
Article:
Harrington, Lauren, Rhodes, Richard William and Hughes, Vincent orcid.org/0000-0002-4660-979X (2021) Style variability in disfluency analysis for forensic speaker comparison. International Journal of Speech, Language and the Law. pp. 31-58. ISSN 1748-8885

https://doi.org/10.1558/ijsll.20214

Harrington, L., Hughes, V. and Rhodes, R. (2021) Style variability in disfluency analysis for forensic speaker comparison. *International Journal of Speech, Language and the Law* 28(1): 31-58.

1 Introduction

In the last few years, disfluency phenomena have attracted attention due to their potential usefulness in forensic speaker comparison (FSC) casework. Disfluencies, such as filled pauses and self-interruptions, are disruptions in the flow of speech caused by the speaker. They occur in the temporal domain and are unlikely to be affected by poor quality audio and bandwidth limitations in the same way that measurements such as vowel formants and fundamental frequency may be. Furthermore, they are a mostly subconscious phenomenon which might mean that they are less affected by disguise than other features of speech. A framework for empirically analysing disfluency features was developed, called the Taxonomy of Fluency Features for Forensic Analysis or 'TOFFA' (McDougall & Duckworth, 2017). Where descriptions in casework had previously been mostly impressionistic, TOFFA offers a more systematic and objective approach to the quantification of disfluency features.

The framework has been adopted by forensic analysts at J P French Associates (JPFA), the UK's largest forensic speech and acoustics laboratory, and is now used in a proportion of FSC cases. There are a number of requirements that analysts must consider before using disfluency analysis: the total duration of the sample (ideally 90-120+ seconds), the length and naturalness of the utterances, and the degree to which the samples are matched for speaking style/situation. The latter requirement comes about due to a lack of empirical data on the extent of within-speaker variation across different speaking styles; mismatched samples would only be analysed in rare cases where particularly striking disfluency behaviour is observed.

The issue of style mismatch is prevalent in FSC casework, particularly in the UK, where police interviews are the typical source of reference material, while questioned recordings include: bugged car conversations, fraudulent telephone calls, CCTV recording or voice messages, for example. Alongside different audio qualities, there will likely be differences including interlocutor dynamics, speech formality, and the topic and structure of conversations.

TOFFA is a relatively new analytical technique which, although it has been tested with controlled recordings, has not yet undergone rigorous validity testing across different speech styles and types of speakers. McDougall & Duckworth (2018) provide some evidence that individuals' disfluency profiles demonstrate a degree of consistency across different speaking styles. The study compared TOFFA features from a police interview to a phone call with an 'accomplice', and the findings of consistency seem promising for FSC casework.

The present study aims to further investigate the potential use of disfluency analysis across samples drawn from a range of situations; it will also apply TOFFA to different speakers to those used in the previous research. Variation in disfluency production at both a group-level and an individual-level is examined across three forensically-relevant tasks: a mock police interview, a paired conversation and a voicemail message; the interview represents a typical FSC reference recording, and the other conditions are analogous to FSC questioned samples.

2 Previous research

2.1 Individual variation

Between-speaker variation in disfluency usage was established as early as the 1950s. Research considered the different types of disfluencies used and the rate at which those types are used. A range of terminology has been used over the years, including "speech disturbances" (Mahl, 1957) and "hesitation phenomena" (Maclay & Osgood, 1959), but all studies have reached the same general conclusion: individual speakers differ in the types of disfluencies they produce most frequently and in the distribution of disfluency types (see section 1 in McDougall & Duckworth (2017) for a more detailed overview of this work).

2.2 Forensic application

The application of structured disfluency analysis to FSC is relatively recent; prior to this, analysis was carried out in an ad hoc manner with no general guidelines for practitioners to follow. McDougall and Duckworth (2017) analysed disfluency profiles - the frequency and types of disfluencies produced by an individual - to assess discriminatory power. The TOFFA framework (see Table 1) incorporates features from previous studies on disfluent speech in normally-fluent speakers, and speakers with speech pathologies.

Table 1: TOFFA categories and subcategories, adapted from McDougall et al. (2019).

	Subcategories and examples
Filled Pauses	● *er* **[er]** ● *erm* **[erm]** ● others, e.g. *ah* **[fpo]**
Silent Pauses	● 'grammatical' **[pg]** ● 'other' **[po]**
Repetitions	● part-word **[pwr]** *on the road I park my car th- there's* ● whole word **[wrep]** *but she- she's also* ● phrase **[prep]** *on your- on your left there's a reservoir* ● multiple (i.e. more than 2 iterations) **[mrep]** *a hairdresser at the- at the- at the- at the-*
Prolongations	● vocalic, e.g. vowel, nasal, lateral **[prov]** ● fricative **[prof]** ● plosive closure duration or affricate closure or release duration **[prop]**
Interruptions	(speaker interrupts self and discontinues the utterance, or continues with a modification) ● phrase **[pint]** *pighty road which- and then then you…* ● word **[wint]** *I thi- I probably recognise like the bar lady*

In McDougall and Duckworth (2017), disfluency profiles were analysed for 20 young adult male speakers of Standard Southern British English (SSBE); the recordings were drawn from the Dynamic Variability in Speech (DyViS) database (Nolan et al., 2009) and involved participants taking part in a mock police interview activity. Using TOFFA, disfluency rates per 100 syllables were calculated for each category. The results showed that individuals varied in how often they used disfluencies, and in which types they used. Discriminant

analysis was carried out to assess the speaker-specificity of these profiles, i.e. how successfully a speaker can be identified by their disfluency profile. All disfluency types were found to discriminate between speakers at an above-chance level; combining disfluency types achieved a classification rate of 29.4%. It was concluded that disfluency analysis could be used alongside other tools in FSC cases. The researchers state that overall disfluency profiles, rather than features in isolation, should be considered.

McDougall and Duckworth (2018) continued their research by comparing disfluency profiles across different speaking situations. Two simulated tasks from DyViS were compared: the first was a mock police interview in which participants were provided with visual prompts on a computer screen and instructed to deny all knowledge about certain details. The second task was a telephone call with an 'accomplice' in which participants were asked to discuss information about the mock crime from the previous task. In both cases, the interlocutor was one of three DyViS researchers. Based on rates per 100 syllables, no significant differences were found between the styles. Many speakers remained relatively consistent across the tasks and significant correlations were found between all but three features. Some features performed better than others, e.g. **[erm]** and **[wrep]**, while less frequently-occurring features generally resulted in lower correlations. McDougall and Duckworth discussed the speaker-specificity of the profiles and the consistency across speech styles, highlighting their potential in FSC cases where recordings differ in style.

2.3 Situational variation

The findings of McDougall and Duckworth (2018) suggest that disfluency profiles are likely to remain consistent across styles. However, previous literature suggests otherwise: that a speaker's disfluency usage differs according to factors such as the interlocutor, the task at hand or their state of anxiety.

2.3.1 Interlocutor

The presence or absence of an interlocutor has been shown to have an effect on disfluency usage. Monologues have consistently been found to be less disfluent than dialogues. Moniz et al. (2014) investigated disfluency production in conversational dialogue during a map task and in monologic university lecture recordings. A higher percentage of disfluency phenomena were observed in the conversational speech with double the amount of fragments (i.e. truncated or incomplete linguistic material), though fewer deletions than the lecture speech. Broen and Siegel (1972) analysed rates in three tasks: a monologue, a monologue spoken as though to TV cameras or an audience, and a conversation. The highest rates of disfluencies were found in the conversation task, while both of the conditions without an interlocutor showed similar rates. Oviatt (1995) also compared three tasks: a face-to-face dialogue, a phone dialogue, and a monologue. The monologue task was found to be the least disfluent of the three, while the telephone conversation was the most disfluent speaking situation.

