Progressivity for Voice Interface Design

Joel E. Fischer
Stuart Reeves
Martin Porcheron
firstname.surname@nottingham.ac.uk
School of Computer Science
University of Nottingham
Nottingham, UK

Rein Ove Sikveland
R.O.Sikveland@lboro.ac.uk
School of Social Sciences
Loughborough University
Loughborough, UK

ABSTRACT

Drawing from Conversation Analysis (CA), we examine how the orientation towards progressivity in talk—keeping things moving—might help us better understand and design for voice interactions. We introduce progressivity by surveying its explication in CA, and then look at how a strong preference for progressivity in conversation works out practically in sequences of voice interaction recorded in people's homes. Following Stivers and Robinson's work on progressivity, we find our data shows: how non-answer responses impede progress; how accounts offered for non-answer responses can lead to recovery; how participants work to receive answers; and how, ultimately, moving the interaction forwards does not necessarily involve a fitted answer, but other kinds of responses as well. We discuss the wider potential of applying progressivity to evaluate and understand voice interactions, and consider what designers of voice experiences might do to design for progressivity. Our contribution is a demonstration of the progressivity principle and its interactional features, which also points towards the need for specific kinds of future developments in speech technology.

CCS CONCEPTS

• **Human-centered computing → Natural language interfaces**.

KEYWORDS

Voice, Speech, Conversation Analysis, Design, VUI

ACM Reference Format:
Joel E. Fischer, Stuart Reeves, Martin Porcheron, and Rein Ove Sikveland. 2019. Progressivity for Voice Interface Design. In *1st International Conference on Conversational User Interfaces (CUI 2019), August 22–23, 2019, Dublin, Ireland.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3342775.3342788

1 INTRODUCTION

Progressivity is a driving feature of conversation, so pervasive that if a recipient of a question does not respond, or has difficulty in doing so, the participants will provide an account for this, or another participant might respond instead (cf., [10, 24, 29]). In this work we show that progressivity is also at the core of voice interaction. The progressivity of everyday talk—as articulated in exemplary fashion by work in Conversation Analysis (CA)—refers to the fundamental principle that in conversation "the interactants are concerned with the progress of talk in interaction" [29, p. 387]. It is this sense of getting something done or accomplished that conversationalists relentlessly work towards, so as to continuously move a conversation on. To this end, CA finds talk to be saturated with features that orient towards talk's progressivity, including repairs, assessments, and understanding checks. In this paper we: a) foreground the conceptual primacy of *progressivity* for Voice User Interfaces (VUIs), and b) detail how design might support it.

It would be misleading to suggest progressivity is excluded from current CUI literature. There is a well-recognised need to provide VUI designers with design knowledge to create usable and useful voice user experiences, an aim that is indeed intimately tied up with establishing progress in human-computer interactions. But this is currently a hodgepodge of approaches, design guidelines, and recommendations for the design of VUIs. Instead, we wish to propose a *principled* approach to the CUI community, where the first principle should be to support the interactants' progress towards completion of the interactional sequence. We would also argue that a sensitivity towards progressivity is ultimately more respectful of what is interactionally at stake for participants.

Next, we begin by elaborating CA's concern with progressivity, and then review CUI literature to outline some of the main design thinking currently available (e.g. guidelines, recommendations, etc.). Following this we use the observations of Stivers and Robinson [29] on progressivity in talk as a way of organising examples of VUIs in use. Our study draws on a corpus of naturalistic recordings of people using the Amazon Echo (a "smart speaker") in their homes. Finally, we conclude with a sketch of how the concept of progressivity might be used to drive voice interface design forwards.

2 RELATED WORK

Two bodies of literature frame our work: (i) the Conversation Analysis literature is reviewed to introduce progressivity, and (ii) work on design approaches to VUIs is summarised to position the contribution within the CUI community.

2.1 Progressivity in talk-in-interaction

Progressivity is a central feature of everyday conversation, as examined by Conversation Analysis. CA studies the practical organising structures of talk-in-interaction [24]. Heritage described progressivity as a "great principle of conversational organization" [10, p. 308] and thus of chief concern for CA. When people are referring to

CUI 2019, August 22–23, 2019, Dublin, Ireland

place or person names, for example, participants privilege progressivity unless or until a failure to achieve a shared understanding of the reference is indicated, at which point the participants further negotiate the reference [9]. Kuroshima shows how participants privilege progressivity when ordering sushi at a restaurant, involving the service provider's repetition of the ingredient named by the customer. The repetition serves "to convey what the speaker registers from the previous turn, which thus becomes available for the previous speaker to repair if it is not correct" [12, p. 857]. Schegloff, who probably introduced the term to CA, frames the principle in the following way:

> "Among the most pervasively relevant features in the organization of talk-and-other conduct-in-interaction is the relationship of adjacency or 'nextness.' [...] Moving from some element to a hearably-next-one with nothing intervening is the embodiment of, and the measure of, progressivity." [24, p. 14]

It is worth emphasising that Schegloff frames progressivity as "moving from some element to a hearably-next-one *with nothing intervening*", and that anything that does intervene with this progress is monitored for its relevance to the interactional process—and whether there is some trouble to resolve before proceeding. Schegloff makes this explicit when framing the "trouble problem" as:

> "how to deal with trouble in speaking, hearing and/or understanding the talk so that the interaction *does not freeze in place* when trouble arises, that intersubjectivity is maintained or restored, and that the turn and sequence and activity can *progress to possible completion*?" [24, p. xiv, emphases added].

