



# VISION

Making talking to an Agent  
more natural



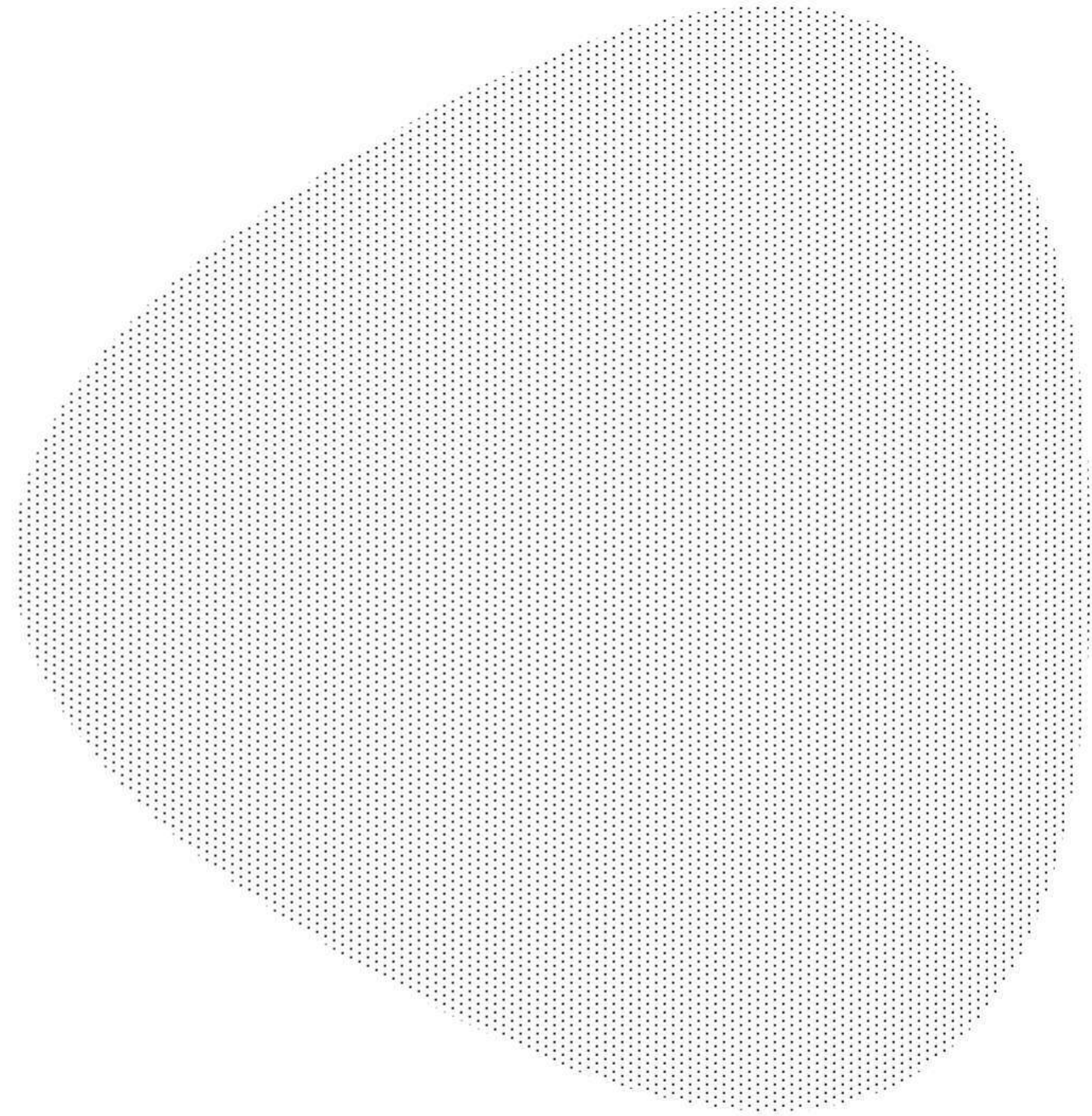
# GROUNDING IN COMMUNICATION WITH AN A.I.