The physical presence of an interlocutor also has an influence on disfluency usage, with generally slightly higher rates when the interlocutor is not visible. Oviatt (1995) found that a telephone dialogue had higher overall disfluency rates than a face-to-face dialogue. Kasl and Mahl (1965) carried out interviews with participants who were encouraged to talk freely about themselves. For half of the interview the interviewer was in the same room and for the other half they were in the adjacent room, communicating remotely. The portion of the

interview where the interlocutors were in separate rooms yielded higher rates of filled pauses. Branigan, Lickley and McKelvie (1999) investigated the effect of eye contact by carrying out collaborative map tasks, one where participants could see each other and another when a screen was installed between participants. Though overall rates remained similar across the tasks, a much higher number of repetitions were produced in the no-eye contact condition.

2.3.2 Task

The task being performed also influences the use of disfluencies. Lickley (2001) investigated how the distribution of disfluency phenomena varies across speech containing instructions and speech containing answers. When performing a collaborative map task, participants produced more hesitations and self-corrections when giving instructions and more pauses and repetitions when providing answers. Lickley also observed that more disfluencies were produced if an answer was difficult to find or formulate.

Similarly, the topic of conversation has been shown to affect disfluency usage. Unfamiliar or anxiety-inducing topics tend to have higher overall disfluency rates. Bortfeld, Leon, Bloom, Schober and Brennan (2001) conducted research involving picture-matching tasks, one with photos of children and the other with a series of monochromatic geometric forms called tangrams. These topics represent familiar and unfamiliar domains, respectively. Participants exhibited higher overall rates when discussing the tangrams, mostly due to a rise in the use of repeats and restarts. It was also found that participants used more fillers in the task involving the familiar domain.

Siegman and Pope (1965) also investigated the topic of conversation, focussing on the effect of anxiety. It was found that the use of all disfluency categories (except filled pauses) increased significantly in the high anxiety-inducing portions of the interview, and similar findings are reported in Kasl and Mahl (1965). Furthermore, Bello (2006) found that situational conflict (in which participants were faced with two undesirable options: outright truth or outright deceit) had a significant impact on overall disfluency usage. Participants were more disfluent in situations of conflict than non-conflict, although once again filled pauses were found to behave in a unique way, such that the highest number of filled pauses occurred in the formal ‘non-conflict’ condition.

2.4 Summary

Results consistently demonstrate varying disfluency behaviour across speech situations, and most of these situational differences are applicable to FSC: one might expect heightened levels of anxiety or conflict in an interview situation (and indeed in some criminal situations), and almost always a different topic of conversation to that of the questioned sample(s). The interlocutor dynamic is also likely to differ across samples, in relation to both the power-relationship between speakers and whether the interlocutor is physically present; for example, if the questioned recording is a telephone call or voice message.

Overall, there are many ways in which the situation may affect the use of disfluencies and it therefore seems unlikely that disfluency profiles will remain consistent across different tasks. It should be noted that while the tasks used in McDougall and Duckworth’s (2018) study are nominally different, there are some similarities: both consist of a roleplay about the same mock crime, fairly detailed prompts were provided, and the interlocutor dynamic is also similar across tasks, with the ‘police officer’ and ‘accomplice’ roles played by DyViS

researchers. The resemblance between the two tasks could be responsible for the finding that speakers' disfluency production remains relatively stable. The results must therefore be interpreted with caution when related to forensic casework, where samples are likely to have been recorded in much more divergent settings.

3 Research focus

The present study aims to evaluate how well disfluency analysis performs as a method for FSC, and when it should be employed. It builds on McDougall and Duckworth (2018) in a number of key ways: it involves a wider range of tasks with distinct interlocutor dynamics, the speakers are from West Yorkshire rather than the South East of England, and it will apply closer comparisons of individual speakers' disfluency profiles in the same way that FSC analysts might for speaker profiles. Furthermore, as voice messages (by phone or social media apps) are becoming increasingly common as questioned samples in FSC, the present study will assess this type of recording situation. This is important for FSC practitioners as an understanding of disfluency behaviour in different situations is vital for a) choosing whether to use disfluency analysis and b) interpreting the results of disfluency analysis; if practitioners wrongly assume features will remain stable in different situations, it may lead to incorrect conclusions.

The overarching aim is to test the validity of the TOFFA framework, since this is currently in use in casework in the UK. The validation of methods in forensic science is currently a requirement of the UK Forensic Science Regulator's (FSR) guidance, where laboratories are required to gain accreditation to the laboratory standard ISO/IEC 17025 (2017) and the FSR *Codes of Practice and Conduct* (2021). Both standards introduce strict requirements for methods to be validated in order for evidence to be used in the criminal justice system, even when such methods are considered standard and in widespread use.

We pose two main research questions:

1) At a group level, to what extent is disfluency production consistent across the tasks?
2) At an individual level, to what extent are speakers' disfluency profiles stable across the tasks?

4 Methodology

4.1 Framework

This study adopted the TOFFA framework as described by McDougall and Duckworth (2017, 2018) with some modifications. It comprises five main categories (such as Filled Pause or Repetition), each with multiple sub-categories (such as **[erm]** and **[wrep]**). Some features of the framework were adapted for this study. A threshold of 200ms is used for prolongations according to TOFFA. For the current study this was raised to 250ms for diphthongs since many natural-sounding diphthongs would have otherwise been defined as disfluent. Grammatical pauses were excluded from analysis in this study due to doubts about their status as a disfluency phenomenon. A time-basis was used to calculate rates rather than a syllable-basis; disfluency counts were calculated per minute as a time-based method has proved to be more efficient while providing similar results, and this has since been adopted by the creators of TOFFA (McDougall et al., 2019).

The modified categories and subcategories are presented below (see section 2 of McDougall & Duckworth (2017) for more detail).

Table 2: Modified TOFFA categories and subcategories used in present study.

	Subcategories and examples
Filled Pauses	● *er* **[er]** ● *erm* **[erm]** ● others, e.g. *ah* **[fpo]**
Unfilled Pauses	● 'non-grammatical' [**po]**
Repetitions	● part-word **[pwr]** *it's r- really cool* ● whole word **[wrep]** *one of my- my mates* ● phrase **[prep]** *I've been doing it- doing it here* ● multiple (i.e. more than 2 iterations) **[mrep]** *it's it's it's one of the more famous accents*
Prolongations	(duration > 200 msecs) ● vocalic: monophthongal vowel, nasal, lateral **[prov]** ● fricative **[prof]** ● plosive closure duration or affricate closure/release duration **[prop]** (duration > 250 msecs) ● vocalic: diphthongal vowel **[prov]**
Interruptions	(speaker interrupts self and discontinues the utterance, or continues with a modification) ● phrase **[pint]** *you see a- you walk past a group* ● word **[wint]** *you can get pi- scouted to play*

4.2 Data

Recordings of 20 young adult (aged between 18-30) male speakers from Kirklees, West Yorkshire were analysed. These were selected from the West Yorkshire Regional Accent Database 'WYRED' (Gold, Ross & Earnshaw, 2018). Three recordings were analysed for each speaker, each from a separate task. The first task analysed was a mock police interview which replicated the methodology used in the DyViS project, whereby participants were interrogated by a 'police officer' (played by Research Assistant 1) about a mock crime they had committed. Before the interview began, participants were provided with a slideshow containing background information on the role they would be playing and the 'crime' they had committed. During the interview, slides appeared on screen containing information about the events in question, and participants were advised to be cooperative but to deny or avoid mentioning any incriminating facts which were shown in red (Nolan et al., 2009).

The second task was a paired conversation where participants were coupled up, advised to act as if they were having a conversation with a friend and to talk freely about any topic for twenty minutes. Prompt cards containing questions on a range of topics (e.g. work, hobbies, education, etc.) were provided and participants could use these if desired. The third task was a voicemail message for which participants were instructed to leave a message for their brother 'John', requesting that he hide or destroy evidence relating to the aforementioned mock crime and contact their 'accomplice' immediately. Participants were provided with a timer and advised that their message should be around 2 minutes in length. They were also given a list of four bullet-point examples of pieces of evidence to be hidden or destroyed.