Our emphases in the above highlight halting ('freezing') interactional progress as the ultimate outcome to avoid, and in turn frames 'progress' as the online resolution of trouble. This has been explicated for a variety of conversational phenomena, but perhaps none is as well-documented as 'repair' practices. Schegloff's work has shown that there is a preference for progressivity in that each repair initiation and resolution furthers the production of the speaker's turn [23]. Abandoning the turn and reformulating (restarting) is not uncommon when progress ceases [ibid.].

Another important concept relevant to progressivity is *preference organisation*. Stivers and Robinson demonstrate an "observable preference for progressivity in interaction" by showing that conversationalists prioritize a preference for answers over a preference for a response by the selected next speaker [29, p. 367]. Whether a response is 'preferred' or 'dispreferred' can be understood by members' analysis of the position and their design of the response turn (which itself offers an account of how they are treating a prior turn). The authors make a range of observations (that we will return to in our own analysis) which demonstrate the importance of the response for progressivity. This includes:

- **Answers are preferred to non-answer responses**. "answers [are] actions that further the progress of the activity" (p. 371).
- **Non-answer responses may impede progress**, especially in cases where "the non-answer response does not further the activity" (p. 372), and where "non-answer responses fail to collaborate with promoting the progress of the activity through the sequence." (p. 373)
- **Accounts are often offered for non-answers.** (e.g. "I don't know because I've not looked at it"). Accounts reveal what is potentially problematic about non-answer responses (p. 373)
- **Participants work to receive and provide answers**. Further evidence that "answers are preferred over non-answer responses can be seen in the way that participants struggle to receive and provide answers if at all possible" (p. 374).
- **The pressure for an answer to be provided over a non-answer response is greater than the pressure for an answer to be provided by the selected speaker** (p. 380). Especially relevant in multi-party environments: e.g. when the selected next speaker fails to respond at the transition relevance place, or claims an inability to answer, or vocally displays difficulty in answering the question.

Note that Stivers and Robinson's framework makes a distinction between *answers*—responses that fit with the previous turn (e.g. a question)—and *responses*—anything that happens (or not) following the previous turn. We use this framework to structure our analysis of VUI interaction in section 3.

2.2 VUI design approaches

The commercial release of VUIs on smartphones and 'smart speakers'—standalone VUI products designed for multi-party use in the home—has precipitated a notable surge in HCI research on the use of—and design for—voice interfaces (including, of course, the CUI conference itself). Approaches to designing VUIs vary from broad categorisations of methodical process, such as iterating and evaluating the design of a voice interface [16], to specific arguments including sociophonetics in the design of synthesised voice responses [32]. In terms of creating cohesive experiences that help users 'get things done', Pearl [17, pp. 16–19] proposes the notion of "conversational design" as one in which interaction with a VUI is designed across multiple requests, employing techniques such as producing visual mockups of interfaces and schematics of conversational 'flows'. Clark et al. [6] drill down on this notion through an elaboration of how "human-agent interaction [should be regarded] as a new genre of conversation" rather than as a simulation of human-human conversation. Further still, Reeves et al. [20] argue that interaction with VUIs is not to be misunderstood as a conversation, and instead better understood as "sequentially organized moves around request and response". In the same vein, Porcheron et al. [18] propose designing with the notion of responses from VUIs as *resources* for further user interaction with the VUI.

Given the commercial motivator behind recent VUI research, it is perhaps unsurprising that there is a plethora of industry-developed guidelines for designing with voice, each concentrating on a distinct set of principles for designers. Google's Conversation Design guidelines [8] calls for designers to "[c]raft conversations that are natural and intuitive for users" and that "conversation design is about the flow of the conversation and its underlying logic". The Alexa Design Guide [1] calls for designers to embrace "situational design" by creating interfaces that are personal, adaptable, available, and relatable to users. A third guide, and perhaps most congruent with the notions proposed in this paper, is the IBM Conversational UX guidelines [5], which stipulates three principles of VUI design: *recipient design* (design for *this* user by creating adaptive scripts that react to the conversation thus far), *minimization* (design for minimal use of talk), and *repair* (design to allow people to fix interactional troubles–which are recurrent in human-human talk–with ease). This last set of notions follow from core CA [22] findings of talk-in-interaction, but due to technical limitations of creating a 'conversational' machine, still require the preparation of flows by designers. This paper seeks to provide additional insights to such guides through the lens of progressivity and its utility in both the evaluation and design of VUIs.

3 PROGRESSIVITY IN VOICE INTERACTION

We have already mentioned the importance of progressivity as a concept for CA. Broadly, our approach in this paper towards progressivity assumes a perspective drawn from Ethnomethodology and Conversation Analysis (EMCA) [7, 21]. Ethnomethodology has a strong relationship with CA in that it examines how members of social settings act as 'everyday sociologists', analysing what one another is doing and making those analyses readily available as embedded features of their own actions. CA takes this basic observation and applies it primarily to talk (and secondarily, bodily action). EMCA has previously been applied to examine VUI use in talk-in-interaction [18, 20]. We present fragments of data which make use of the Jeffersonian transcript notation [2]. The notation denotes short `(.)` pauses in talk and pauses for a specific time, such as `(1.4)` being 1.4 seconds, show where talk is `LOUD`er or emphasised, and where a sound is `elong:::ated`. Overlapping talk is represented using indentation and `[square brackets]`.

