented by a filler proper.

4.2. Repairs

Self-repairs are considered the second subcategory of disfluencies. They are distinguished from the hesitations in that they happen *post factum*, i.e., something has already gone wrong and needs to be fixed. This is why Levelt (1983) calls repairs overt corrections as opposed to covert ones, which happen before an error is manifest.

4.2.1. Lexical Corrections

The first category to be identified are self-corrections (SCs), which describe a situation in which a lexical item that was incorrectly chosen or pronounced is substituted by one that is deemed more suitable by the speaker. A twofold example is given in (27): the definite article (m) *der* is replaced by the demonstrative (f) *diese*, and the preposition *in* 'in' is replaced by the preposition *auf* 'on'.

(25) [. . .] *und dann wird der- diese Käsemasse da rein geschüttet in den in den äh ja also auf den Boden.* '[. . .] and then, the- this cheese mixture is poured in there in the in the uh well on the baking sheet'. (S #2)

4.2.2. Insertions

Insertions constitute different mechanisms, i.e., a lexical item is added which was missing in the original utterance. An example of this is (26), which actually comprises a lexical correction first (*son Stück* 'such a bit' is replaced by *so ne Größe* 'such a size') and then an insertion (the adjective *gewisse* 'certain' is inserted before the noun *Größe*).

(26) *Und wenn die dann nachher son Stück so ne Größe e gewisse Größe erreicht ham* [. . .] 'And once they have reached a bit a size, a certain size' [. . .]. (S #2)

4.2.3. Change in Sentence Structure

A change in sentence structure comprises examples in which the speaker makes a change to the syntactic structure of the (remaining) utterance and starts anew. In example (27), the intended direct object (*ne Grun-* 'a foun-', which was probably going to be *ne Grundierung* 'a foundation') is replaced by the prepositional phrase *mit nem Bleistift* 'with a pencil'. This is the most far-reaching repair mechanism conceivable.

(27) *Also erst mal würd ich ne Grun- erst mal würd ich mit nem Bleistift mir mal vormalen, was ich überhaupt malen will.* 'So to start with I would a put a foun- to start with I would sketch what I want to paint with a pencil'. (S #2)

5. Use of Hesitation Markers

5.1. Frequency of Occurrence

The next step in determining individuality in disfluency behavior is to establish the proportion of the various disfluency phenomena. This can be measured in different ways. The easiest way of establishing the frequency of disfluencies is to express it in terms of number per time unit, i.e., second or minute. This, however, fails to take account of the individual speaking tempo. Therefore, the average number of linguistic units (words or syllables) between disfluencies may be more appropriate. These could be words or syllables. The former are easier to measure, but in languages such as German, which make frequent use of multiple compounds (consider examples such as *Donaudampfschiffahrtskapitänswitwe* 'Widow of a steamboat captain on the Danube'), it may be more advisable to use the syllable as a unit unless the topic is kept constant. In the literature, it is common to express the frequency of disfluencies by the number of occurrences per 100 words (Clark and Fox Tree 2002; Bortfeld et al. 2001; Corley et al. 2007) or—in corpus studies—as a percentage of the total number of words (Kjellmer 2003). Due to the between-language differences in word length, (linguistic) syllables as opposed to phonetic syllables or words should definitely be used in cross-linguistic studies.

5.2. Combining Hesitation Markers

Hesitation markers are by no means confined to single events. They may in fact be combined in various ways: The same marker may simply be repeated (in practice, this applies almost exclusively to fillers), as in (28), or different markers may be combined and possibly also repeated, as in (29) and (30).

(28) [. . .] *der Bademeister hat dann gesagt (Lachen) ähm ähm ich werde mit dem reden.*
'[. . .] the pool master said [laughs] *um um* I will talk to him.' (S #1)

(29) *Was mir zu kurz kommt bei dem Buch sind so en bisschen ähm äh die Schilderungen doch auch der Umwelt mh äh* [. . .] 'What comes up short in this book is in my opinion *um uh* the description of the environment *mh, uh* [. . .]' (S #1)

(30) *Die [Käsekuchenhilfe] wird mit Milch und Eier ähm ja mit mit ähm mit, mit Eigelb wird es vermischt zu einer Masse.* 'The [cheesecake mix] is blended with milk and eggs *um uh* with with *um* with with eggyolk' (S #1).

5.3. Combining Hesitation Markers and Text

Another way of distinguishing the use of hesitation markers and thus establishing individual patterns is the way they are attached to the text before and after. They may simply be imbedded in pauses. Those could be termed *isolated* as in (33).

(31) *This is # uh # not a very good idea.*

They may also be either preceded or followed by pauses, i.e., linked up with the preceding or following lexical item. These could be termed *connected*, as in (34) and (35).

(32) *This is uh # not a very good idea.*

(33) *This is # uh not a very good idea.*

They may finally be attached to text at both ends, i.e., with no pauses at all: these could be termed *imbedded*.

(34) *This is uh not a very good idea.*

With the forensic perspective in mind, it would certainly be useful to establish whether or not speakers have an individual preference for imbedding hesitation markers into the surrounding text without pauses.

5.4. Phonetic Properties of Hesitation Markers

The next step in the quest for the individuality of hesitation behavior is to look at the phonetic properties of the markers. When analyzing them, it seems advisable to distinguish the laryngeal from the supralaryngeal domain.

5.4.1. Fundamental Frequency

At the laryngeal level, the fundamental frequency of the fillers, particularly in relation to the surrounding text, is of interest. Previous research has shown that as a rule, the F0 of fillers tends to be lower than that of the surrounding text (cf. Shriberg and Lickley 1993; Batliner et al. 1995; Swerts 1998; Zhao and Jurafsky 2005; Rosin 2011; Braun and Rosin 2015). This touches upon a more general point: it would be worthwhile to see whether the fillers other than *uh* and *um* differ from these two in this respect. This would serve as an indication of whether or not the fillers really fall into two distinct categories.

5.4.2. Duration

Another element to be studied is the duration of isolated fillers and the total pause duration consisting of pause–filler–pause. This would, at the same time, allow us to check Clark and Fox Tree's (2002) claims about there being longer pauses after *um* than after *uh*.

5.4.3. Voice Quality

Furthermore, the voice quality of the hesitation markers may be a source of individual variation. In the present context, creak(y voice), breathy voice, and whispery voice are obvious candidates. Figure 4 shows a typical example of creak.

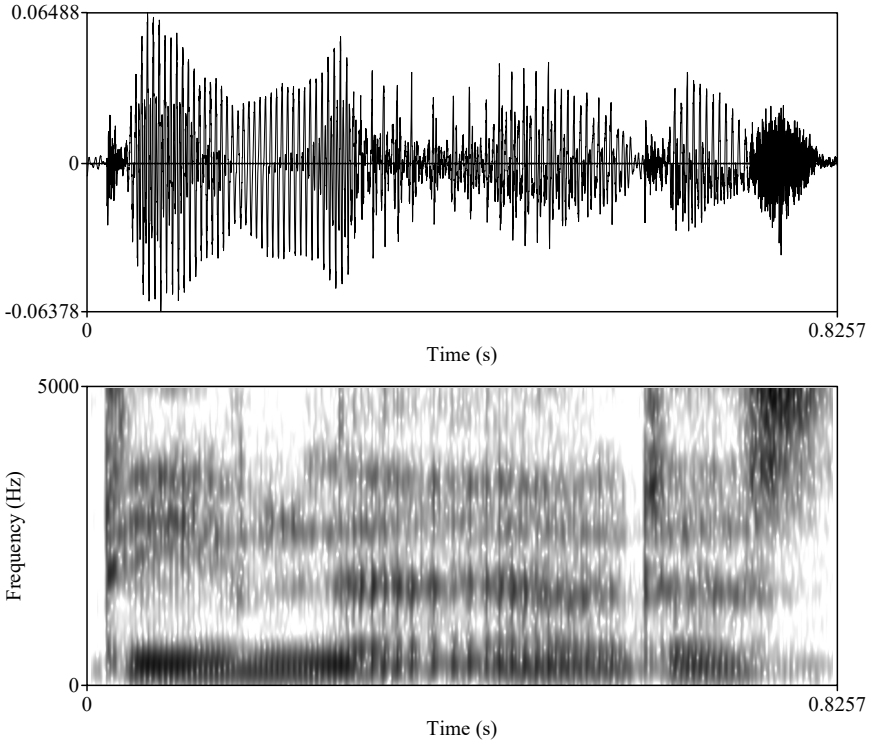

Figure 4. Typical example of creak (S #7). The utterance is [*nach*]*dem äh er das* [. . .].

5.4.4. Spectral Composition

In the supralaryngeal domain, formants are the method of choice for measuring the acoustic quality of vowels in fillers. In this context, it is of interest whether a subset is used of those vowels which form part of the phonemic inventory of a given language or whether there is a separate "hesitation phoneme", or even an idiosyncratic way of forming hesitation sounds. At the same time, within-speaker variability of the hesitation vowels is of interest. It may be useful to distinguish between isolated fillers and connected ones, because the latter may show more coarticulation with the preceding or following word and thus be more variable.

Part B: A Sample Analysis

Having considered the quest for individuality in disfluency behavior from a systematic point of view, a sample is (re-)analyzed to demonstrate what a comprehensive analysis could look like. There were some promising results in a relatively small pilot study from some years back (Braun and Rosin 2015), which analyzed different types of "conventional" hesitation markers only. Preliminary analyses indicated that for these, the type of hesitation alone is not sufficient to discriminate between speakers. Therefore, additional aspects of hesitation behavior fulfilling the criteria outlined above are included in the analysis.