Recordings for the interview and conversation were roughly 20 minutes in length, and samples of 4-5 minutes from each speaker were extracted. The recordings had previously been orthographically transcribed in a TextGrid in Praat (Boersma & Weenink, 2019). A Praat script was designed to extract and concatenate parts of the recording where the aligning interval tier contained text. Recordings for the voicemail task were between 1 and 3 minutes, containing only the participants' speech. Praat was used to examine each audio file and mark the disfluencies using a TextGrid, and TOFFA guidelines were followed as closely as possible.

Due to the constraints of the present study, we were unable to employ a second independent auditor and therefore provide inter-annotator reliability scores. However, each file was coded twice by the first author to ensure consistency throughout. As a result of some subjectivity within the framework, procedures were developed throughout the first pass of coding to deal with complex situations. The second pass took place immediately after the first and some changes were made to the earlier files where there had been some confusion, e.g. between repetitions and interruptions. For example, the utterance 'footpith footpath' was originally coded as a repetition but it was later decided that this should be marked as an interruption since the speaker "changes what had been started" (McDougall & Duckworth, 2017, p. 20).

The number of coded disfluency features were calculated, and a count for each subcategory and the overall total was calculated. For the interview and conversation tasks, the samples were between 4 and 5 minutes in length, meaning that there were 4 full one-minute sets for each speaker. Speaker 15 was the only exception to this, as their sample for the conversation was between 3 and 4 minutes and therefore only contained 3 one-minute sets. For the voicemail task, speakers had either 1 or 2 one-minute sets depending on the length of the sample; at least 90 seconds of speech was elicited from all participants in this task, excluding Speakers 9 and 26 whose voicemail messages were 66 and 79 seconds in length respectively. Using the one-minute sets, an average count per minute of all subcategories, main categories and the overall total was calculated for each speaker and for each task.

4.3 Statistical analysis

4.3.1 Linear mixed effects models

To investigate the effect Task had on disfluency rates, linear mixed effects models were fitted for each of the following disfluency categories: filled pause, unfilled pause, repetition, prolongation, interruption and overall total. The data input consisted of the one-minute sets described in section 4.5, however for the voicemail task, only 1 one-minute set was used for each speaker. Models were created in R studio using the lme4 (Bates, Mächler, Bolker & Walker, 2015) and lmerTest (Kuznetsova, Brockhoff & Chistensen, 2017) packages. In each model, Task was included as a fixed effect and, in order to account for the fact that some participants may behave differently from the average trend and that disfluency counts may vary randomly from minute to minute within a sample, Speaker and Minute were included as random effects.

A separate "null" model for each disfluency category was fitted, identical to the corresponding full model except for the exclusion of the fixed effect, Task. Each full model was then compared with the null model using Chi-Square tests with the anova function in R. Model comparison indicates whether the full model (i.e. the one including Task) provides a better fit than the null model, and is therefore better at accounting for the variability in the

data. The results of the comparison include a *p*-value which can be used to provide a measure of statistical significance for the variable of interest.

Each model conducted a three-way comparison across the tasks. The interview was used as a baseline condition when compared with the conversation and voicemail, and the conversation was used as a baseline in the conversation-voicemail comparison. Summary output containing an Estimate, Standard Error rate, and a *p*-value for each model was examined. The Estimate indicates the direction and magnitude of the effect, such that an Estimate of +3 suggests that a participant would produce on average 3 more disfluencies per minute in Condition 2 compared with Condition 1. A threshold of $\alpha = 0.05$ was used to determine statistical significance.

4.3.2 Individual profile analysis

Three speakers displaying extremely above- or below-average rates of a particular category were chosen for specific analysis. For each speaker, a disfluency summary was produced in the style of that used by forensic analysts, containing side-by-side profiles with one displaying average frequency per minute of each TOFFA subcategory and another demonstrating proportions of each main category. The profiles within each summary were visually compared across Task, and average frequencies and proportions of the overall total were considered, to evaluate relative stability of disfluency features.

5 Results

The consistency of the group's disfluency production across tasks will be considered in section 5.1 and the stability of individuals' disfluency profiles will be considered in section 5.2.

5.1 Disfluency use across tasks

Summary finding: Every disfluency category was found to be significantly affected by Task.

Table 3: Results of model comparison for each disfluency category. Significant results are highlighted with green text.

Disfluency category	ChiSq	Df	*p*
Overall	16.069	2	0.000324
Filled pause	129.96	2	<0.0001
Unfilled pause	29.227	2	<0.0001
Repetition	36.391	2	<0.0001
Prolongation	16.029	2	0.0003306
Interruption	76.787	2	<0.0001

Table 3 provides an overview of model comparisons between full and null models for each disfluency category. For all categories, model fit was significantly improved when including Task as a fixed effect. The strongest effects were found for filled pauses ($\chi^2 = 129.96$; $p < 0.0001$) and interruptions ($\chi^2 = 76.787$; $p < 0.0001$).

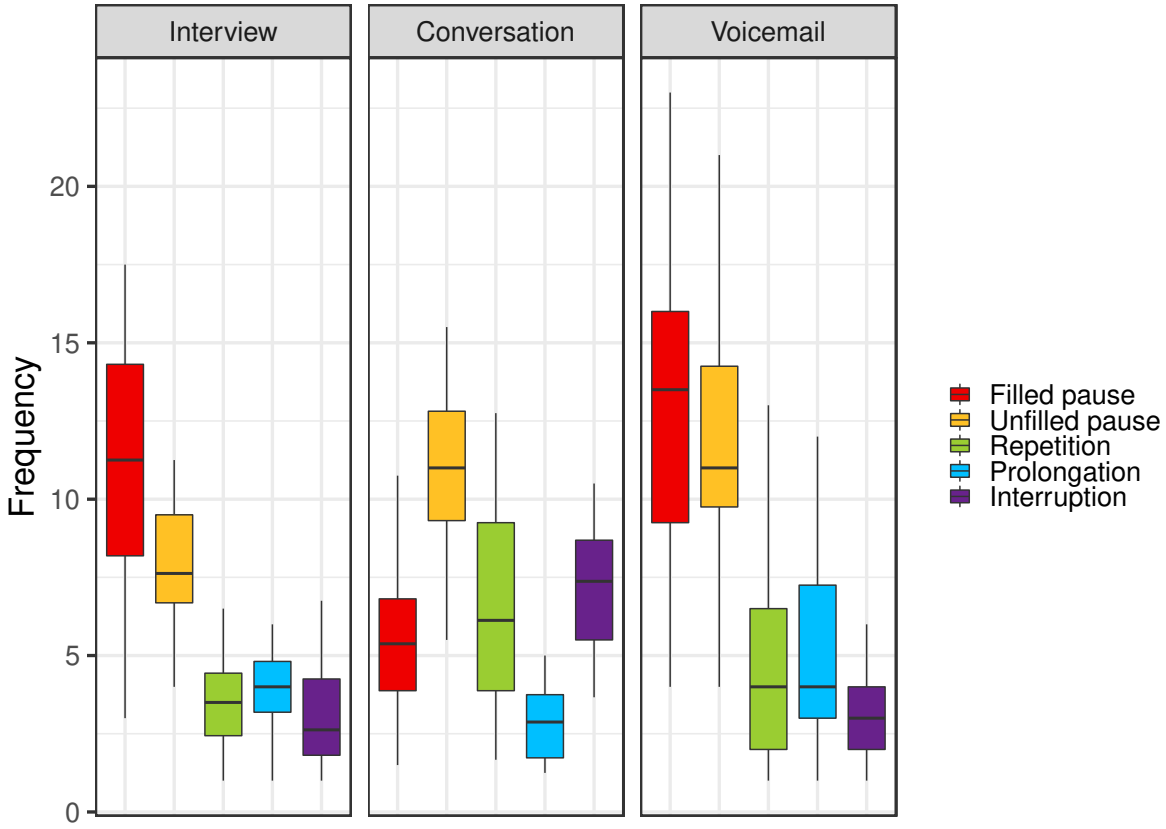

Figure 1: Summary of speakers' average production per minute of each TOFFA category within all three tasks.