The data we draw from as vivid exhibits [3] of progressivity are from a corpus made up of six hours of recordings of five households using the Amazon Echo[1]. Each household provided informed consent to participate in a one-month study in which their interactions with an Echo were selectively audio recorded by a conditional voice recorder reported on in related work [18]. This corpus contains 883 distinct 'Alexa-relevant' utterances (where such utterances are either formulated as or commands to, or questions looking for answers from, the device).

To understand how progressivity features in talk-in-inter- action around voice interfaces, we examine what we take as a three-part sequence of interaction between a device user (INTeractant) and the device (ALExa):

```
01 INT  question / command
02 ALE  response
03 INT  evaluation
```

A question or command addressed to the device is followed by a response and then subsequent evaluative turn uttered by the interactant. It is in this third turn that the interactant's response makes their own assessment of the progressivity of the sequence analytically available—this may be silence or, in many cases, a further turn that evaluates prior turns.

3.1 Progressivity in action

In this section we unpack progressivity through a range of Alexa-interaction sequences, following Stivers and Robinson's framework.

3.1.1 1. Preferred answer response. Our first case presents a straightforward sequence with Alexa where a preferred response is produced by the device, with the subsequent analysis of that response being evaluated positively by the interactant. We are joining participant Rob, who lives with his wife and two children, but is currently alone with Alexa.

```
01 ROB  Alexa (0.5) play all FM
02      (3.3)
03 ALE  all FM (.) on tune in ((radio plays))
```

Rob produces a command (line 1) and after a pause (3.3) the device produces a response. The sequence demonstrates progressivity through the interaction by virtue of a preferred response from Alexa, i.e., successfully playing the desired radio station. The response on line 3 from Alexa embeds an 'analysis'[2] of what Rob said: specifically that it is both correctly formatted (i.e. has used appropriate syntax), and is semantically correct (i.e. invokes a known action—"play"—from the relevant dialogue management processes, and uses corresponding defined variables—"all FM"). That there is no third position (after line 3), i.e. that Rob remains silent, is itself hearably an assessment of the success of the prior turn from Alexa. This demonstrates Rob's orientation to there being 'nothing left to do' given that the desired radio station is now playing.

Note that although pauses in human dialogue are often indicative of trouble, they are common in VUI interaction (line 2). Participants demonstrably expect some delay following both the wakeword and any subsequent utterances directed towards the device. Like many VUIs, the Alexa's pause is accompanied by a visual progress indicator, displaying that the device is processing the input (however, longer pauses can be treated as indicators of trouble).

Alexa's response (line 3) indicates it has provided the 'conditionally relevant' answer (cf. [29]) by repeating the equivalent requested object ("all f m"). As we have already noted, the absence of the third-part evaluation turn suggests that Rob has analysed Alexa's response as the preferred response, facilitating progress. In our corpus, preferred Alexa-responses generally do not attract further responses. Instead, in case of the absence of an evaluative third turn the sequence can typically be seen as successfully reaching completion.

Unfortunately only a minority of sequences in our corpus containing 883 requests are this straightforward. Instead, a preferred response from the device is often absent, thus leading to further turns.

3.1.2 2. Non-answer response. The device also produces non-answer responses which impede progressivity, although sometimes these provide in their formulation some resources with which to move forwards. In these cases a third-part evaluative response is introduced, i.e. a further turn after the Alexa-response. As part of this we also examine how participants analyse in some measure the 'accountability' of device responses in order to progress the sequence.

[1] Note that our examples of progressivity are somewhat limited by Alexa's current command-response model, and should thus only be read as preliminary. Future work is needed to examine interaction with other VUIs to further qualify our findings.

[2] We note that our use of the term 'analysis' here strictly refers to the computational capabilities of Alexa and should not be confused with the ongoing analysis of talk that conversationalists do.

Case 2a. Non-answer responses impede progress. We are joining a family of four at the dinner table, where Emma (the 11-year old daughter) has just asked her mum Susan to ask Alexa for a "normal family quiz". We can see here two overlapping three-part sequences.

```
01 SUS  Alexa? (0.7) set us a family quiz.
02      (2.5)
03 ALE  sorry. (.) I can't find the answer to the question
        I heard
04      (0.4)
05 EMM  Alexa:? (1.0) Set (0.3) a family quiz
06      (2.3)
07 ALE  sorry. (.) I don't have the answer to that question.
08      (0.4)
[transcript continues in Case 3a]
```

Alexa's response (line 3) follows on from Susan's initial command to Alexa. This is treated as a non-answer response by Emma (line 5) in that she produces a nearly identical command to Susan. Emma's turn (line 5) initiates a further response from Alexa, which itself is another, similar non-answer response (line 7). This is yet again analysed as a (dispreferred) non-answer response by members of the family, leading to further reformulations by others present (line 8 onwards, transcript continues in Case 3a).