6. Materials and Methods

Materials consisted of recordings from eight middle-aged female speakers who were all from the same part of the country. This eliminated gender, age, and dialect as influencing factors. Subjects' age varied between 45 and 65 years. They were thus between the middle-aged and the older group studied by Bortfeld et al. (2001). Those researchers found "only slightly higher disfluency rates" (p. 123) in their older speakers than in their middle-aged ones, the difference amounting to 1 filler per 100 words. In the present data set, no significant correlation was found between the frequency of disfluencies and speaker age ($r = 0.286$; Spearman rank correlation). Handedness was controlled for because the retraining of left-handers may affect speech fluency (Kushner 2012). All speakers were right-handed. There were two smokers among the participants, who reported smoking

3–4 cigarettes daily. All speakers were working and—despite varied degrees of formal training—held jobs requiring them to communicate regularly with customers or students. Speakers were recorded talking spontaneously about a fixed set of topics (their recent vacation, books they had read, soccer, their opinion about the ban on smoking in restaurants and pubs, what they would tell Angela Merkel if they had a chance to meet her, etc.). They were prompted by the investigator once they ran out of things to say on a given topic, but other than that, the speech material was monological. Thus, certain factors which have been found to influence disfluency rates, such as relationship, topic, and syntactic complexity (Bortfeld et al. 2001), could be ruled out for this data set. There were three recording sessions per speaker, which took place about a week apart. This was conducted in order to be able to test within-speaker variability against between-speaker variability. Prior to each recording session, speakers were asked if they felt fatigued. They were recorded only of they said that they did not. Sessions lasted between 7 and 15 min, depending on the number of disfluencies produced and the speaking tempo. A total of 100 filled pauses were aimed for in each recording session. This aim was surpassed.

Recordings were analyzed by creating Praat (Boersma and Weenink 2022) TextGrids and annotating the type of pause (breath, silent, filled), the type of filler, the way fillers were worked into the text (preceded and/or followed by a pause), glottal constrictions, verbal fillers, as well as restarts and repairs. The steady phase of the "classical" fillers was noted, and it was subsequently used for formant measurements. All annotations were carried out manually. The following 15 parameters were analyzed by way of Praat scripts: voice fundamental frequency; number of filled and unfilled pauses; number of breath pauses; number and type of hesitations; connection of fillers with surrounding text; frequencies of the first four formants of the "classical" fillers *uh* and *uhm*; voice quality of hesitations; glottal constrictions; and repairs.

Statistical analyses included *t*-tests and linear mixed-effects models for assessing differences in the use of various disfluency markers. However, considering parameters in isolation does not exploit the full potential of the measures proposed. Instead, patterns need to be examined. For those, we are lacking appropriate background data. We therefore tackled the problem from two different angles: a machine-learning approach using all of the results established, and a mathematical approach utilizing Hellinger distances to estimate classification error probabilities as derived by Fazekas and Liese (1996) for Bayes risks. Both test how well the different speakers can be discriminated based on disfluency patterns.

7. Results

It is quite clear from the above list that there is no room for reporting all the findings in this contribution. Therefore, the results focus mostly on parameters which have not been studied (extensively) in previous work and on speaker specificity. The approach in reporting the results is largely descriptive. The emphasis is not so much on generalization and statistical testing of individual findings, but rather on examining the speaker specificity of elements of disfluency patterns. Results are reported by speaker and by session. This replicates the forensic setting more closely than averaged results would and will therefore provide a realistic impression of intra- and interspeaker variability as it might appear in actual casework.

In the present study, a total of 442 min of recordings were analyzed. They contained a total of 18.436 disfluencies, 10.826 of which were pauses. Of the pauses, 8.117 were unfilled, and 2.709 were filled. Averaged over the 3 recordings, speakers produced a total of 28.4 disfluencies, 1.8 verbal fillers, and 5.3 prolongations per 100 words. Table 1 shows the details.

Table 1. Disfluencies, verbal fillers, and prolongations per 100 words and per minute for individual speakers.

Speaker No.	Disfluencies per 100 Words/per Minute	Verbal Fillers per 100 Words/per Minute	Prolongations per 100 Words/per Minute
#1	45.9/48.1	1.8/1.8	7.5/7.8
#2	17.8/33.4	1.3/2.5	2.1/4.1
#3	30.6/45.6	1.8/2.8	5.8/8.8
#4	36.2/41.5	0.1/0.1	11.6/13.5
#5	29.9/42.1	1.1/1.6	4.7/7.1
#6	32.5/48.5	3.7/5.5	3.0/4.5
#7	29.6/34.1	0.8/0.9	3.9/4.6
#8	28.4/43.1	1.3/1.9	5.3/7.9
Mean	28.4/41.7	1.8/2.2	5.3/7.1

Results are reported in terms of occurrence per 100 words and per minute. Differences between the two are owed to between-subject variation in speaking tempo and perhaps also to differences in preferred word length. The ranges are quite large, with speakers #1 and #2 representing the extremes of the disfluency rates per 100 words and all other speakers exhibiting comparable numbers. These numbers are much higher than those in other studies (for German, see, e.g., Belz (2021, p. 168), who reported between 1.9 and 11.3 filled pauses per minute), but they are not really comparable with most previous results because pauses, as well as hesitation markers and prolongations, are covered, which were not always included in previous work.

Furthermore, for the first time, extensive reference data for the frequency of verbal fillers in German are presented. Once again, individuals vary greatly in how frequently they use this type of disfluency marker. Speakers range from 0.1 to 5.5 per minute, and they can be grouped as follows: speaker #4, who uses basically no verbal fillers; speakers #1, #2, #3, #5, #7, and #8, who use them rarely; and speaker #6, who uses them frequently. It should be noted that the groups vary, i.e., speakers who are similar in one respect differ greatly in others.

7.1. Disfluency Markers

Figure 5 shows the proportion of the various disfluency markers relative to all vocalized disfluencies, i.e., without unfilled pauses.

Since sessions differ in length, proportions are given as opposed to absolute numbers. For most speakers, the bulk of disfluency markers are made up of fillers and prolongations. We found 5.3 prolongations per 100 words on average, with individual speakers ranging from 2.1 (speaker #2) to 11.6 (speaker #4). There were five speakers who exhibited more prolongations than fillers. This is in line with results reported by Eklund (2001). Speaker #5, though, showed a more even distribution of disfluency markers than the rest. In this case, glottal constrictions and lexical corrections equaled the number of fillers and prolongations.

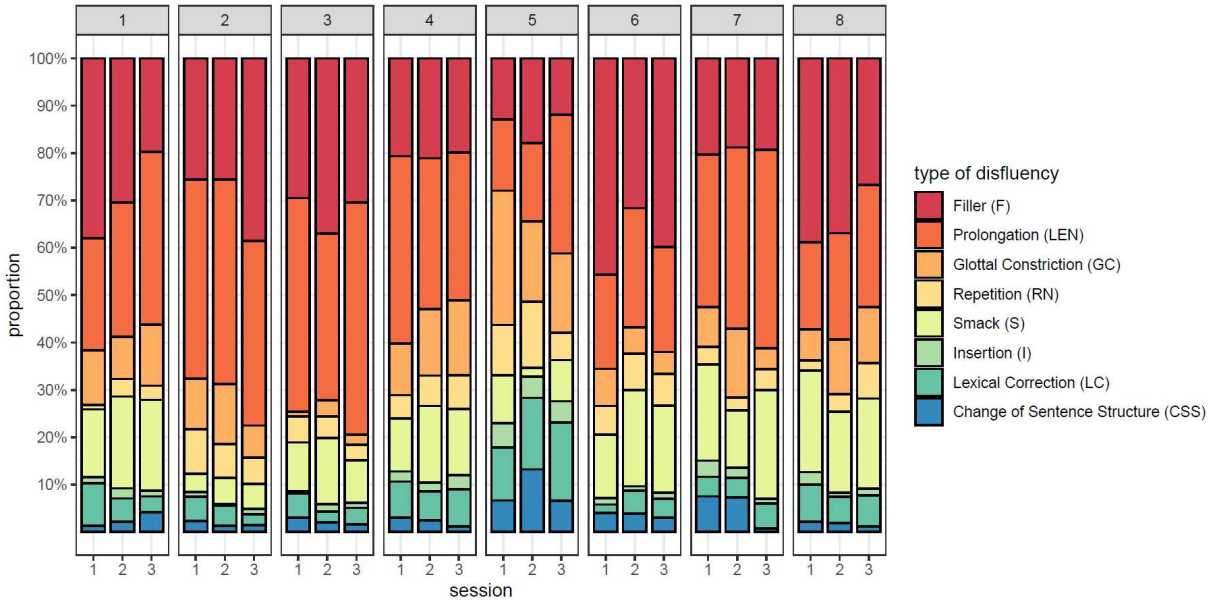

Figure 5. Proportion of different disfluency markers in relation to all vocalized hesitations. (F = filler, LEN = prolongation, CSS = change in sentence structure, GC = glottal constriction, I = insertion, RN = repetition, S = smack, LC = lexical correction (word or pronunciation)). Numbers at the top of the columns refer to subjects; numbers at the bottom refer to sessions.

The most striking result is probably that our speakers generally showed high within-subject consistency, i.e., the proportions in the different sessions look very similar. This impression needs to be confirmed by further statistical analysis. On the other hand, there are notable between-speaker differences. These apply, for instance, to the relation of filled pauses to prolongations. Speakers #3 and #4 differed sharply in this respect, whereas speaker #6 stood out for her rare use of prolongations. Considering prolongations only, the speakers can be grouped as follows: speakers #2, #6, and #7, who use them rarely; speaker #4, who uses them frequently; and speakers #1, #3, #5, and #8, who are somewhat in the middle (see Table 1). This is yet another indication that speakers differ in their hesitation behavior. These findings underline the importance of including prolongations in the analysis of hesitation patterns.