Figure 1 shows the distributions of average disfluency counts of the main TOFFA categories for each speaker across the tasks. Having found that Task does have a significant effect on disfluency usage, a three-way comparison of the interview, conversation and voicemail was conducted. The differences across tasks were investigated using the output of each full model summary, focusing particularly on the direction and magnitude of the differences.

5.1.1 Interview vs Conversation

Summary finding: All disfluency categories were significantly affected by Task. Not only were overall rates different, but the distribution of the TOFFA categories used across the tasks also differed.

The average overall rate in the conversation was 32.8 disfluencies per minute compared with 30.0 per minute in the interview; this difference was found to be significant ($\beta = 2.86$, SE = 0.97, $p < 0.005$). Examination of the summary output for the model indicated that disfluency counts of every category were in fact significantly different in these tasks. The most striking difference between the tasks was found in filled pause usage. There were on average 6.15 fewer filled pauses per minute in the conversation than in the interview ($\beta = -6.15$, SE = 0.45, $p < 0.0001$). Another prominent difference was the much more frequent use of interruptions in the conversation, where the count per minute was on average 4.38 higher than in the interview ($\beta = 4.38$, SE = 0.47, $p < 0.0001$).

Unfilled pauses and repetitions were significantly more frequent in the conversation than in the interview, with an estimated 2.72 more unfilled pauses ($\beta = 2.72$, SE = 0.53, $p < 0.0001$) and 2.83 more repetitions ($\beta = 2.83$, SE = 0.44, $p < 0.0001$) per minute in the conversation. Lastly, prolongations were used significantly less frequently in the conversation than in the interview, with prolongation counts per minute being on average 1.11 higher in the interview ($\beta = -1.11$, SE = 0.33, $p < 0.005$)

5.1.2 Interview vs Voicemail

Summary finding: With the exception of overall rates and unfilled pauses, no significant effects of Task were observed in this comparison, though raw frequencies were generally higher in the voicemail.

The voicemail task was found to be significantly more disfluent than the interview, with an average overall rate of 35.1 disfluencies per minute compared with only 30 per minute in the interview ($\beta = 6.14$, SE = 1.60, $p < 0.0005$). Interestingly, almost all TOFFA categories exhibited more frequent usage in the voicemail task, although many of the differences in usage between the tasks were not significant. Only unfilled pause counts were found to be significantly different across the tasks, with an average of 3.37 more unfilled pauses per minute in the voicemail ($\beta = 3.37$, SE = 0.84, $p < 0.0005$). Filled pauses, repetitions and prolongations were all produced slightly (but not significantly) more frequently in the voicemail, while interruptions were produced slightly (but not significantly) more frequently in the interview.

5.1.3 Conversation vs Voicemail

Summary finding: All categories except unfilled pauses were found to significantly differ across the tasks. As with the interview-conversation comparison, overall rates as well as the distribution of TOFFA categories within the tasks were different.

Overall disfluency rates were higher in the voicemail, with a group average of 35.1 disfluencies per minute compared with 32.8 per minute in the conversation, and this difference was found to be significant ($\beta = 3.27$, SE = 1.61, $p = 0.045$). The use of almost every TOFFA category was significantly affected by Task. The only exception was unfilled pauses which were produced slightly (but not significantly) more frequently in the voicemail. The remaining four TOFFA categories showed significantly different usage across the tasks, with two categories produced more frequently in the voicemail and two categories produced more frequently in the conversation. Filled pauses were produced at a strikingly higher rate in the voicemail, with count per minute being an estimated 6.66 higher than in the conversation ($\beta = 6.66$, SE = 0.77, $p < 0.0001$). Prolongations were also more frequent in the voicemail task which contained on average 1.76 more prolongations per minute ($\beta = 1.76$, SE = 0.52, $p < 0.005$).

In the opposite direction, interruptions and repetitions were significantly less frequent in the voicemail than in the conversation. A prominent difference was observed in the production of interruptions, where count per minute in the voicemail task was on average 4.87 lower ($\beta = -4.87$, SE = 0.78, $p < 0.0001$). Less marked but still significant was the difference in Repetition counts, with an estimated 1.62 fewer repetitions per minute compared with the conversation task ($\beta = -1.62$, SE = 0.68, $p = 0.018$).

5.1.4 General trends: filled pauses and interruptions

Summary finding: Filled pauses were significantly less frequent in the conversation than in the other tasks, and interruption counts were significantly higher in the conversation than in the other tasks. No significant difference in usage of either category was found between the interview and voicemail.

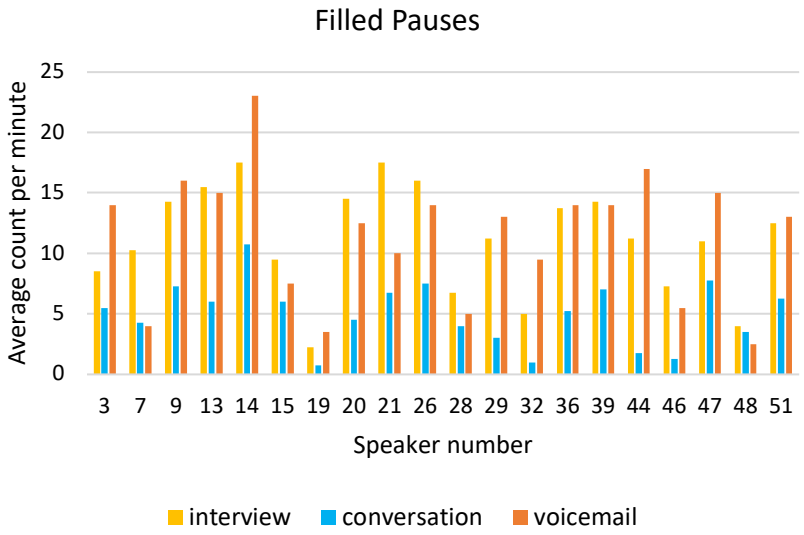

Figure 2: Average count per minute of filled pauses for each speaker.

A number of general trends were observed across the three tasks. Figure 2 shows the average count per minute of filled pauses for each speaker (speaker numbers here refer to the numbers assigned to speakers within the WYRED database). For all but two speakers, this category was much less frequent in the conversation than in the other two tasks. The group average in the conversation was 5.0 filled pauses per minute which was significantly lower than the group averages of 11.1 in the interview ($\beta = -6.15$, SE = 0.45, $p < 0.0001$) and 11.4 in the voicemail ($\beta = 6.66$, SE = 0.77, $p < 0.0001$). No significant difference was found between filled pause production in the interview and voicemail tasks.

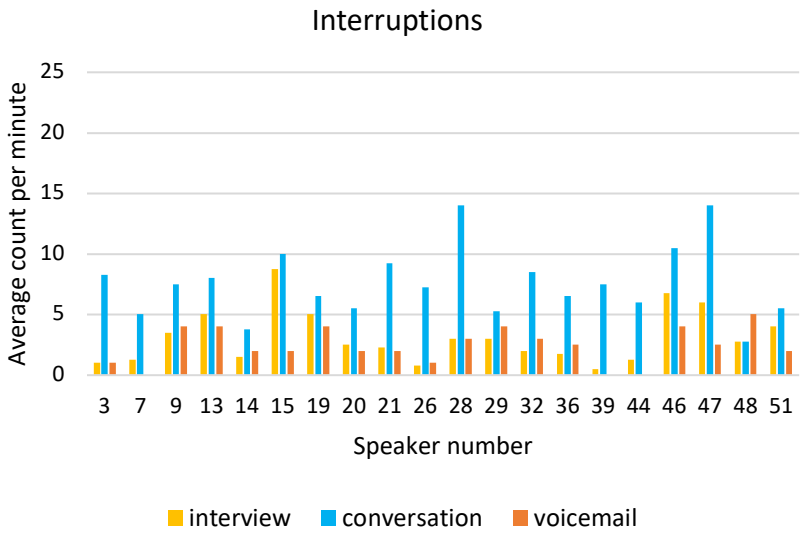

Figure 3: Average count per minute of interruptions for each speaker.