The point here is that progress through the sequence is clearly impeded through the repeated production of similar non-answer responses by the device. This fragment lets us begin to see the features that indicate problems with interactional progressivity in talk with VUIs. We thus see how various conversational methods are recruited by participants to break the impasse: pauses, restarts, and variations of prior commands follow on from non-answer responses in an attempt to move on through the sequence towards possible completion.

Case 2b. Accounts offered for non-answer responses can lead to recovery. People frequently offer *accounts* in their non-answer responses in everyday conversations [29]. These accounts can reveal or suggest possible resolutions and progress the conversation in turn. Somewhat analogously with accounts, we find output from VUIs to sometimes includes relevant interactional resources to aid progressivity in the midst of trouble. Let's consider a sequence in which the resources embedded in the turn from Alexa seems to contribute to a reformulation, resolution and progress by the participant. Here we join Rob again.

```
01 ROB  Alexa (1.7) star::t (.) white noise
02      (1.6)
03 ALE  sorry (0.3) I'm having trouble understanding
        ((beep))
04      (2.9)
05 ROB  Alexa (1.8) start white noise
06 ALE  ((skill starts))
```

Alexa provides another non-answer response (line 3) to Rob's initial command (line 1), however this time further resources that *could* account for the failure seem to be built in more explicitly (than Case 2a) to that response ("trouble understanding", line 3). Regardless of whether this sequence unfolds as a designer might have intended, Rob nevertheless appears to treat the response from Alexa as matter of 'hearing'[3] trouble and repairs it accordingly. Thus, following the characteristic trouble-indicative pause (line 4) Rob subtly reformulates his original utterance by shortening his production of the command "start", from the earlier prolonged "star:::t". Next, Alexa starts the skill (line 6), and thus the sequence progresses to completion with no immediate further turns from Rob (subsequent interaction with the skill marks the start of a new sequence).

Although Alexa's response output of "I'm having trouble understanding" offered only a vague insight into what went wrong, Rob's corrective action of targeted alteration of the repeated initial command suggests that this response provided enough resources for Rob to analyse his prior pronunciation as a source of trouble. Thus, the provision of those resources on the device's part may have played a part in helping to resolve the issue and further the progression through the sequence.

We can now return to the prior fragment, Case 2a, and note that something similar is happening here. Emma's attempt on line 5 builds a more truncated variation of Susan's original command ("set us a family quiz"). Emma omits "us a" but also emphasises the beginning of "Set" coupled with a pause (0.3). This seems to treat the problem as a matter of 'hearing' once again, possibly in response to the very formulation of the prior non-answer from Alexa (line 3) that itself introduces a suggestion of the trouble source with "I heard" (i.e. technically that a command component has not been successfully processed by speech-to-text).

3.1.3 3. Interactants work to provide and receive answers. In order to progress through the sequence, parties to the interaction do considerable work to provide and elicit answers. By 'do work' we mean that, as Stivers and Robinson point out, there is often great "pressure for an answer", sometimes even if conversationalists are "in no position truly to answer them" [29, p. 364]. First let's look at a case that demonstrates how participants work to elicit answers from Alexa.

Case 3a. Working to receive an answer from Alexa. The following sequence again joins the family dinner table, picking up where the sequence in Case 2a left off. Now, the 10-year old Liam and dad Carl join in.

```
01 LIA  Alexa:? (0.9) please set (0.3) a [family quiz.    ]
02 ALL                                   [((laughter))    ]
03      (1.2)
04 ALE  I wasn't able to understand [the question I heard.]
05 EMM                              [ ((laughs))          ]
06 LIA  beep
07      (0.9)
08 CAR  ALEXA, (0.7) FAMily quiz.
```

After Alexa produces two non-answer responses (in Case 2a), Liam alters the prior query by inserting "please" (line 1), which of course amuses Eric and Carl whose laughter (line 2) anticipates Alexa's third non-answer (line 4). Next Emma laughs as well, while Liam responds mockingly with a sound imitating Alexa (line 6). Finally, Carl takes his turn at producing an even more pared down variation, leaving out the action-keyword altogether (i.e. "set", line 7).

Taken together with Case 2a, this extended sequence demonstrates the collaborative work participants must perform to get the desired response from Alexa. All four family members have a go at getting Alexa to work, revising the prior query slightly with each subsequent attempt. Porcheron et al. have previously observed

[3] Again, we must be careful to note that troubles of 'hearing' here are only analogous in a limited way with routine conversational hearing troubles. As a supposed 'conversationalist', Alexa's 'hearing' troubles are a complex result of various facets of microphone design, device placement, overlapped talk, and the speech-to-text pipeline. Such troubles tend to be treated as one might any other troublesome input device, as problems to be worked with, like a stuck key or a shattered capacitive phone screen.

this kind of "collaborative repair" in VUI use in multiparty settings [18, 19]. Despite the family's best efforts, progress through the sequence is persistently impeded, but they nevertheless struggle on with reformulation after reformulation, which seems preferred by the family to complete abandonment of the interaction.