**8 week** Internship Project  
May-July 2019  
Anshuman Dhar / IIT Guwahati  
Voice UX Team

# Grounding in Communication

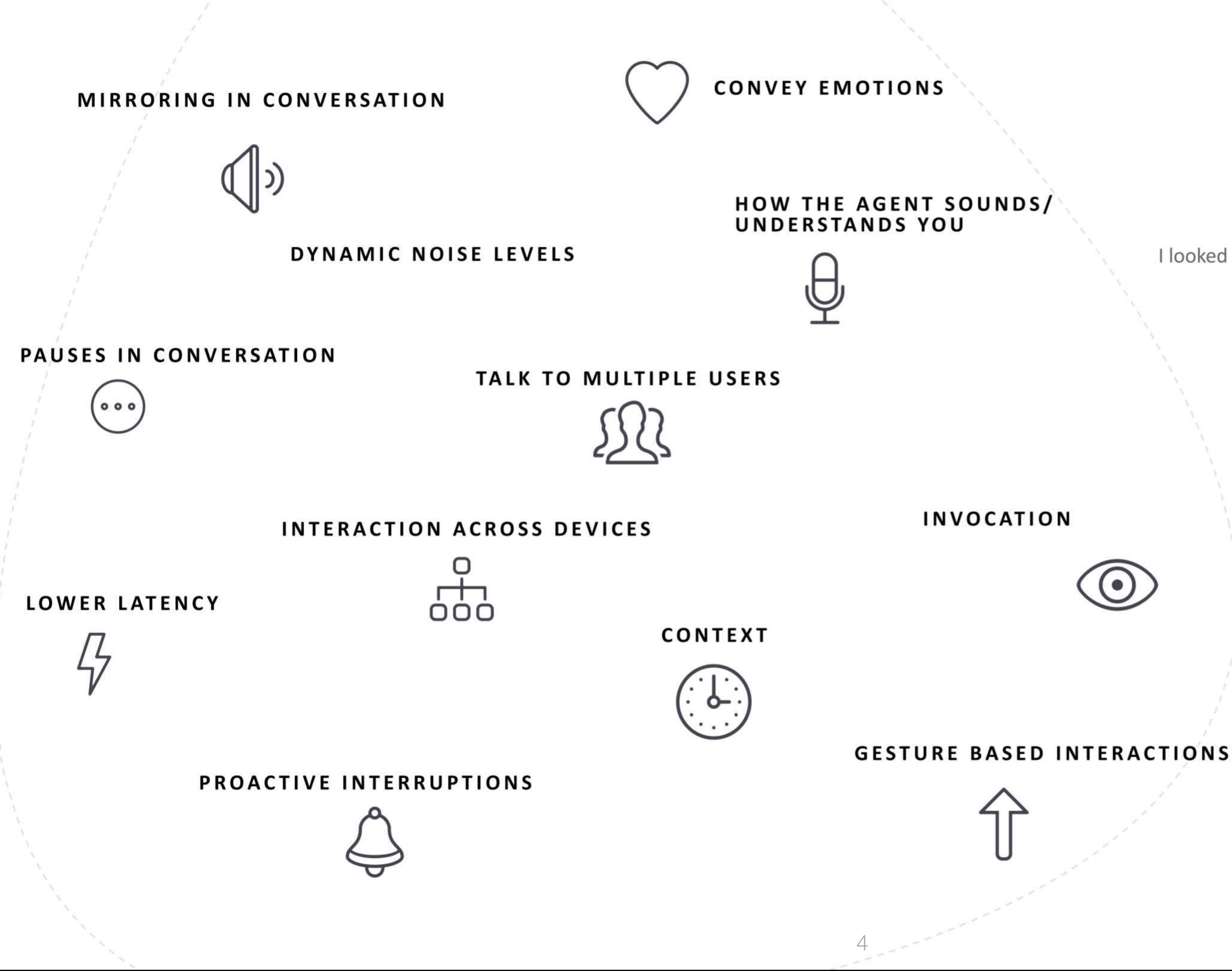
---

The untaught rules of communication which all humans have.