Figure 5 also shows the frequency of occurrence of clicking sounds, denoted by the letter <S> ('smacks'). It is evident from the distribution that there are speakers who use this hesitation marker quite frequently. In fact, speaker #7 uses them almost as often as fillers, whereas speakers #3 and #6 hardly exhibit any clicks at all.

7.2. Proportion of Fillers

In addition to *äh* and *ähm*, which are commonly analyzed, results on *mh* and verbal fillers are provided here. Figures 6 and 7 summarize the results for the "classical" fillers *äh* (V) and *ähm* (VC), as well as *mh* (C) and verbal fillers (termed "O" in the caption).

Figure 6 shows the number of the "classical" fillers (*äh*, *ähm*) and *mh* per 100 words per speaker and session. They range from 1.8 to 6.4. This considerably exceeds the numbers given by Fox Tree (2001), who found a median of 1.73. One explanation for this difference may be that German is more prone to hesitations than English due to its morphological and syntactic complexity.[10] The number of these fillers per minute ranges from 2.2 to 11.2 in our materials and corresponds very well with those reported by Belz (2021), i.e., 1.9 to 11.3 per minute.[11] The nasal filler is hardly, if ever, used by four of our speakers, but two use it about as often as *äh*. This alone underlines the necessity of including both *mh* and verbal fillers in the analysis of fillers.

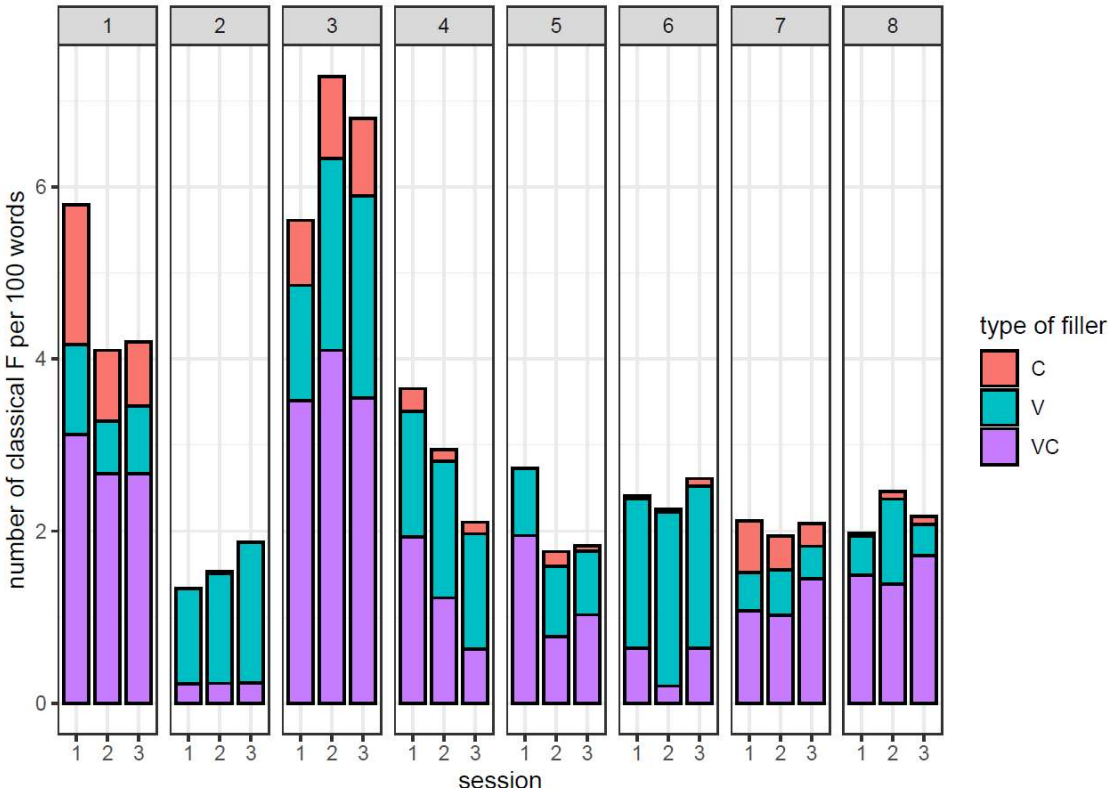

Figure 6. Number of "classical fillers", including *mh* per 100 words according to speaker and session (C = consonantal filler *mh*; V = vocalic filler *äh*; VC = filler consisting of vowel plus consonant *ähm*).

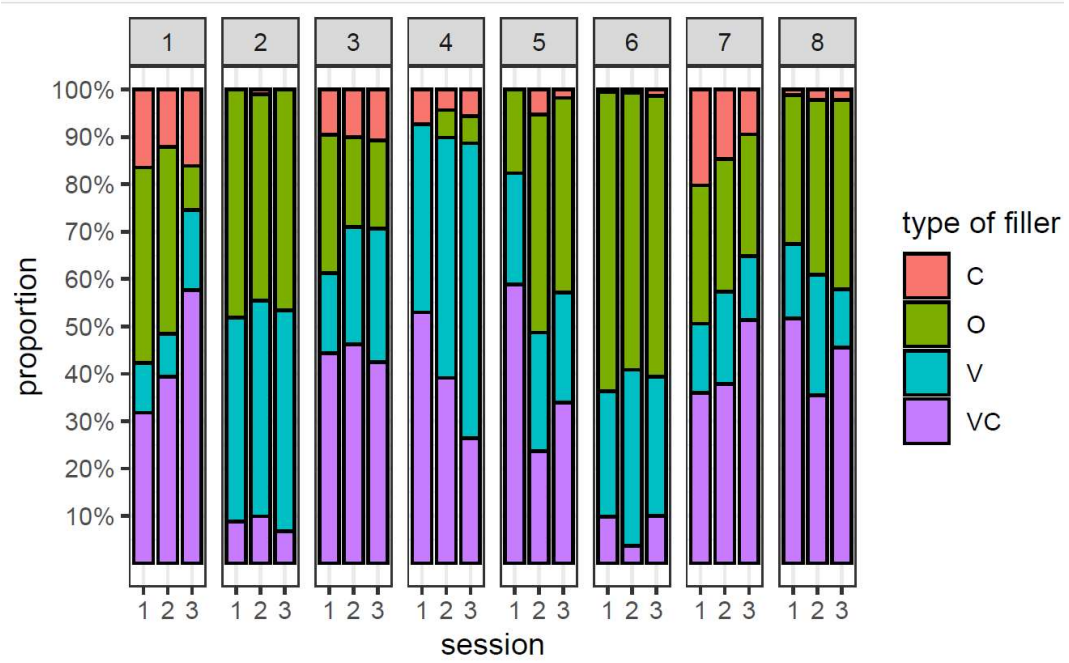

Figure 7. Proportion of various types of fillers in relation to all fillers (C indicates consonantal fillers; V indicates vocalic fillers; VC indicates a combination of vocalic and consonantal elements; and O indicates verbal fillers).

The preference between *äh* and *ähm* is clearly speaker-specific in our materials. Speaker #1, for instance, uses very few *äh*s, whereas speakers #2 and #6 very rarely use *ähm*. Unpaired two-tailed *t*-tests per speaker yield significant results for six speakers, four of whom prefer *ähm*, two of whom prefer *äh*, and two speakers who show no clear preference (speaker #1: $t = -3.8$, df = 2.39, $p = 0.046$; speaker #2: $t = -6$, df = 3.916, $p = 0.016$; speaker #3: $t = -6.007$, df = 2.4, $p = 0.016$; speaker 4: $t = 1.134$, df = 3.899, $p = 0.321$; speaker #5: $t = -1.425$, df = 2.011, $p = 0.289$; speaker #6: $t = 6.09$, df = 3.428, $p = 0.006$; speaker #7: $t = -4.974$, df = 2.566, $p = 0.022$; speaker #8: $t = -4.314$, df = 3.873, $p = 0.013$).

Two things are evident: most speakers (#2, #4, #5, #6, #7, and #8) are extremely consistent across sessions, and only #1 and #3 show a larger degree of variability. Secondly, a high degree of between-speaker variability can also be observed. This concerns the use and frequency of the consonantal filler, but also the proportion of the three remaining ones.

The proportions of the three classical fillers and the verbal fillers were analyzed using linear mixed-effects models (*lmers*) in R. Each model contained the speakers as fixed effects and sessions as a random factor. The *p*-values were obtained by using each subject as the reference level successively. Cross-tables containing the *p*-values for each speaker pair (N = 28) were produced, and the number of significant differences was counted. The results show that *ähm* discriminate best between the speakers (22 out of 28 pairs in the cross-table are significant), while the verbal fillers discriminate worst (13 out of 28 comparisons falling short of significance).

7.3. Fundamental Frequency

Support for considering the nasal filler and verbal fillers as hesitation markers as opposed to interjections comes from the F0 data. Previous researchers found a lowering of the fundamental frequency of hesitation markers relative to their immediate context (Shriberg and Lickley 1993; Batliner et al. 1995; Rosin 2011; Swerts 1998; Belz 2021). For our materials, results depend on how the measurements are carried out. If we consider the F0 values averaged over all fillers per speaker and session and compare them with the averaged values of the sequences immediately preceding the filler, then fillers are higher in F0 in our materials. If, however, differences between the individual fillers and their immediate context are looked at, the F0 of the verbal fillers differs from that of the surrounding text by a sizeable lowering (cf. Figure 8). The general declination of F0 throughout an utterance may serve as an explanation for this finding. This shows that there is no such thing as a typical "filler frequency", but that the contrast to the immediate environment is crucial. There is no significant difference in this respect between *äh* and *ähm*, on the one hand, and *mh* and verbal fillers on the other (unpaired two-tailed *t*-test; $t = 1.064$, $p = 0.293$). This can be taken as yet another indication that these two constitute fillers in their own right. Once again, though, there is one speaker (#7) who does not follow this "pattern". This is in accordance with Belz (2021, p. 126), who also mentions that one speaker behaved differently.