Case 3b. Working to provide an answer. Sometimes responses from the device can themselves play into the work of providing an answer. We now turn to one such instance, again with the family of Case 3a. Here, Emma is trying to invoke another quiz game called "Beat The Intro".

```
01 EMM  Alexa? (.) (1.0) play beat the intro
        ((line ommitted))
02      (1.9)
03 ALE  you want to hear a station for b b intro
        [(0.4) right?]
05 EMM  [ no         ]
06      (1.1)
07 EMM  no (.) I don't Alex(h)uh (0.5) (h)No,
08      (1.3)
09 ALE  alright.
```

Here, the device does something we have not seen so far, which is that it produces a clarification question (line 4) in response to Emma's initial request (line 1). This does two things: (i) it 'proposes' that Emma wanted to "hear a station", and (ii) it makes the speech-to-text transcription of Emma's prior request hearably available to Emma as "b b intro". Her immediate overlapping answer (line 5), albeit negative, and affirmation (line 7), treats the Alexa response as a proposition of next action. It also serves the purpose of progressing the sequence, drawing to a close with a simple and apt response by Alexa (line 9).

As we have indicated on other points throughout this paper, we are not suggesting that Alexa 'works to provide an answer' as a co-conversationalist. This would set up a conceptual confusion between devices that just so happen to produce turn-by-turn vocalisations, and competent conversationalists [18]. However, there *is* an analogy here in that the Alexa response of line 4 is treated in a way by participants that progresses the sequence more substantively than prior fragments in which impediments to progressivity are encountered. Alexa here produces something that could be considered an insertion sequence. Insertion sequences have been argued to be produced in favour of intersubjectivity (i.e. mutual understanding) over progressivity, as the insertion itself can be seen to impede progress [12]. However, here Alexa's 'insertion' does seem to be analysed by Emma as offering her concrete moves that are not just repeating or reformulating prior commands as in Cases 2a and 3a (specifically her rapid negative response to close down the sequence with "no", thus quickly bringing matters to a closure).

3.1.4 4. Progressivity is more important than provision of an answer.

As Stivers and Robinson observe particularly for multiparty settings, the preference for progressivity trumps the preference for an answer from the selected speaker. This means that someone other than the selected speaker may produce a next turn in cases where none is seemingly forthcoming from the selected speaker. The following sequence demonstrates that for VUI interactions too: should the response from Alexa not appear, then the interactants may well just move on to produce their own completion of the sequence by whatever means possible. Consider the following fragment, again joining the same family as the prior cases. This occurs sometime after Case 3b, where they are still trying to start the quiz game:

```
01 SUS  yeaherr:: Alexa skills (.) beat the intro
02      (4.5)
03 SUS  uh::
04 EMM  she didn like tha:t
```

This time, Susan's initial command (line 1) is met with 4.5s of silence, a non-response (line 2). Susan's evaluation of this response is a sigh, which also acts as a pause or hesitation (line 3) that is inserted before any more substantive next move by herself or others. Then, on line 4 we get Emma's assessment of why the non-answer response was received ("she didn't like that"). This effectively works to move the sequence on (line 4).

Taken together with the prior sequence in Case 3b, it appears that for VUI interaction, if the response is not forthcoming, participants prioritise progression over the need to elicit an answer from the VUI. In some sense, the VUI will be 'left behind' if it cannot facilitate progressivity (which in many ways suggests that little has perhaps been learned since Suchman's *Plans and Situated Actions* which clearly articulated misalignments between machine design rationales and situational features only available to users [31]).

4 DISCUSSION

We think that a focus on progressivity has potential value for the CUI community. However, to reiterate some caveats before we turn to design considerations: first, Stivers and Robinson's analysis was focused on 'information questions' in conversations [29], which may limit how applicable their framework is to human-machine interactions that often follow instruction-response patterns. Second, we stress that the concept of progressivity can't offer well-defined guidelines because what counts as progress in a sequence of interaction is formulated *in and as* the unfolding of the sequence itself. However, we can explore different strategies to design *for* progressivity, in view of the design goal of helping 'people to get things done' in interaction.

Next we take this discussion of our study of progressivity in two directions: first to consider what the implications of progressivity are for the technologies driving VUIs; and second, how design approaches can take progressivity into account practically.

4.1 Progressivity to evaluate VUIs

We saw in our fragments how the position and design of the interactants' third turn was critical to understanding how this might act as an *evaluation* of what came before (i.e. the VUI's turn). From this we can see what characteristic *features* make unfolding trouble with progressivity observable-and-reportable (i.e. analysable to members of the setting and us authors as 'spectators'). This includes gaps (pauses) before the evaluative turn (Cases 2-4), hesitations (Case 4), negative responses (Case 3b), and overlapping talk (Cases 3a&b).

CA research has demonstrated how speakers use a combination of syntactic, prosodic, and pragmatic features to constantly work together to move the conversation forward, e.g. by monitoring the other person's concurrent feedback (backchannels, nods, posture, continued attention), and revise their turn on-the-fly if necessary to resolve trouble in agreement and understanding (e.g. [29]). Thus, there is an abundant understanding of how particular features are used in the work of speaking. In Table 1 we have summarised

Table 1: Progressivity features

Function	Features
1. Pre-turn progressivity	
Projecting next speaker	Address terms [14], gaze and body movement [30]
2. Within-turn progressivity	
Same speaker hesitations / turn holding	Sound stretches [33]; acoustically stable prolonged speech segments; glottal stop [15]; discourse marker "I mean".
Other-speaker talk in overlap	Competitive overlap; loud, high pitch speech compared to non-competitive overlaps [27]
3. Post-turn progressivity	
Other-speaker expressions of failed understanding	Marking previous talk as inapposite; long gaps [11], turn-initial "uhm", laughter.

some of these features from the literature, along with their common 'functions' in conversations regarding their turn-respective position (pre-, within- and post-turn).