Figure 3 shows another clear pattern observed, which is the increased use of interruptions in the conversation compared with the other two tasks. The average count per minute in the conversation was 7.6 interruptions, which was significantly higher than the averages of 3.1 per minute in the interview ($\beta = 4.38$, SE = 0.47, $p < 0.0001$) and 2.4 per minute in the voicemail message ($\beta = -4.87$, SE = 0.78, $p < 0.0001$). No significant difference was found between use of this disfluency category in the interview and voicemail.

5.2 Disfluency profiles across task

In FSC casework, analysts are interested in the profiles of individuals, and while general trends can be informative, they do not reveal much about the stability of each speaker's disfluency production. In this section, we will consider the stability of individuals' disfluency profiles across the tasks by examining the range of within-speaker variation exhibited, and by carrying out individual profile analysis.

5.2.1 Within-speaker variation

Figure 4 shows the standard deviation of each speaker's average overall disfluency rate across the three tasks. Some individuals displayed relatively stable overall rates in the interview, conversation and voicemail, such as Speaker 7 who produced 24.25, 25.75 and 23.00 disfluencies per minute in the three tasks respectively. Others showed much more variability, such as Speaker 28 who produced 25.00, 41.75 and 31.00 disfluencies per minute in the three tasks.

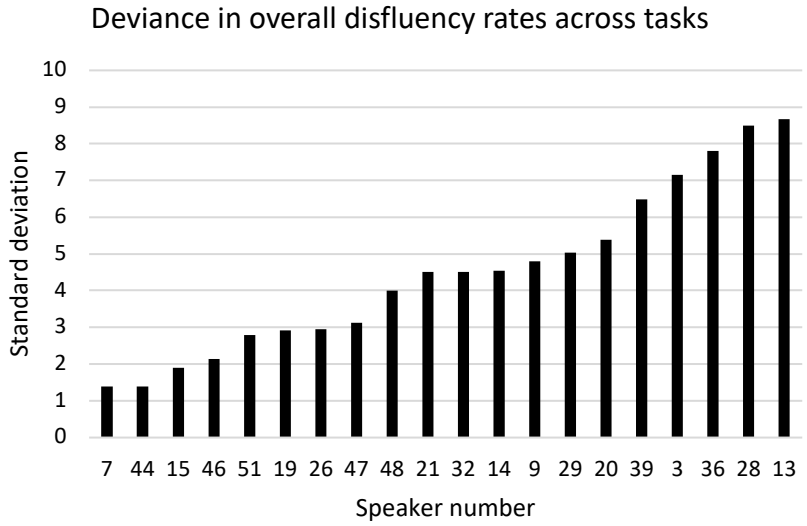

Figure 4: Standard deviation measurements for each speaker's average overall disfluency count per minute in all tasks. Participants have been ordered from low to high.

Two participants were chosen to highlight varying rates of speaker-stability within the TOFFA categories. Figure 5 shows disfluency counts in each task for Speaker 19 who displayed relatively stable production of most categories. With the exception of unfilled pauses which were produced considerably more frequently in the voicemail than in the other tasks, Speaker 19's production proved to be fairly consistent across the interview,

conversation and voicemail. This speaker follows the general trends observed in section 5.1.3, producing fewer filled pauses and more interruptions in the conversation; considering that disfluency counts of these categories are relatively stable across the interview and voicemail, the overall trends could account for why the speaker doesn't produce a consistent account across all three tasks for filled pauses and interruptions. In this sense, then, this speaker is quite predictable from group norms.

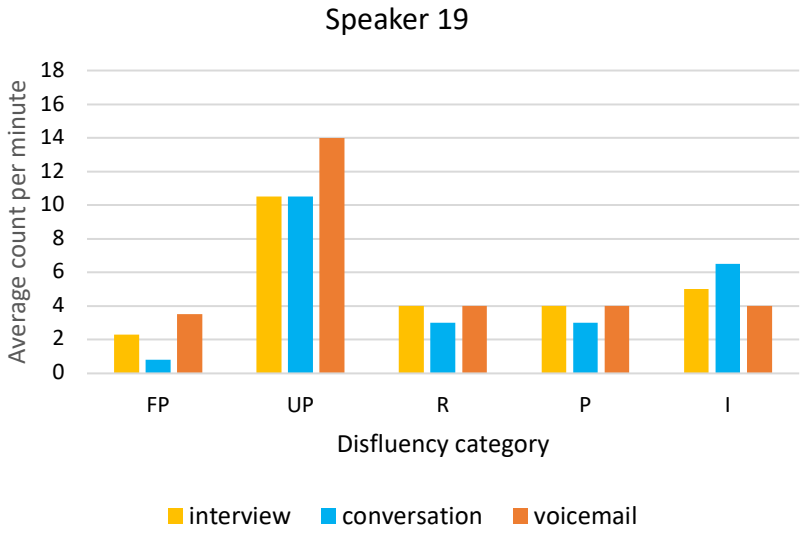

Figure 5: Average count per minute of each TOFFA category for Speaker 19.

Speaker 3, on the other hand, demonstrates a much higher rate of between-task variability, as seen in Figure 6. Their production of filled pauses and repetitions varies considerably across all three tasks, and unfilled pause counts are also fairly unstable across the tasks, though a weak similarity is found between the conversation and voicemail.

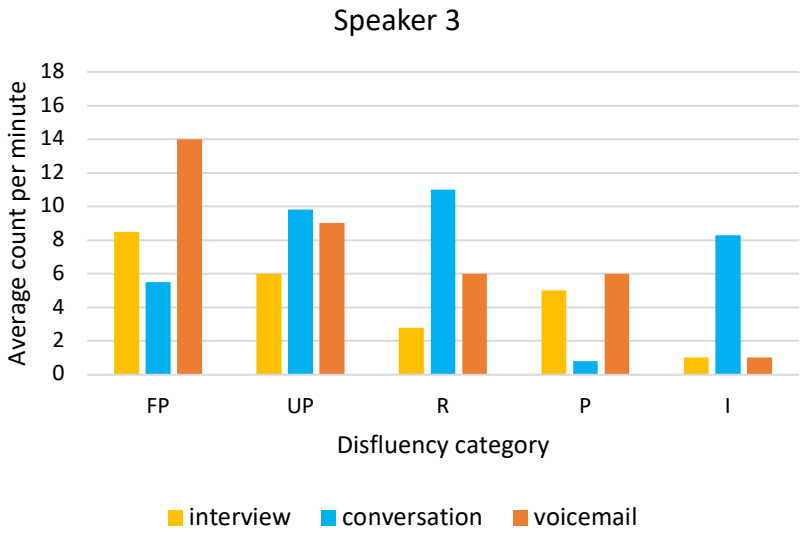

Figure 6: Average count per minute of each TOFFA category for Speaker 3.

Similar frequencies of prolongations and interruptions are observed within the interview and voicemail, although compared with these tasks, Speaker 3 produced a much lower number of

prolongations and a much higher number of interruptions in the conversation. Overall there is no clear stability of any category across all of the tasks, and the variation is not predictable.

5.2.2 Individual profile analysis

Forensic analysts do not use disfluency analysis in all cases. It is more likely that TOFFA analysis will be used if there is an 'interesting' or 'unusual' disfluency behaviour exhibited; this may be a qualitative judgement made by the analyst based on their experience or knowledge of the variety spoken in the samples. To replicate this, three speakers displaying extremely above or below average rates of a particular disfluency category have been chosen for individual profile analysis. This involves detailed analysis of disfluency profiles to investigate if and how behaviour varies across tasks. Each profile is made up of two bar graphs, of which the top graph represents raw frequencies of each TOFFA subcategory. The lower graph converts the raw frequencies to proportions of the overall disfluency total within each task and combines the subcategories into the five main categories. This method allows for a more detailed overview of the speaker's disfluency production, as both the frequency of each disfluency type and its distribution is presented.