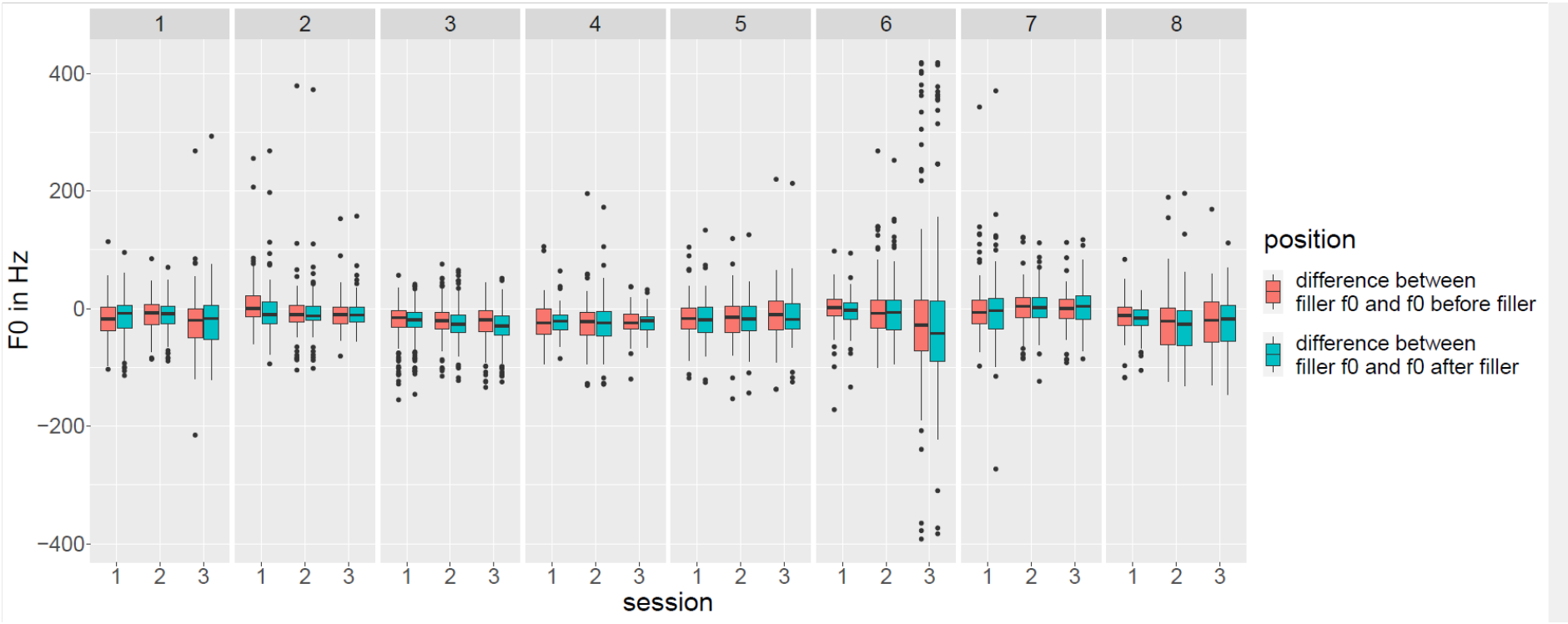

Figure 8. Differences in F0 between the filler and its immediate context (3 s before and after) for subjects individually.

7.4. Multiple Fillers

Fillers do not always appear in isolation. They may be repeated or combined with other fillers. The frequency of occurrence of multiple fillers has not to our knowledge been reported. Our results (see Figure 9) show that the use of multiple fillers is again individual. Speakers #2, #3, and #6 make quite frequent use of this option, while speakers #4, #5, #7, and #8 very rarely do. Speaker #1 is very variable and can be counted in the first group.

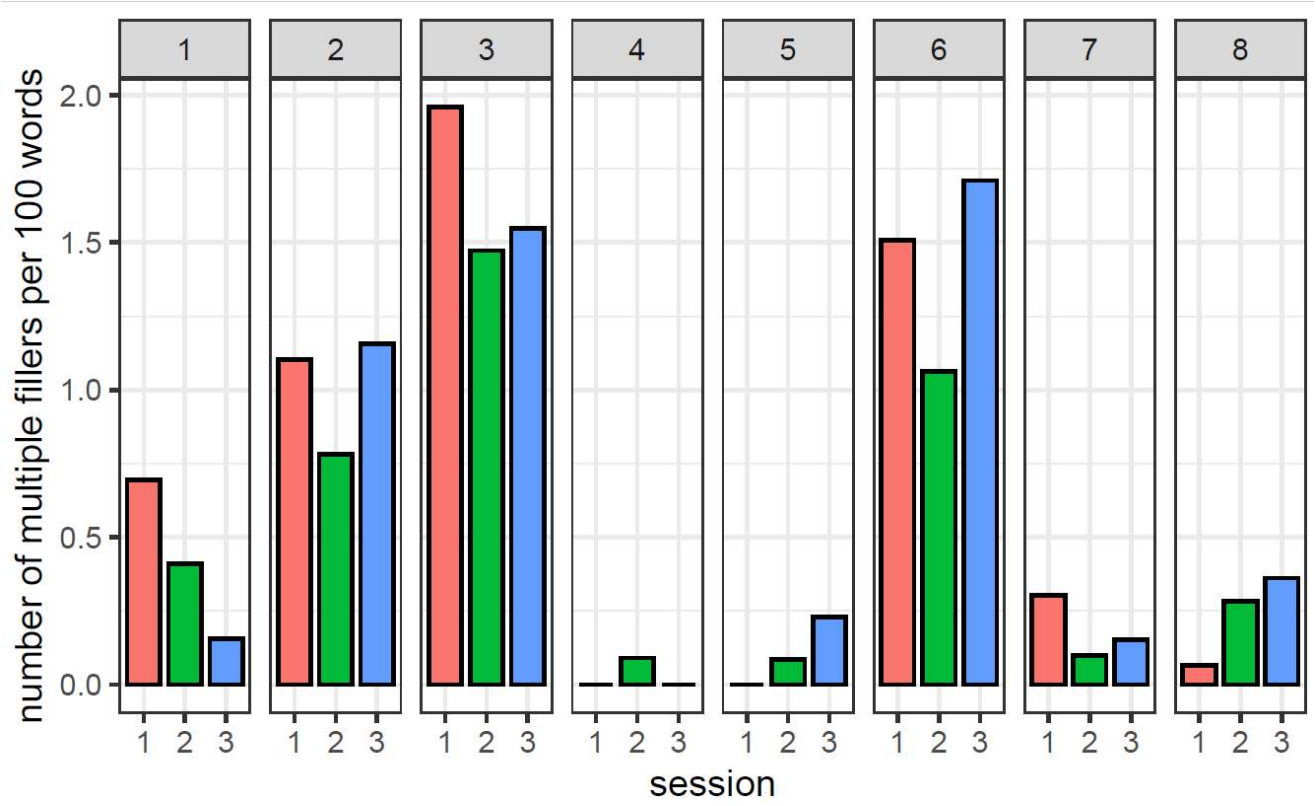

Figure 9. Multiple fillers per 100 words for subjects individually.

The calculation of a linear mixed-effects model focusing on speaker individuality rendered significant differences for 17 out of 28 pairwise comparisons.

7.5. Linking of Fillers

A parameter which has only rarely been addressed earlier on (see, however, Jessen 2012) is the way fillers are linked to the text around them (see Section 5.3 above). Four cases can be distinguished: there are pauses both before and after the filler (isolated filler); the filler is preceded by a pause; the filler is followed by a pause; and there are no pauses before or after the filler (imbedded filler). Figures 10 and 11 show the results.

Once again, a high degree of intra-speaker consistency and, at the same time, inter-speaker differences can be observed. This time, speakers #1 and #7 stand out by exhibiting very few cases of imbedded fillers. They differ from each other by the proportion of <F> and <M>, i.e., speaker #1 shows more isolated fillers, whereas speaker #7 exhibits more fillers preceded by pauses only.