It is worth considering the turn-respective position of features in Table 1 in relation to the structure of our three-part base sequence (request, response, evaluation).

4.1.1 1. Evaluating the request turn. Regarding the question / command turn addressed at the VUI, observation of pre-turn progressivity features can be used to evaluate address and initiation of interaction with the VUI. Multi-turn interactions eschewing the wakeword come to mind as a challenge, as does work that looks to incorporate, for instance, gaze for turn-taking in human-robot interaction [28].

Secondly, within-turn progressivity features produced by the speaker addressing the device are currently ignored by the 'one shot' model of current Automatic Speech Recognition (ASR), although we note that incremental ASR is increasingly available (e.g. IBM Watson Speech to Text[4]). After initiation through the wakeword, current VUI ASRs 'listen' for input, but do not process the input until a pause of a certain duration is detected. The current model means that the moment-by-moment production of talk is unavailable for machine 'reasoning'.

Research seeking to address the shortcomings of the current model is far from new. For instance, incremental dialogue processing could bring a range of "phenomena into reach that cannot otherwise be modelled", including, "concurrent feedback ('uh-huh', 'yeah'), fast turn-taking, [and] collaborative utterance construction" [25, p. 85]. Incremental ASR and dialogue management are at the cutting edge of current research in speech technologies research communities [4]. Results show that an incremental semantic parser outperforms state-of-the art retrieval models on datasets which contain "spontaneous incremental dialogue phenomena such as restarts and self-corrections" [26, p. 98]. Future work should bring together speech technologists working on incremental and spontaneous speech with CA-driven research in voice interaction to create a step-change towards the development of VUIs capable of supporting aspects of progressivity.

4.1.2 2. Evaluating the evaluation turn. Examining the evaluation turn—in particular post-turn progressivity features—can give insight into other-speaker expressions of failed understanding, making available the interactant's own analysis of the VUI response, for example through long gaps (Cases 2-4), turn-initial hesitations (Case 4), and laughter (Case 3a).

There is much more to be done here. One avenue of future work could adopt the CA approach to examine progressivity in greater depth, and using different data. Beyond studying recorded instances of VUI interaction, there is an opportunity to develop technical contributions to ASR and dialogue systems that evaluate the evaluation turn in real-time. There is a paucity of evaluation metrics that indicate, for instance, how well a dialogue system supports progressivity on a turn-by-turn basis [13]. We also need to equip dialogue systems with the capability to infer from the interactant's evaluation turn whether the prior response produced by the system *could* be seen as supporting or hindering progressivity. For instance, the absence of an evaluation turn or a short gap may indicate support, while a long gap, hesitation, overlap and / or features of negative sentiment may indicate interference. Coupled with online learning approaches, this turn-by-turn evaluation might result in dialogue systems that more regularly enable people to progress through the interactive sequence to completion.

4.2 Design for progressivity: Response design

Without access to the ASR and other core components, how might designers build progressivity support into their voice experiences and skills? Designers can provide experiences that drive the interaction towards completion through *response design*. To this end we offer **five questions** designers may ask of their designed responses to reflect on how that design takes progressivity into account.

What is 'response design'? As demonstrated by our fragments of audio data, VUIs can frequently involve further action beyond the initial question / command and response pairing—done in the third part of a sequence. This evaluative turn embeds the conversationalist's reasoning about the preceding device response, and thus uses the response as a resource in successive actions. Therefore, the utility of the response from the VUI is a core concern when designing to support users 'getting things done'; put simply, response design can support or impede the user's progress. 'Response design' should thus really be *progressive response design*. Our questions outline what this practically means.

1. Could this response be delivered minimally, allowing users to progress to their next move earlier? There is a progressivity trade-off between providing minimal responses, and offering more comprehensive indications of potential trouble. For the latter, checking the understanding of some element of the previous turn can show that things are moving, while also providing the ability to check that such movements are in the right direction. This then provides the user the resources to rephrase their first attempt rather than starting the sequence all over again. However, sometimes it may be beneficial to adopt minimalism of response, so as to quickly 'start again' rather than provide a semblance of coherence if user is repeatedly unable to progress.

2. Does this response support or impede the user to be sure of the VUI's 'understanding' of spoken talk? Device responses which are

[4] https://www.ibm.com/watson/services/speech-to-text/

incongruent with a user's prior utterance (e.g. responding about a question when a command is given as per Case 2a) do not necessarily support progression (even if well intended as part of the persona). Repeated issuance of an identical or near-identical non-answer can impede progress because no useful resources are available to a user to understand if the adjusted requests are 'understood' differently by the VUI.

3. Could other resources for providing users something to move on with help, e.g. accounts of what went wrong? In this, can repeating the terms the ASR transcribed—even if they do not align with possible next actions—support users' analysis of their interactive sequences with the VUI (as per Case 3b)? Furthermore, could these responses identify these terms that are out of context, for example as the family vary the verb to start the skill in Case 2a and 3a?

4. How could the user provide more information to this response than expected? If users' goals are to complete the task minimally, this may mean answering a question not specifically asked by the VUI in a response. Given the preference for progressivity in interaction, does the response provide users the options to minimally complete the sequence? Consider how in everyday conversation people may answer questions not specifically asked, so as to progress interaction (e.g. skipping ahead to answer an anticipated follow-up question).