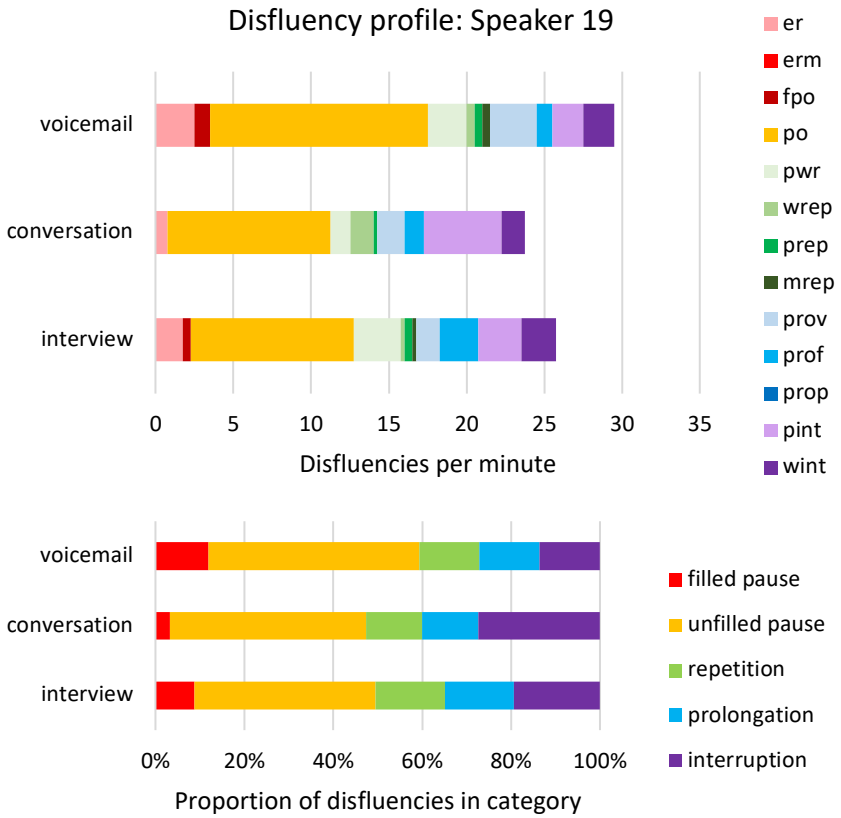

Figure 7: Disfluency summary for Speaker 19.

Figure 7 shows a disfluency summary for Speaker 19. This speaker consistently produced very low rates of filled pauses (shown in shades of red), averaging 2.3, 0.8 and 3.5 per minute in the interview, conversation and voicemail; these averages are considerably lower than the group averages of 11.1, 5.0 and 11.4 respectively. With the exception of the voicemail, where Speaker 48 produced an average of 2.5 filled pauses per minute, Speaker 19 consistently produced the lowest number of filled pauses within the tasks. Another striking feature is the

absence of **[erm]** from this individual's speech since they are the only speaker to not produce this type of filled pause at all within their sample.

Speaker 19 demonstrates fairly stable production of filled pauses in the interview and voicemail, and while the frequency per minute is considerably lower in the conversation, this pattern is to be expected given the general trend for decreased filled pause production in this task. However, when considering filled pauses as a proportion of overall disfluency production, considerable variation is observed between the tasks; filled pauses represent 8.7% of all disfluencies in the interview, as compared with 3.2% in the conversation and 11.9% in the voicemail.

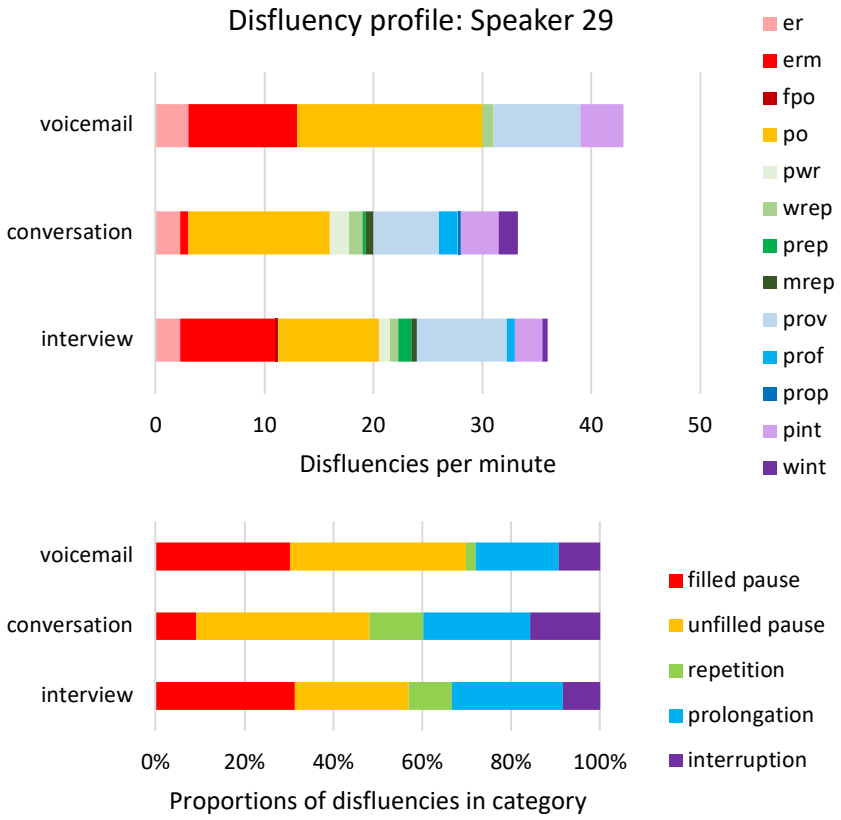

Figure 8: Disfluency summary for Speaker 29.

Figure 8 contains a disfluency summary for Speaker 29, who consistently produced very high frequencies of prolongations. Average counts of 9.0, 8.0 and 8.0 prolongations per minute were observed for this speaker, which are all much higher than the group averages of 3.9, 2.7 and 4.6. It is clear from the distribution of TOFFA subcategories that this is mainly due to the frequent production of prolonged vocalic segments. In the interview and conversation, this individual produced the highest number of prolongations per minute within the group, but in the voicemail there were three speakers who produced between 8.5 and 9.5 prolongations per minute (Speakers 13, 15 and 21); however, these speakers did not produce a consistently high frequency of prolongations across all three tasks. Regarding distribution of disfluency types within the profiles, prolongations represented a very similar proportion of the overall disfluency production in the interview (25.0%) and conversation (24.1%), while in the voicemail, prolongations made up a smaller percentage of the overall disfluency production (18.6%).

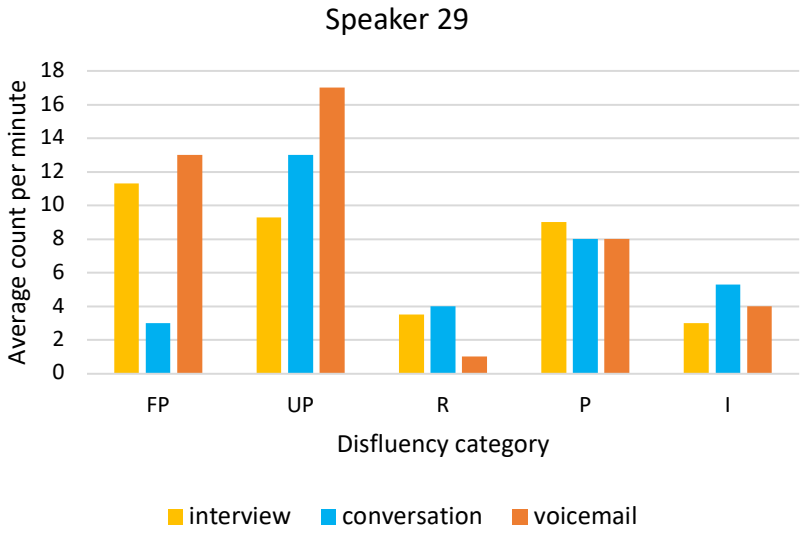

Figure 9: Average count per minute of each TOFFA category for Speaker 29.