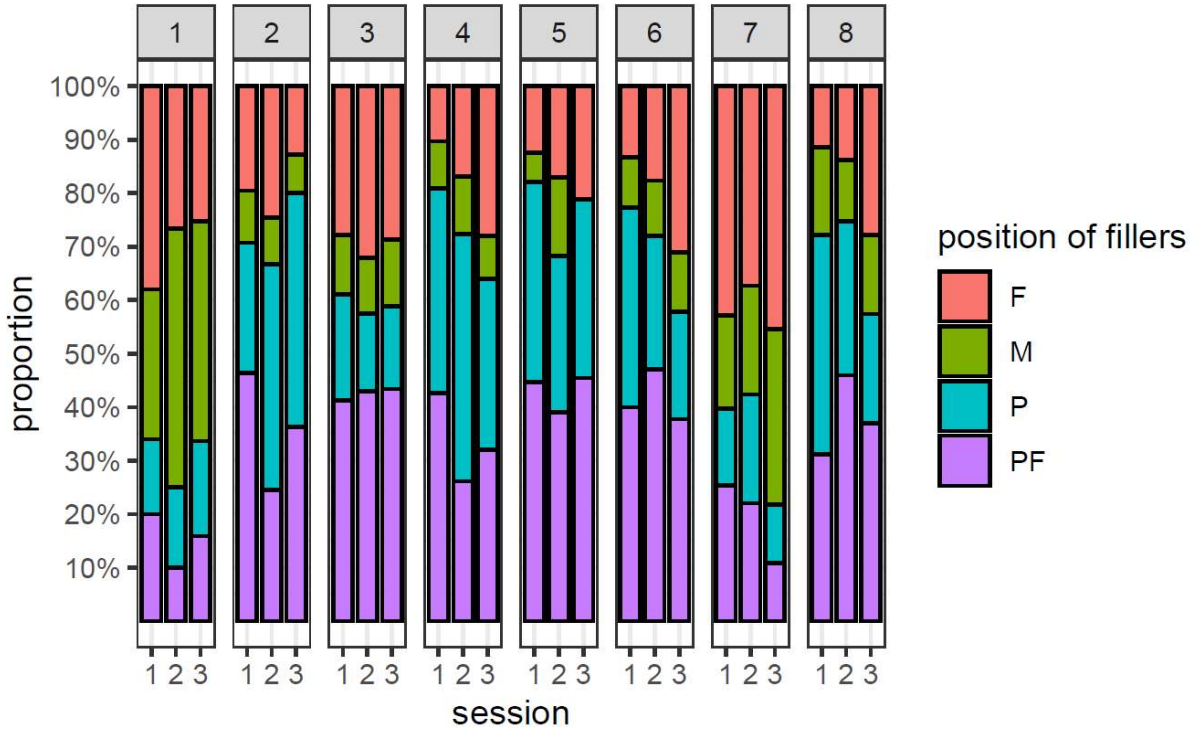

Figure 10. Linking of all fillers to surrounding text. < M> means that the filler is isolated, <P> means that the filler is followed by a pause, <F> means that it is preceded by a pause, and <PF> means that it is imbedded.

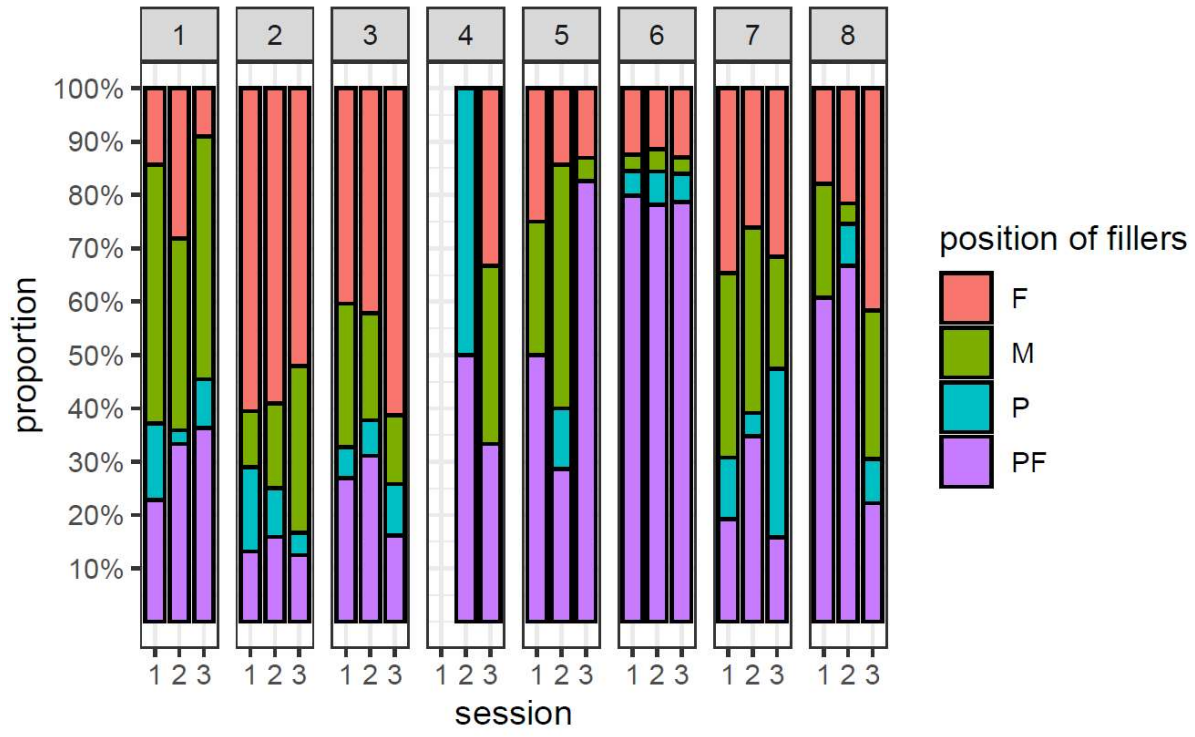

Figure 11. Linking of verbal fillers to surrounding text.

If one looks at the results for verbal fillers only, the picture is a lot clearer.

The verbal fillers differ from other fillers in that speakers either do not pause before and after or pause in both instances. Pauses before or after only are relatively rare. Speaker

#6 pauses only rarely before and after a filler, whereas speakers #2 and #3 articulate fillers with a preceding pause only in about half of cases. Speaker #1 prefers isolated fillers. Again, this feature is apt to distinguish speakers. The four types of linkage between fillers and text were analyzed using linear mixed-effects models (*lmers*) in R. Each model contained the speakers as fixed effects and the sessions as random factor. The *p*-values were obtained using each subject as the reference level successively. Cross-tables show 18 pairs out of 28 with the PF type to be significant (no pause before or after a filler), whereas that was true for only 12 of the P type (filler followed by a pause).

7.6. Formants

Figure 12 shows formant plots of the vocalic consonantal (VC) and vocalic (V) fillers for each speaker. The central part of the vowel was measured. Values derived from vocalic fillers are represented by the letter <V>, whereas <VC> denotes *ähm* fillers.

The formants are very consistent within speakers for the most part and vary between speakers. They generally cluster very well. For all speakers, the formant plots reveal differences between the vocalic and the VC fillers in that vocalic fillers have a higher F2 than VC fillers. A *t*-test revealed this difference to be significant for all speakers but speaker #4 (speaker #1 $t = 2.6478$, df = 46.151, $p = 0.011$; speaker #2: $t = 2.330$, df = 54.699, $p = 0.023$; speaker #3: $t = 3.879$, df = 251.51, $p = 0.000$; speaker #4: $t = 0.333$, df = 152.41, $p = 0.7393$; speaker #5: $t = 3.763$, df = 102.84, $p = 0.000$; speaker #6: $t = 2.825$, df = 73.158, $p = 0.006$; speaker #7: $t = 5.046$, df = 68.244, $p = 0.000$; speaker #8: $t = 3.173$, df = 109.96, $p = 0.002$).

This means that seven out of eight speakers articulated the vowel in the vocalic filler further front than the vowel in the VC filler. The remaining formants do not render a clear picture: the difference in F1 reaches significance for speakers #5 and #7 only, the formant being lower in the VC filler than in the V filler. F3 differs significantly for speaker #1 only, while F4 differs for speakers #1, #2, and #6 (unpaired *t*-test, two-tailed). These results show that among our speakers, there is no specific "hesitation vowel" distinct from the vowels of German. Instead, different vowel qualities have to be assumed for *äh* and *ähm*, the former suggesting a more peripheral quality, and results for higher formants are individual. While the present findings confirm previous research, which has demonstrated that filler formants are a valuable parameter in speaker comparison (Hughes et al. 2016), it is not advisable to pool formant values of different types of fillers in actual casework.

A statistical challenge in this study consists in the fact that in the present context, one is trying to examine what is normally considered a confounding factor, i.e., that individual speakers differ. This also involves comparing within-speaker variability and between-speaker variability. With filler formants, this can be conducted by way of likelihood ratio (LR) analysis, cf., e.g., Hughes et al. (2016). So far, the statistical procedures applied involved only one disfluency marker at a time. However, in order to determine speaker specificity, disfluency patterns need to be considered. We used two different approaches to address this problem. The first method employed was a random forest model. To train the model, the single takes[12] at all three points in time were treated as separate events. All measurement results in 70% of the takes were used for training. The remaining 30% were used in the classification task. The accuracy achieved was 78.5%. All data which had been determined (N = 245) were fed into the model indiscriminately. The parameters that were most important for the classification task were F1, F3, and F4, as well as the duration of the inspiration noise. Figure 13 shows the details.

Figure 12. Formant plots per speaker and session. Rows represent the eight speakers; columns show the results for each session per speaker. (V = vocalic filler *äh*; VC = vocalic-consonantal filler *ähm*.).

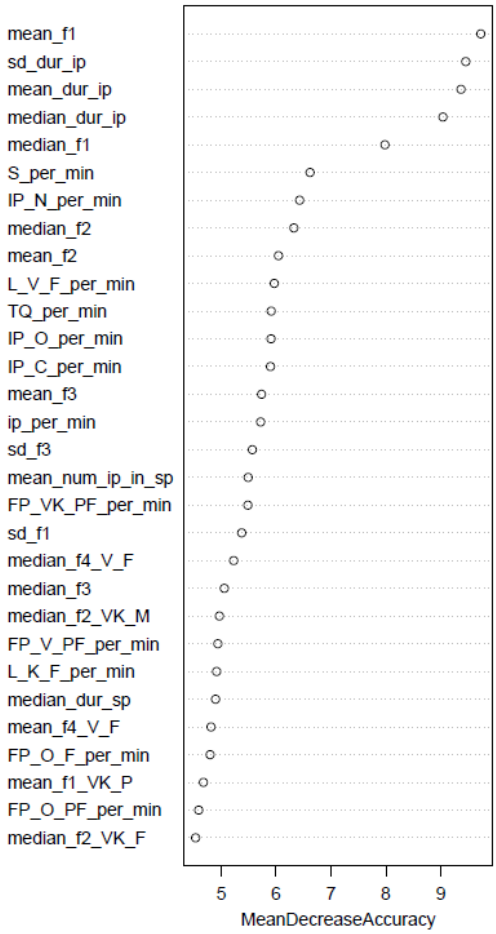

Figure 13. The 30 most important factors influencing attribution accuracy in the random forest model.