5. How do users themselves work to support, and halt, progress of a sequence in response to the VUI, either in overlap with or following the VUI's response? In this, we flag up for consideration that users' intent to accomplish a sequence can be assumed. In conversation, troubles are a routine occurrence, and are routinely repaired [22]. Is it possible that the user will want to undertake repair as a result of a given response? Or perhaps users may select to end a sequence for whatever reason. Conversation between people is driven by moment-by-moment analyses of talk-in-interaction. The asymmetry of a VUI's access (vs. the user's access) to resources to analyse this is well documented [31, pp. 77–122]. While it is reasonably obvious to humans when we become exasperated or want to end a conversation, for VUIs, in lieu of being able to sense this, how can we enable users to halt a sequence with ease?

Consideration of these questions should only be applied subsequent to initial design work, following platform specific guidelines to create coherent user experiences. These questions form part of the iterative stages [16] of designing a VUI, after initial conversation flows, as per most industry design guide recommendations.

We acknowledge that many of our questions relate to existing principles in VUI design guides. For example, delivering responses to 'confirm user's input' in cases including where a candidate action is to unfold is a recommendation in Google's guidelines [8]. The Amazon guide [1] argues for offering next actions when the VUI did not understand input, including re-asking the question, and the IBM guide calls for systems to "fail gracefully" by "[d]isplay[ing] what the agent does understand so the user can better diagnose and repair the trouble" [5, section "Practices"]. Our purpose in posing these questions is thus to *augment* such extant guidelines by introducing VUI designers to the utility of the concept of progressivity. We have shown how existing understandings of talk in the social sciences can support designers' conceptual approach to supporting progressivity by design.

Finally, we propose the above questions as a provocation, to further the debate on approaches for designing and evaluating VUI use.

5 CONCLUSION

The (re)emergence of so-called 'conversational' interfaces has led to increased focus on the role of natural language in interaction. Within this broader context, the adoption of voice interfaces in everyday circumstances has led us to consider how research on forms of talk—particularly Conversation Analysis—might offer new design opportunities for the CUI community. Drawing from established work in CA, we have articulated just some of the methods leveraged in aid of the fundamental orientation to progressivity, as displayed by conversationalists with and around voice interfaces so as to keep interactions moving forwards. Our study has also pointed us towards significant challenges for future work in voice interfaces. Not only does it suggest the need for new ways of blending insights from CA with advances in ASR, natural language understanding (NLU) and dialogue management pipelines, but also that there are key design lessons to learn from plumbing the central organising features of everyday talk.

ACKNOWLEDGMENTS

This work is supported by the Engineering and Physical Sciences Research Council [grant numbers EP/N014243/1, EP/M02315X/1, EP/K025848/1].

REFERENCES

[1] Amazon Inc. 2019. Alexa Design Guide. https://developer.amazon.com/docs/alexa-design/get-started.html

[2] J. Maxwell Atkinson and John Heritage. 1984. Transcript Notation. In *Structures of Social Action: Studies in Conversation Analysis*. Cambridge University Press, ix–xvi. https://doi.org/10.1017/CBO9780511665868

[3] Liam Bannon, John Bowers, Peter Carstensen, John A. Hughes, Kari Kuutii, James Pycock, Tom Rodden, Kjeld Schmidt, Dan Shapiro, Wes Sharrock, and Stephen Viller. 1993. Informing CSCW System Requirements. In *COMIC Deliverable 2.1.*

[4] Timo Baumann, Casey Kennington, Julian Hough, and David Schlangen. 2017. Recognising conversational speech: What an incremental asr should do for a dialogue system and how to get there. In *Dialogues with Social Robots*. Springer, 421–432. https://doi.org/10.1007/978-981-10-2585-3_35

[5] Bob Moore and Raphael Arar. 2019. Conversation design. https://conversational-ux.mybluemix.net/design/conversational-ux/

[6] Leigh Clark, Nadia Pantidi, Orla Cooney, Philip Doyle, Diego Garaialde, Justin Edwards, Brendan Spillane, Emer Gilmartin, Christine Murad, Cosmin Munteanu, Vincent Wade, and Benjamin R. Cowan. 2019. What Makes a Good Conversation?: Challenges in Designing Truly Conversational Agents. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, Article 475, 12 pages. https://doi.org/10.1145/3290605.3300705

[7] Harold Garfinkel. 1967. *Studies in Ethnomethodology*. Prentice-Hall.

[8] Google Inc. 2019. Conversation design. https://designguidelines.withgoogle.com/conversation/conversation-design/welcome.html#

[9] John Heritage. 2007. *Intersubjectivity and progressivity in person (and place) reference*. Cambridge University Press, Cambridge, 255âĂŞ280. https://doi.org/10.1017/CBO9780511486746.012

[10] John Heritage. 2009. Conversation Analysis as Social Theory. In *The new Blackwell companion to social theory*, Bryan S. Turner (Ed.). Blackwell Oxford, 300–320. https://doi.org/10.1002/9781444304992.ch15

[11] Kobin H Kendrick and Francisco Torreira. 2015. The timing and construction of preference: A quantitative study. *Discourse Processes* 52, 4 (2015), 255–289. https://doi.org/10.1080/0163853X.2014.955997

[12] Satomi Kuroshima. 2010. Another look at the service encounter: Progressivity, intersubjectivity, and trust in a Japanese sushi restaurant. *Journal of Pragmatics* 42, 3 (2010), 856–869. https://doi.org/10.1016/j.pragma.2009.08.009

[13] Oliver Lemon. 2018. Designing engaging open-domain conversational AI for the Amazon Alexa Prize. (2018). Keynote at SICSA workshop on Conversational AI.