Figure 9 shows the average count per minute of each TOFFA category within the three tasks. The only disfluency type which demonstrates some stability across the tasks is prolongations, the category which this speaker produces at an above-average rate. Counts per minute of all other types vary considerably according to Task, particularly for filled pauses and unfilled pauses.

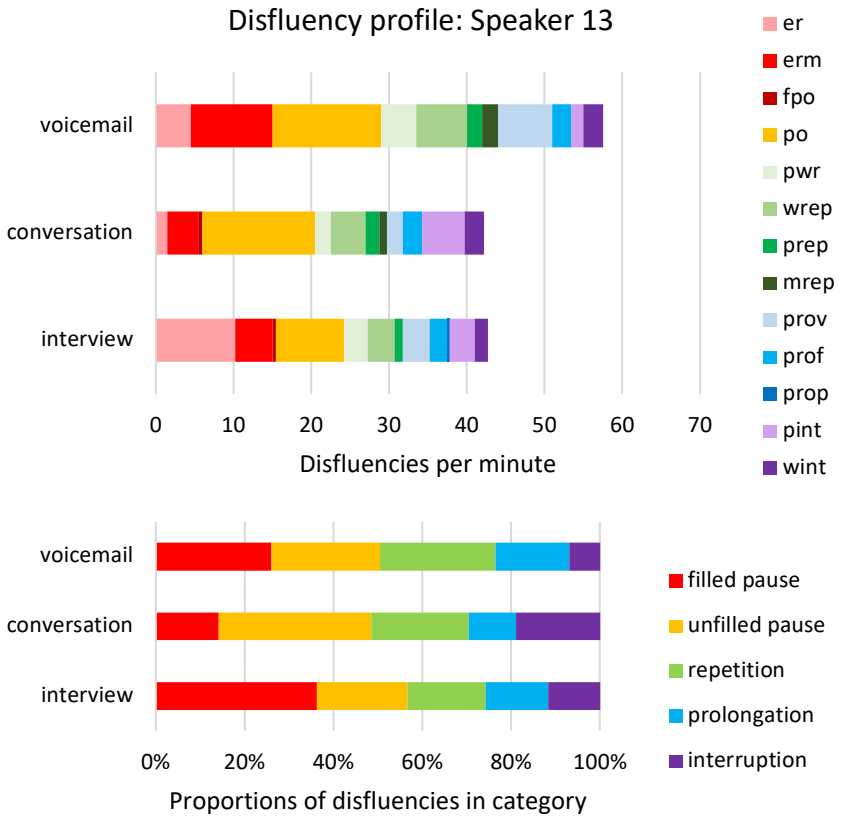

Figure 10: Disfluency summary for Speaker 13.

Finally, Figure 10 shows a disfluency summary for Speaker 13, who exhibited the highest use of repetitions in the voicemail with an average count of 15.0 repetitions per minute, almost three times the group average of 5.1. This individual also demonstrated the highest rate of Repetition production in the interview task, producing 7.5 per minute, over double the group average of 3.6. Despite producing the highest numbers of repetitions within the group for both of these tasks, the rate of production did not remain stable, with the voicemail containing twice as many repetitions as the interview. Furthermore, although Speaker 13 produced a slightly above average frequency in the conversation, with 9.3 repetitions per minute as compared with the group average of 6.5, there were 4 participants who produced a higher frequency of repetitions in this task. While the previous speakers demonstrated that 'extreme' disfluency behaviour can be preserved across different tasks, Speaker 13 shows that this is not always the case. This speaker's disfluency production also demonstrates that inconsistent frequencies and proportions across tasks does not indicate that the speech must have been produced by different speakers.

With regard to proportions of overall disfluency production, repetitions represented roughly 20% of the total disfluencies in all tasks: 17.5%, 21.9% and 26.1% in the interview, conversation and voicemail respectively. No relative stability was observed for any other disfluency type across all three tasks.

6 Discussion

The central question in this study is whether Task affects disfluency production. The results of this study overwhelmingly indicate that it does. This section will address each research question in turn and compare the current findings with those of McDougall and Duckworth (2018). Finally, some recommendations for practice and future research are made.

6.1 Disfluency production across tasks

RQ1: On a group level, to what extent is disfluency production consistent across tasks?

The results of model comparisons revealed that every disfluency category is significantly affected by Task. A three-way comparison across tasks of each category was then carried out, and revealed that the interview and voicemail were similar in many ways. Overall disfluency count was significantly higher in the voicemail, as was the frequency of unfilled pauses; all other disfluency categories were not found to be significantly affected by Task, though the raw frequencies were generally higher in the voicemail. These findings suggest that the increased use of unfilled pauses in the voicemail is responsible for the significantly different overall disfluency production between that task and the interview.

This finding was rather unexpected, given that there was no interlocutor in the voicemail task, and numerous previous studies have found that the absence (either physical or total) of an interlocutor can significantly influence disfluency production. Monologue speech has generally been found to be less disfluent than conversational (i.e. dialogue) speech (Broen & Siegel, 1972; Oviatt, 1995), and hence it was predicted that the interview would be more disfluent than the voicemail. A review of the activities involved in the study reveals a number of similarities between the interview and voicemail tasks which could have overridden the difference in interlocutor. Firstly, both tasks contained semi-spontaneous speech as the result of visual prompts presented throughout the task. Though instructed to answer as naturally as possible, due to the prompts it was not necessary for participants to spontaneously formulate ideas for their answers or requests in the mock police interview and voicemail message

respectively. Furthermore, both of these tasks centred on the same topic, i.e. a mock crime that the participant had committed. It has previously been found that the topic of conversation can have an effect on disfluency production (Bortfeld et al., 2001), therefore this could also have played a role in the similarity between tasks.

The similarity between the interview and voicemail tasks most likely contributed towards two general trends observed in the data; speakers produced significantly lower frequencies of filled pauses and significantly higher frequencies of interruptions in the conversation as compared with the other tasks, in which production did not differ significantly. The much higher number of interruptions in the conversation is likely a result of this task involving fully spontaneous speech, where participants were only prompted by question cards if necessary. Compared with the interview and voicemail, where specific content was discussed, the conversation task required participants to talk freely about any desired topic, e.g. holidays, interests, political news. Because there were no guidelines on what had to be mentioned, speakers were free to reformulate their utterances or change topic whenever desired, including in the middle of a word or phrase. In the mock police interview, this could have come across as "acting suspiciously" and in the voicemail there was a specific amount of information to be conveyed within a short amount of time. This trend was observed for 18 of the 20 participants, with the remaining speakers (Speakers 9 and 48) producing more interruptions in the voicemail instead.

The conversation task also elicited lower use of filled pauses compared to the other tasks. Many previous studies have demonstrated higher frequencies of filled pauses when discussing familiar topics (Bortfeld et al., 2001), yet in this study the tasks involving a much less familiar topic (a mock crime) were found to induce significantly higher rates. However, these findings fall in line with previous research on conflict where an informal 'non-conflict' condition elicited the lowest number of filled pauses (Bello, 2006). The pseudo-crime tasks (interview and voicemail) could be interpreted as both involving a degree of situational conflict whereby the speaker must choose between truthfulness or deceit when neither option is desirable; for example, in the police interview, participants were advised to be cooperative but also to conceal certain facts. There were many other factors present which may have interfered with filled pause production. The question-and-answer format of the interview could be responsible for the higher use of filled pauses in that task, as many speakers produced a filled pause at the beginning of almost every utterance, possibly in order to gain time to formulate answers, as suggested by Lickley (2001). Speakers may have also needed more time to formulate answers in the voicemail since they needed to ensure that their message contained the prompted information in an efficient way, given the time constraints.