The second approach was to estimate error probabilities in binary and multi-speaker classification tasks by upper bounds on the minimax risks. For that purpose, we used a statistical model composed of multinomial and normal distributions, with the former modelling the proportions of types of "classical" and nasal fillers, and the latter modelling the distributions of F2, in an attempt to keep the model as simple as possible. In this case, a sequence containing 50 hesitations is correctly assigned to one of the eight speakers with an error probability of <0.174 using an appropriate classification procedure. This corresponds to an expected accuracy of >0.826 in the multi-speaker classification task. In order to derive these bounds, we applied a method based on Fazekas and Liese (1996) with the idea that Hellinger distances between distributions are large if speakers differ greatly with respect to the modeled disfluency markers. Bounds on the probabilities of error were also calculated for the simpler binary classification problems as listed in Table 2.

We also tested a two-dimensional multinomial model. Parameters entered this time were type of fillers (dimension 1) and linking of the fillers (dimension 2). A chi2-test was used to examine within-speaker homogeneity. It emerged that the assumption of homogeneity with respect to the linking of the fillers could not be rejected for any of the speakers. This confirms the impression of within-speaker consistency. In this model, a sequence containing 50 hesitations is correctly assigned to one of the eight speakers 60% of the time in a minimax approach. This corresponds to a probability of error of <0.4 for a multi-speaker classification and was computed using the same method as in the previous model. With regard to the binary classification, the maximum pairwise error probability is <0.283.

Table 2. Bounds for the probabilities of error in a binary multinomial-normal model based on 50 hesitations.

	2	**3**	**4**	**5**	**6**	**7**	**8**
1	$8 \cdot 10^{-21}$	$1 \cdot 10^{-2}$	$2 \cdot 10^{-6}$	$2 \cdot 10^{-4}$	$2 \cdot 10^{-16}$	$15 \cdot 10^{-1}$	$3 \cdot 10^{-8}$
2		$4 \cdot 10^{-11}$	$1 \cdot 10^{-5}$	$2 \cdot 10^{-10}$	$6 \cdot 10^{-3}$	$6 \cdot 10^{-21}$	$3 \cdot 10^{-8}$
3			$18 \cdot 10^{-2}$	$14 \cdot 10^{-1}$	$1 \cdot 10^{-8}$	$42 \cdot 10^{-2}$	$2 \cdot 10^{-3}$
4				$17 \cdot 10^{-2}$	$2 \cdot 10^{-5}$	$2 \cdot 10^{-5}$	$6 \cdot 10^{-3}$
5					$2 \cdot 10^{-7}$	$5 \cdot 10^{-3}$	$15 \cdot 10^{-2}$
6						$7 \cdot 10^{-15}$	$9 \cdot 10^{-5}$

The table shows that most probabilities are very small when classifying between each two of the eight speakers, i.e., the speakers can be distinguished very well using a small set of hesitation markers alone. The only speaker pairs who are not classified at the 5% level are #3 and #5, as well as #1 and #7. The boundary values depend on the number of hesitations. The success of a classification increases with the amount of information available. At 30 hesitations, e.g., the expected accuracy goes down to >0.36.

Given that the chance level is at 12.5%, both results are quite impressive. It was not to be expected that speaker discrimination would be perfect. Under these circumstances, 78.5% in the random forest model underlines the speaker specificity of disfluency behavior. Homogeneity ratings are particularly encouraging. This underlines the potential of including the analysis of disfluency behavior in forensic phonetic reports.

8. Discussion

The first and foremost result of our study is that the total number of disfluencies is much larger than most that are reported in the literature (see, e.g., Belz 2021 or Betz et al. 2017). There are several conceivable reasons for this. First of all, we considered types of disfluency markers that have not normally been addressed in previous research, i.e., verbal fillers, the nasal filler, glottal constrictions, and smacks. Furthermore, the large number could be task-related, due to speaker age, or owed to the annotation process. For instance, Betz et al. (2017) used a semiautomatic tool, whereas we annotated manually without setting a time limit for prolongations. Instead, prolongation was tagged whenever a sound was deemed to be longer than in fluent speech. Because all tagging was carried out by hand, relatively short prolongations were also taken into account.

Looking at individual disfluency behavior opens a new perspective. It makes very clear what was to be expected from the scattered remarks by various researchers about individual variation: disfluency behavior has the potential of being individual. This does not so much apply to each single parameter, but it does apply to the pattern, which is formed by the various observations. Statistical analysis of individual disfluency markers revealed significant differences between individual speakers, but there was no single parameter which would distinguish all speakers.

This contribution has attempted to shed some light on the pattern-forming elements. Statistical analyses have shown that speakers can be discriminated at a high probability. Still, some parameters proved to be better than others at distinguishing speakers in our materials. The random forest model lists the parameters which contribute the most to speaker separation; in this case, formants and breath pauses. Owing to the nature of the process (machine learning), there is no way of knowing how this hierarchy came about and whether it will hold true for other groups of speakers as well.

The multinomial model, on the other hand, aims to keep things as simple as possible. We tested the proportion of "classical" and nasal fillers, the distribution of F2, the distribution of type of filler, the linking of fillers, and the type of filler, and again F2 proved to work

best at discriminating speakers. This shows that there is probably no single "idiosyncratic" parameter, but instead the full pattern has to be taken into account.

Some generalizations which have previously been considered to be well established cannot be confirmed if individual speakers are looked at. Examples are the prevalence of um over uh or the frequency of prolongations and glottal constrictions. The relevance of the nasal filler and the verbal fillers is underlined by the present findings. In other respects, previous results are confirmed, for instance, the F0 of fillers compared to the utterances immediately preceding and following it or the number of fillers per minute.

The ratio of *äh* to *ähm* is of particular interest for two reasons. First, Belz (2021) and Wieling et al. (2016) find that speakers, women in particular, generally produce significantly more VC fillers than V fillers. Our data do not support this conclusion. Three out of out eight speakers produced more vocalic fillers, the difference being significant for two. The findings by the above researchers may be the result of averaging across speakers. Secondly, the results concerning the use of *äh* vs. *ähm* is of interest in the context of Clark and Fox Tree's (2002) argument that speakers actively signal a long delay by using *um* and a short delay by using *uh*. Given the intra-speaker consistency of *äh* vs. *ähm*, it seems hard to conceive that two out of eight speakers would never use long delays at all, while one would want to signal long delays exclusively. Our results depend on the definition of pause. If only real gaps in the signal are taken into account, and the actual pause durations are compared, there is no significant difference between pauses following *äh* and *ähm*. If, on the other hand, a pause of zero duration is assumed whenever the filler is linked to the ensuing utterance without any gap in the signal, the difference is significant for seven out of eight speakers. This result underlines the necessity to think carefully about how measurement results are established.

As far as the general debate about the lexical status of fillers is concerned, our findings tend to oppose Clark and Fox Tree's (2002) theory about them being words. One potential counterargument to this is the high between-speaker variability paired with a low within-speaker variability of those fillers. In fact, we find the preference for *äh* vs. *ähm* to be highly individual. If a speaker has a clear preference for one of the two, it can hardly be argued that speakers make a conscious choice in each individual situation.

9. Conclusions

The questions asked at the outset can be answered as follows.

There are definitely a number of parameters contributing to individual disfluency patterns which have not received the attention which they deserve: verbal fillers; the nasal filler; multiple fillers glottal constrictions; prolongations; repetitions; non-speech vocalizations; spurious sounds as well as the linking with the surrounding text.

Many results show striking intra-subject similarity across the three sessions. This applies, for instance, to the proportion and the linking of fillers. This demonstrates that disfluency behavior is by no means random but instead follows individual patterns, given that the setting is kept constant. That said, while there are speakers who show a large degree of intra-speaker consistency, others are quite variable.

In this small group, between-speaker differences exceed within-speaker differences most of the time. Usually, at least one distinguishing element allowing for speaker discrimination could be identified. It has to be pointed out in this context, though, that similarity is not a sufficient criterion on which to draw conclusions about speaker identity. The typicality criterion also needs to be taken into account (Rose 2002). In order to be able to do this, however, much more material needs to be collected. Typicality will have to be assessed with respect to a speaker pattern as opposed to single disfluency markers. With the forensic application in mind, it is highly desirable to create language-specific databases for disfluency patterns in order to gain a better basis for assessing the typicality of the patterns encountered.

It was not the intention of this research to argue that speaker identification should rely on disfluency behavior alone. However, the present results demonstrate that while

no single parameter suffices to distinguish speakers, looking at the complete disfluency pattern has the potential of achieving just that. It can therefore be a valuable addendum to existing procedures. It is certainly worthwhile to extend the TOFFA framework proposed by McDougall et al. (2019) along the lines suggested here and by Braun and Rosin (2015). It might be argued that this approach is unrealistic because recordings of sufficient duration are rarely available in a forensic setting to carry out extensive analyses of the disfluency behavior. In jurisdictions which do not allow telephone intercepts, mismatches in speaking style between the questioned and reference materials may present a problem. In fact, Harrington et al. (2021) caution against using disfluency behavior under mismatch conditions. The present authors agree in part only. It could well be that we need to distinguish between differences in degree and in kind with respect to speaking style, the pattern staying intact while the frequency varies. In other words, a speaker who has a clear preference for *ähm* over *äh* may exhibit this preference in all speaking conditions, albeit at a different frequency. This claim is hypothetical for the time being. It will need to be substantiated by further study.