[14] Gene H Lerner. 2003. Selecting next speaker: The context-sensitive operation of a context-free organization. *Language in society* 32, 2 (2003), 177–201. https://doi.org/10.1017/S004740450332202X

[15] John Local and John Kelly. 1986. Projection and 'silences': Notes on phonetic and conversational structure. *Human studies* 9, 2 (1986), 185–204. https://doi.org/10.1007/BF00148126

[16] Neil Patel, Sheetal Agarwal, Nitendra Rajput, Amit Nanavati, Paresh Dave, and Tapan S. Parikh. 2008. Experiences designing a voice interface for rural India. In *2008 IEEE Spoken Language Technology Workshop*. 21–24. https://doi.org/10.1109/SLT.2008.4777830

[17] Cathy Pearl. 2016. *Designing Voice User Interfaces: Principles of Conversational Experiences* (1st ed.). O'Reilly Media, Inc.

[18] Martin Porcheron, Joel E. Fischer, Stuart Reeves, and Sarah Sharples. 2018. Voice Interfaces in Everyday Life. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Article 640, 12 pages. https://doi.org/10.1145/3173574.3174214

[19] Martin Porcheron, Joel E. Fischer, and Sarah Sharples. 2017. "Do Animals Have Accents?": Talking with Agents in Multi-Party Conversation. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17)*. ACM, New York, NY, USA, 207–219. https://doi.org/10.1145/2998181.2998298

[20] Stuart Reeves, Martin Porcheron, and Joel Fischer. 2018. 'This is Not What We Wanted': Designing for Conversation with Voice Interfaces. *Interactions* 26, 1 (Dec. 2018), 46–51. https://doi.org/10.1145/3296699

[21] Harvey Sacks. 1992. *Harvey Sacks: Letures on Conversation*. Basil Publishing, Oxford.

[22] Harvey Sacks, Emanuel A. Schegloff, and Gail Jefferson. 1974. A Simplest Systematics for the Organization of Turn-Taking for Conversation. *Language* 50, 4 (jan 1974), 696. https://doi.org/10.2307/412243

[23] Emanuel A Schegloff. 1979. The relevance of repair to syntax-for-conversation. *Syntax and semantics* 12 (1979), 261–286.

[24] Emanuel A Schegloff. 2007. *Sequence Organization in Interaction: A Primer in Conversation Analysis*. Vol. 1. Cambridge University Press.

[25] David Schlangen and Gabriel Skantze. 2009. A General, Abstract Model of Incremental Dialogue Processing. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL '09)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 710–718. http://dl.acm.org/citation.cfm?id=1609067.1609146

[26] Igor Shalyminov, Arash Eshghi, and Oliver Lemon. 2017. Challenging Neural Dialogue Models with Natural Data: Memory Networks Fail on Incremental Phenomena. In *Proceedings of SEMDIAL 2017 (SaarDial) Workshop on the Semantics and Pragmatics of Dialogue*. ISCA, Saarbrucken, Germany, 98–106. https://doi.org/10.21437/SemDial.2017-11

[27] Rein Ove Sikveland and David Zeitlyn. 2017. Using prosodic cues to identify dialogue acts: methodological challenge. *Text & Talk* 37, 3 (2017), 311–334. https://doi.org/10.1515/text-2017-0007

[28] Gabriel Skantze, Martin Johansson, and Jonas Beskow. 2015. Exploring Turn-taking Cues in Multi-party Human-robot Discussions About Objects. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction (ICMI '15)*. ACM, New York, NY, USA, 67–74. https://doi.org/10.1145/2818346.2820749

[29] Tanya Stivers and Jeffrey D Robinson. 2006. A preference for progressivity in interaction. *Language in society* 35, 3 (2006), 367–392. https://doi.org/10.1017/S0047404506060179

[30] Tanya Stivers and Federico Rossano. 2010. Mobilizing response. *Research on Language and social interaction* 43, 1 (2010), 3–31. https://doi.org/10.1080/08351810903471258

[31] Lucy Suchman. 1985. *Plans and Situated Actions: The Problem of Human Machine Communication*. Cambridge University Press, Cambridge. 220 pages.

[32] Selina Jeanne Sutton, Paul Foulkes, David Kirk, and Shaun Lawson. 2019. Voice As a Design Material: Sociophonetic Inspired Design Strategies in Human-Computer Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, Article 603, 14 pages. https://doi.org/10.1145/3290605.3300833

[33] Vasilisa Verkhodanova, Vladimir Shapranov, and Irina Kipyatkova. 2017. Hesitations in Spontaneous Speech: Acoustic Analysis and Detection. In *International Conference on Speech and Computer*. Springer, Springer, 398–406. https://doi.org/10.1007/978-3-319-66429-3_39