Despite the general similarity between the interview and voicemail results, they represent the least disfluent and most disfluent tasks, respectively, when considering overall disfluency rates. The voicemail was found to contain a significantly higher number of disfluencies than the conversation, which in turn contained a significantly higher number than the interview. This is an interesting pattern, given that monologues have previously been found to contain significantly fewer disfluencies than dialogues. However, voicemail messages are a relatively unique type of monologue, and the limited number of studies that have acknowledged disfluencies in voicemails mention high disfluency rates along with high articulation rates (Padmanabhan et al., 1998; Koumpis & Renals, 2000). The present study did not investigate any differences in segmental or suprasegmental variables across the tasks, and it is certainly possible that articulation rate is a contributing factor to this finding, rather than interlocutor or activity.

The findings discussed in this section indicate that disfluency production is significantly affected by Task; overall disfluency rates were significantly different in all tasks and the distribution of disfluency types was also found to differ across tasks, though some similarities were observed between the interview and voicemail. These results suggest that forensic analysts must be conscious of the influence that different situations may have on a speaker's disfluency behaviour. It is important to note that the recordings used in this study are not completely like the samples analysed in FSC casework; for example, the mock police interview exercise in the WYRED project may be similar to a real police interview in terms of the Q&A format, but the level of emotion and anxiety will likely be extremely different. Despite such issues with ecological validity, the overarching message prevails: disfluency production is very unlikely to remain stable across samples from differing situations.

6.2 Disfluency profiles across tasks

RQ2: On an individual level, to what extent are speakers' disfluency profiles stable across tasks?

In general, individuals' disfluency profiles were not stable across all three tasks, and high levels of within-speaker variation were observed. While some participants (e.g. Speakers 7 and 19) showed very stable overall disfluency counts in all tasks, the majority of speakers showed some stability across two of the three tasks, though the tasks which resembled each other varied by speaker. Overall, taking into account the average frequency per minute of the disfluency categories, very few speakers demonstrated stability across all three tasks; and when two tasks exhibited very similar usage for a category, the similarity was not observed for all disfluency types.

A major factor in the high rates of within-speaker variability was the increased use of interruptions and decreased use of filled pauses in the conversation. Such patterns were observed for most speakers, and although the group analysis revealed no significant difference in filled pause and interruption production within the interview and voicemail, many speakers showed unstable production of these categories across all three tasks.

TOFFA is sometimes employed in casework when the speaker demonstrates unusual or extreme disfluency behaviour. Three individuals whose production of one disfluency category was extremely above or below average were selected for individual profile analysis. The findings suggest that 'extreme' behaviour of a disfluency category can be preserved across Task: Speakers 19 and 29 exhibit relatively consistent behaviour in the production of filled pauses and prolongations respectively. However, Speaker 13 did not produce stable rates of repetitions across any of the tasks, despite consistently producing extremely above-average rates. Furthermore, many other participants were found to produce very high rates of prolongations or repetitions in one task, but very average rates in the others. These speakers demonstrate that extremely above-average production of a disfluency type in one task does not necessarily mean that a speaker will produce above-average rates in another task.

This also raises the issue of what constitutes 'extreme' or 'unusual' disfluency behaviour, and thus when could disfluency analysis be employed by forensic analysts if there is a situation mismatch between samples. Speaker 19 consistently shows very low production of filled pauses, and the complete exclusion of **[erm]** from their speech is very rare; disfluency analysis could therefore be a useful tool in a FSC in this case. Speaker 29's extremely high use of prolonged vocalic segments could also be useful in a FSC, although more population

data is needed to establish whether this behaviour is 'unusual' enough. However, the findings of this study suggest that employing disfluency analysis for Speaker 13 would not be successful, since multiple other speakers displayed similar behaviour, and since the average frequency per minute was very different in each task. Simply having consistent above-average usage of a particular disfluency type does not seem to be 'unique' enough behaviour to reliably distinguish a speaker from the rest of the group.

6.3 Comparison with McDougall and Duckworth (2018)

The present study followed the approach of McDougall and Duckworth and their research into the speaker-specificity of disfluency profiles (2017), as well as their preliminary investigation into the stability of disfluency production across speaking styles (2018). The TOFFA framework was employed in the present study and a comparison across different styles of speech was conducted, though there were a number of differences between the studies regarding methodology and findings: the speakers were from a different area, and this study considered an additional voice message condition.

The findings of the studies differ quite considerably; this study supports the findings of other research (see 2.3) which, in contrast to McDougall and Duckworth, suggests that situational differences will lead to differences in disfluency usage. Where McDougall and Duckworth noted that individuals' disfluency rates were relatively consistent across the two styles, the present study finds significant variation in disfluency production across the tasks. However, it should be noted that both tasks in McDougall and Duckworth's study involved a mock crime, and similarities were found between the pseudo-crime tasks in the present study. The inclusion of detailed individual profile analysis in the present study highlights the instability of speakers' disfluency profiles across the three tasks, and we conclude by suggesting an air of caution when employing disfluency analysis in FSC casework and a reconsideration of what constitutes 'unusual' behaviour.

6.4 Recommendations

The findings of this study lead to two recommendations, one for practice and the other for future research. Firstly, forensic analysts must apply caution when employing disfluency analysis as a method in FSC casework when there is situational mismatch between samples. In some cases, this technique can be informative and aid the analyst in their investigation, yet in others it could be inconclusive or even misleading. It is clear that some individuals retain similar disfluency profiles across different tasks, though it is also apparent that extreme behaviour exhibited in one task does not guarantee the same behaviour in others. Analysts must be aware that certain disfluency categories can be significantly affected by Task, such as filled pauses and interruptions. They should also be aware that these patterns are not consistent across different speakers, and so it is not always possible to make valid predictions about what changes might be expected when comparing samples from different situations and with different interlocutors.

We therefore recommend that disfluency analysis should be considered in cases where there is no significant mismatch in situation; for example, across multiple questioned samples in a relatively similar situation, e.g. multiple telephone calls with similar interlocutor dynamics or multiple covert recordings of conversations within a similar group. It may therefore be the case that disfluency behaviour within known and questioned samples will rarely be compared, but an overly-cautious approach may be necessary until more research using

ground truth data has been conducted on how methods of disfluency analysis cope with realistic case conditions, including situational mismatch. We should also highlight that, as with many types of testing, FSC analysts would only employ disfluency analysis within the context of a wide range of other analyses; our results highlight the importance of this practice and the risks associated with the use of disfluency analysis in isolation.

Secondly, the gathering of population data on disfluency behaviour would be extremely helpful in identifying what constitutes 'unusual' behaviour and therefore determining when disfluency analysis could be applied in a FSC with sample mismatch. It would also help inform typicality judgments more widely, and be useful for further empirical validation. The comparison of an individual's disfluency profiles across tasks can clearly be advantageous in some cases, though further research into disfluency production and disfluency profiles would be able to solidify the opinion that this technique is a worthwhile type of analysis.

7 Conclusion

The present study expanded upon previous work on disfluency analysis to evaluate its efficacy as a method for use in FSC casework. Results showed significant differences in disfluency production between a mock police interview, a paired conversation and a voicemail message. A number of interesting trends were also identified; for example, in the conversation, participants produced significantly lower frequencies of filled pauses and significantly higher frequencies of interruptions than in the other tasks. Extensive within-speaker variation was observed, demonstrating that disfluency analysis should be avoided in cases of situational mismatch between samples since any differences between samples could be due to style rather than necessarily being different speakers. In such cases, forensic analysts would only proceed with disfluency analysis as part of a FSC casework where there was 'unusual' disfluency behaviour. The findings of the current study support this practice, as individual profile analysis revealed that 'unusual' production of a particular disfluency category (e.g. filled pauses or prolongations) could remain stable across different tasks. However, identifying such behaviour is currently a subjective process, and we recommend that population data on disfluency production is collected to assist forensic analysts identify what constitutes 'unusual' disfluency behaviour and to inform typicality judgements.