In jurisdictions which allow telephone intercepts, there are often many minutes of both questioned and reference speech material available from the same telephone surveillance measure. A detailed analysis of the disfluency behavior is therefore possible. Yet there is another field of practical forensic application to disfluency analysis. Sometimes the task for the forensic phonetician is not voice comparison, but the decoding of the content of surreptitious recordings. This also involves attributing relatively short utterances to one out of a closed set of speakers. Since the number of speakers and their names are known, it is fairly easy to tell them apart by their disfluency patterns, particularly if these are distinct. In these cases, it may prove extremely helpful to use disfluency behavior to make speaker attributions.

Author Contributions: Conceptualization, A.B.; methodology, A.B., L.W., N.E.; software, N.E., L.W., A.B.; validation, L.W., N.E., A.B.; data curation, N.E.; writing—original draft preparation, A.B.; writing—review and editing, A.B.; visualization, N.E.; supervision, A.B.; project administration, A.B., N.E.; funding acquisition, A.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Acknowledgments: The authors would like to extend a warm word of thanks to Melissa Hildebrand and Vivien Meyer for very patiently tagging many minutes of speech.

Conflicts of Interest: The authors declare no conflict of interest.

Notes

1 This debate may seem unrelated to the topic of this contribution, but it does have a bearing on the status of verbal fillers, where a distinction needs to be drawn between usage as filler and as lexical item.

2 Another reason why the auditory acoustic method of voice comparison is still widespread in Europe is that certain jurisdictions (e.g., England) do not allow the use of automatic systems.

3 In the taxonomy established here this corresponds to disfluency behavior.

4 This is not to be confused with other uses of *mh*, such as the reaction to excellent food, a backchannel asking the speaker to repeat his utterance, but also an expression of doubt or incredulous amazement. The decision about the nature of *mh* has to be made by the trained listener at every single instance.

5 Examples taken from our dataset are referenced indicating the speaker number. All other examples are fictitious.

6 See e.g., Brackhane (2022) for a corpus-based overview.

7 For an exhaustive account cf. e.g., Hentschel (2011) and Imo (2013).

8 This is the case for one of the "weathermen" on German TV, Donald Bäcker, who will almost invariably swallow while presenting the weather report.

9 In a word-by-word translation, the "really" would correspond to the word that is lengthened in German, but that of course ends in a vowel. Therefore, we have chosen the word "stand" to exemplify the prolongation.

[10] Compounding is excessive, and the finite verb is moved to the end of a dependent clause. A study on interpreting showed that compounding increases cognitive load, and pausing is a symptom of cognitive overload (Defrancq and Plevoets 2018).

[11] The corresponding graph is not presented here for reasons of space.

[12] One "take" is defined as a part of a recording dealing with one topic.

References

Batliner, Anton, Andreas Kießling, Susanne Burger, and Elmar Nöth. 1995. Filled pauses in spontaneous speech. Paper presented at XIIIth ICPhS, Stockholm, Sweden, August 13–19; vol. 3, pp. 472–475.

Belz, Malte. 2021. *Die Phonetik von äh und ähm. Akustische Variation von Füllpartikeln im Deutschen*. Berlin: Metzler.

Benchetrit, Gila, Steven A. Shea, Truong Pham Dinh, S. Bodocco, Pierre Baconnier, and A. Guz. 1989. Individuality of breathing patterns in adults assessed over time. *Respiration Physiology* 75: 199–210. [CrossRef] [PubMed]

Benchetrit, Gila. 2000. Breathing pattern in humans: Diversity and individuality. *Respiration Physiology* 122: 123–29. [CrossRef] [PubMed]

Betz, Simon, Robert Eklund, and Petra Wagner. 2017. Prolongation in German. In *DiSS 2017 The 8th Workshop on Disfluency in Spontaneous Speech, KTH, Royal Institute of Technology, Stockholm, Sweden*. Stockholm: KTH Royal Institute of Technology, pp. 13–16.

Betz, Simon. 2020. Hesitations in Spoken Dialogue Systems. Ph.D. dissertation, Bielefeld University, Bielefeld, Germany.

Blankenship, Jane, and Christian Kay. 1964. Hesitation Phenomena in English Speech: A Study in Distribution. *Word* 20: 360–72. [CrossRef]

Blau, Eileen Kay. 1991. More on Comprehensible Input: The Effect of Pauses and Hesitation Markers on Listening Comprehension. Paper presented at the Puerto Rico TESOL, San Juan, Puerto Rico, November 15.

Boersma, Paul, and David Weenink. 2022. *Praat: Doing Phonetics by Computer* [Computer Program]. Version 6.3.03, retrieved 17 December 2022.

Bortfeld, Heather, Silvia D. Leon, Jonathan E. Bloom, Michael F. Schober, and Susan E. Brennan. 2001. Disfluency Rates in Conversation: Effects of Age, Relationship, Topic, Role, and Gender. *Language and Speech* 44: 123–47. [CrossRef] [PubMed]

Brackhane, Fabian. 2022. Beobachtungen zu Frequenz und Funktionen von ja in deutscher Spontansprache. *Deutsche Sprache* 50: 335–63. [CrossRef]

Braun, Angelika. 1998. Voice Analysis. Paper presented at the 12th International Forensic Science Symposium, Lyon, France.

Braun, Angelika. 2020. Nonverbal vocalizations—A forensic phonetic perspective. *Proceedings Workshop on Laughter and other Nonverbal Vocalisations*, Bielefeld 2020.

Braun, Angelika. 2021. Forensische Sprach- und Signalverarbeitung. In *Handbuch des Fachanwalts Strafrecht*, 8th ed. Edited by Jan Bockemühl. Köln: Carl Heymanns Verlag, pp. 1890–914.

Braun, Angelika, and Annabelle Rosin. 2015. On the speaker specificity of hesitation markers—A pilot study. Paper presented at XVIIIth International Congress of Phonetic Sciences, Glasgow, UK, August 10–14.

Bühler, Karl. 1934. *Sprachtheorie. Die Darstellungsfunktion der Sprache*. Jena: G. Fischer.

Butcher, Andy. 1973. *Aspects of the Perception and Production of Pauses in Speech*. Arbeitsberichte of the Phonetics Institute of the University of Kiel, Nr. 1. Kiel: University of Kiel.

Candéa, Maria. 2000. Contribution à l'étude des pauses silencieuses et des phénomènes dits "d'hésitation" en fran-çais oral spontané. Etude sur un corpus de récits en classe de français. Ph.D. dissertation, Université de la Sorbonne Nouvelle, Paris, France.

Candea, M., Ioana Vasilescu, and Martine Adda-Decker. 2005. Inter- and intra-language acoustic analysis of autonomous fillers. Paper presented at the DiSS'05, Disfluency in Spontaneous Speech Workshop, Aix-en-Provence, France, September 10–12; pp. 47–51.

Clark, Herbert H., and Jean E. Fox Tree. 2002. Using uh and um in spontaneous speaking. *Cognition* 84: 73–111. [CrossRef]

Corley, Martin, and Oliver M. Stewart. 2008. Hesitation Disfluencies in Spontaneous Speech: The Meaning of *um*. *Language and Linguistics Compass* 2: 589–602. [CrossRef]

Corley, Martin, and Robert. J. Hartsuiker. 2003. Hesitation in speech can . . . um . . . help a listener understand. *Proceedings of the Annual Meeting of the Cognitive Science Society* 25: 276–81.

Corley, Martin, Lucy J. MacGregor, and David I. Donaldson. 2007. It's the way that you, er, say it: Hesitations in speech affect language comprehension. *Cognition* 105: 658–68. [CrossRef]

Defrancq, B., and K. Plevoets. 2018. Over-uh-Load, Filled Pauses in Compounds as a Signal of Cognitive Load. In *Making Way in Corpus-Based Interpreting Studies. New Frontiers in Translation Studies*. Edited by M. Russo, C. Bendazzoli and B. Defrancq. Singapore: Springer. [CrossRef]

Dejours, P., Y. Bechtel-Labrousse, P. Monzein, and J. Raynaud. 1961. Étude de la diversité des régimes ventilatoires chez l'Homme. *Journal de Physiologie* 53: 320–21.

Duez, Danielle. 1982. Silent and non-silent pauses in three speech styles. *Language and Speech* 25: 11–28. [CrossRef]

Duez, Danielle. 1993. Acoustic correlates of subjective pauses. *Journal of Psycholinguistic Research* 22: 21–39. [CrossRef]

Duez, Danielle. 2001. Acoustico-phonetic Characteristics of Filled Pauses in Spontaneous French Speech: Preliminary Results. Paper presented at the DISS'01, Edinburgh, UK, August 29–31; pp. 41–44.

Eisele, J. H., B. Wuyam, G. Savourey, J. Eterradossi, J. H. Bittel, and G. Benchetrit. 1992. Individuality of breathing patterns during hypoxia and exercise. *Journal of Applied Physiology* 72: 2446–53. [CrossRef]

Eklund, Robert, and Elizabeth Shriberg. 1998. Crosslinguistic Disfluency Modeling: A Comparative Analysis of Swedish and American English Human-Human and Human-Machine Dialogues. Paper presented at the ICSLP '98, Sydney, Australia, December 4; pp. 2631–34.
Eklund, Robert. 2000. Crosslinguistic Disfluency Modeling. A Comparative Analysis of Swedish and Tok Pisin. Human-Human ATIS Dialogues. Paper presented at Proceedings ICSLP'00, Beijing, China, October 16–20; pp. 991–94.
Eklund, Robert. 2001. Prolongations: A dark horse in the disfluency stable. Paper presented at the DISS'01, Edinburgh, UK, August 29–31; pp. 5–8.
Eklund, Robert. 2004. Disfluency in Swedish Human—Human and Human—Machine Travel Booking Dialogues. Ph.D. dissertation, Linköping University Electronic Press, Linköping, Sweden.
Erard, Michael. 2007. *Um . . . : Slips, Stumbles and Verbal Blunders, and What They Mean*. New York: Pantheon.
Fant, Gunnar, Anita Kruckenberg, and Joana Barbosa Ferreira. 2003. Individual variation in pausing. A study in read speech. *PHONUM* 9: 193–96.
Fazekas, I., and F. Liese. 1996. Some properties of the Hellinger transform and its application in classification problems. *Computers and Mathematics with Applications* 31: 17–116. [CrossRef]
Finlayson, Ian R., and Martin Corley. 2012. Disfluency in dialogue: An intentional signal from the speaker? *Psychonomic Bulletin Review* 19: 921–28. [CrossRef]
Fox Tree, Jean E. 2001. Listeners' uses of *um* and *uh* in speech comprehension. *Memory and Cognition* 29: 320–26. [CrossRef]
Goldman-Eisler, Frieda. 1961. A comparative study of two hesitation phenomena. *Language and Speech* 4: 18–26. [CrossRef]
Goldman-Eisler, Frieda. 1968. *Psycholinguistics. Experiments in Spontaneous Speech*. London and New York: Academic Press.
Harrington, Lauren, Richard Rhodes, and Vincent Hughes. 2021. Style variability in disfluency analysis for forensic speaker comparison. *International Journal of Speech Language and the Law* 28: 31–58. [CrossRef]
Henderson, Alan, Frieda Goldman-Eisler, and Andrew Skarbek. 1966. Sequential temporal patterns in spontaneous speech. *Language and Speech* 9: 207–16. [CrossRef]
Hentschel, Elke. 2011. *Funktion und Geschichte Deutscher Partikeln: Ja, doch, halt und eben*. Berlin and New York: Max Niemeyer Verlag. [CrossRef]
Hughes, Vincent, Paul Foulkes, and Sophie Wood. 2016. Strength of forensic voice comparison evidence from the acoustics of filled pauses. *International Journal of Speech, Language and the Law* 23: 99–132. [CrossRef]
Imo, Wolfgang. 2013. *Sprache in Interaktion. Analysemethoden und Untersuchungsfelder*. Berlin and Boston: De Gruyter.
Jessen, Michael. 2008. Forensic Phonetics. *Language and Linguistics Compass* 4: 671–711. [CrossRef]
Jessen, Michael. 2012. *Phonetische und linguistische Prinzipien des forensischen Stimmenvergleiches*. München: Lincom.
Kienast, Miriam, and Florian Glitza. 2003. Respiratory Sounds as an Idiosyncratic Feature in Speaker Recognition. Paper presented at the XVth International Congress of Phonetic Sciences, Barcelona, Spain, August 9; pp. 1607–10.
Kjellmer, Göran. 2003. Hesitation. In defence of ER and ERM. *English Studies* 84: 170–98. Available online: https://oyc.yale.edu/ (accessed on 6 June 2022). [CrossRef]
Kowal, Sabine. 1991. *Über die zeitliche Organisation des Sprechens in der Öffentlichkeit. Pausen, Sprechtempo und Verzögerungen in Interviews und Reden von Politikern*. Bern, Stuttgart and Toronto: Verlag Hans Huber.
Kushner, Howard I. 2012. Retraining left-handers and the aetiology of stuttering: The rise and fall of an intriguing theory. *Laterality* 17: 673–93. [CrossRef]
Lauf, Raphaela. 2001. Aspekte der Sprechatmung: Zur Verteilung, Dauer und Struktur von Atemgeräuschen in abgelesenen Texten. In *Beiträge zu Linguistik und Phonetik. Festschrift für Joachim Göschel zum 70. Geburtstag*. Edited by Angelika Braun. Stuttgart: Steiner, pp. 406–20.
Levelt, Willem J.M. 1983. Monitoring and self-repair in speech. *Cognition* 14: 41–104. [CrossRef]
Levelt, Willem J.M. 1989. *Speaking. From Intention to Articulation*. Cambridge and London: MIT Press.
Lickley, Robin J. 2015. Fluency and disfluency. In *The Handbook of Speech Production*. Edited by M. A. Redford. Hoboken: John Wiley & Sons, pp. 445–74.
Lickley, Robin, and Ellen G. Bard. 1996. On not recognizing disfluencies in dialog. Paper presented at the Intenational Conference on Spoken Language Processing, Philadelphia, PA, USA, October 3–6; pp. 1876–79.
Maclay, Howard, and Charles E. Osgood. 1959. Hesitation Phenomena in Spontaneous English Speech. *Word* 15: 19–44. [CrossRef]
McDougall, Kirsty, and Martin Duckworth. 2017. Profiling Fluency: An Analysis of Individual Variation in Disfluencies in Adult Males. *Speech Communication* 95: 16–27. [CrossRef]
McDougall, Kirsty, and Martin Duckworth. 2018. Individual patterns of disfluency across speaking styles: A forensic phonetic investigation of Standard Southern British English. *International Journal of Speech, Language and the Law* 25: 205–30. [CrossRef]
McDougall, Kirsty Richard Rhodes, Martin Duckworth, Peter French, and Christin Kirchhübel. 2019. Application of the 'Toffa' Framework to the Analysis of Disfluencies in Forensic Phonetic Casework. In *Proceedings of the 19th International Congress of Phonetic Sciences*. Edited by Sasha Calhoun, Paola Escudero, Marija Tabain and Paul Warren. Melbourne and Canberra: Australasian Speech Science and Technology Association Inc., pp. 731–35.
Meuwly, Didier. 2001. Reconnaissance de locuteurs en sciences forensiques: L'apport d'une approche automatique. Ph.D. dissertation, Université de Lausanne, Lausanne, Switzerland.

O'Connell, Daniel C., and Sabine Kowal. 2005. Uh and um revisited: Are they interjections for signaling delay? *Journal of Psycholinguistic Research* 34: 555–76. [CrossRef]
Oviatt, Sharon. 1995. Predicting spoken disfluencies during human-computer interaction. *Computer Speech and Language* 9: 19–36. [CrossRef]
Rose, Philip. 2002. *Forensic Speaker Identification*. London and New York: Taylor and Francis.
Rosin, Annabelle. 2011. Sind Häsitationslaute sprechertypisch? Master's thesis, University of Trier, Trier, Germany.
Schegloff, Emanuel A. 2010. Some other "Uh(m)"s. *Discourse Processes* 47: 130–74. [CrossRef]
Shea, Steven A., and A. Guz. 1992. Personnalité ventilatoire—An overview. *Respiration Physiology* 87: 275–91. [CrossRef]
Shea, S. A., J. Walter, K. Murphy, and A. Guz. 1987. Evidence for individuality of breathing patterns in resting healthy man. *Respiration Physiology* 68: 331–44. [CrossRef]
Shriberg, Elizabeth E. 1994. Preliminaries to a Theory of Speech Disfluencies. Ph.D. dissertation, University of California, Berkeley, CA, USA.
Shriberg, Elizabeth E. 1996. Disfluencies in switchboard. In *Proceedings of International Conference on Spoken Language Processing*. Philadelphia: IEEE, pp. 11–14.
Shriberg, Elizabeth E. 2001. To 'errrr' is human: Ecology and acoustics of speech disfluencies. *Journal of the International Phonetic Association* 31: 153–69. [CrossRef]
Shriberg, Elizabeth E., and Robin J. Lickley. 1993. Intonation of Clause-Internal Filled Pauses. *Phonetica* 50: 172–79. [CrossRef]
Stenström, Anna-Brita. 2012. Pauses and hesitations. In *Pragmatics of Society*. Edited by Andersen Gisle and Karin Aijmer. Berlin and Boston: De Gruyter Mouton, pp. 537–67.
Swerts, Marc. 1998. Filled pauses as markers of discourse structure. *Journal of Pragmatics* 30: 485–96. [CrossRef]
Tannen, Deborah. 1985. Silence. Anything but. In *Perspectives on Silence*. Edited by Deborah Tannen and Muriel Saville-Troike. Norwood: Ablex Publishing, pp. 93–112.
Trouvain, Jürgen, Bernd Möbius, and Raphael Werner. 2019. On acoustic features of inhalation noises in read and spontaneous speech. Paper presented at the 1st International Seminar on the Foundations of Speech: Breathing, Pausing, and Voice, Sønderborg, Denmark, December 1–3.
Trouvain, Jürgen, Camille Fauth, and Bernd Möbius. 2016. Breath and Non-breath Pauses in Fluent and Disfluent Phases of German and French L1 and L2 Read Speech. *Proceedings of Speech Prosody (SP8)* 31: 31–35.
Trouvain, Jürgen. 2014. Laughing, Breathing, Clicking—The Prosody of Nonverbal Vocalisations. *Proceedings of Speech Prosody*, 598–602. [CrossRef]
Wieling, Martijn, Jack Grieve, Gosse Bouma, Josef Fruehwald, John Coleman, and Mark Liberman. 2016. Variation and change in the use of hesitation markers in Germanic languages. *Language Dynamics and Change* 6: 199–234. [CrossRef]
Wolf, Jared J. 1972. Efficient acoustic parameters for speaker recognition. *Journal of the Acoustical Society of America* 51: 2044–56. [CrossRef]
Zhao, Yuan, and Dan Jurafsky. 2005. A preliminary study of Mandarin filled pauses. Paper presented at the DiSS'05, Disfluency in Spontaneous Speech Workshop, Aix-en-Provence, France, September 10–12; pp. 179–82.