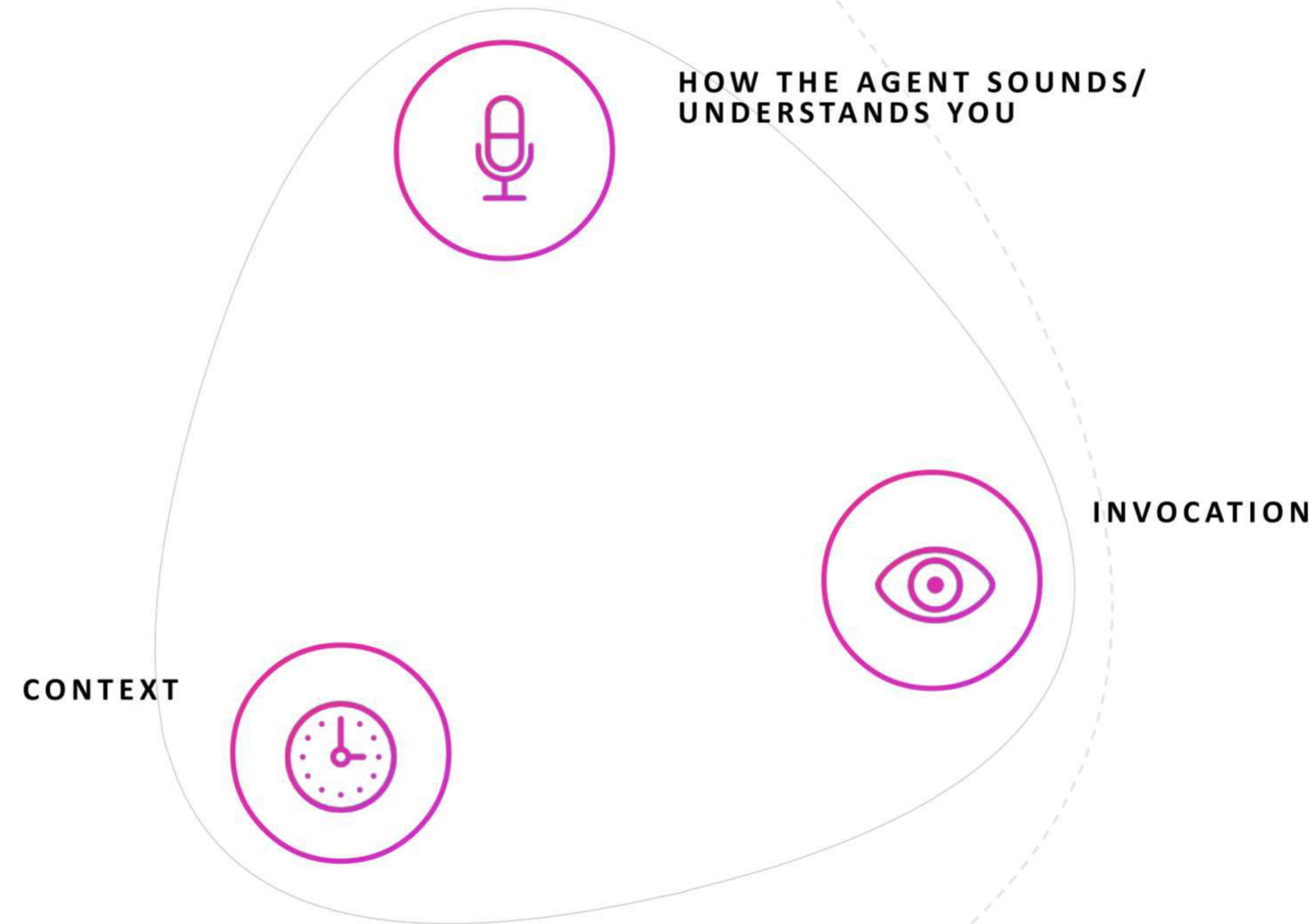
# What is more natural?

I looked into the different areas which could make an agent more natural to talk to.



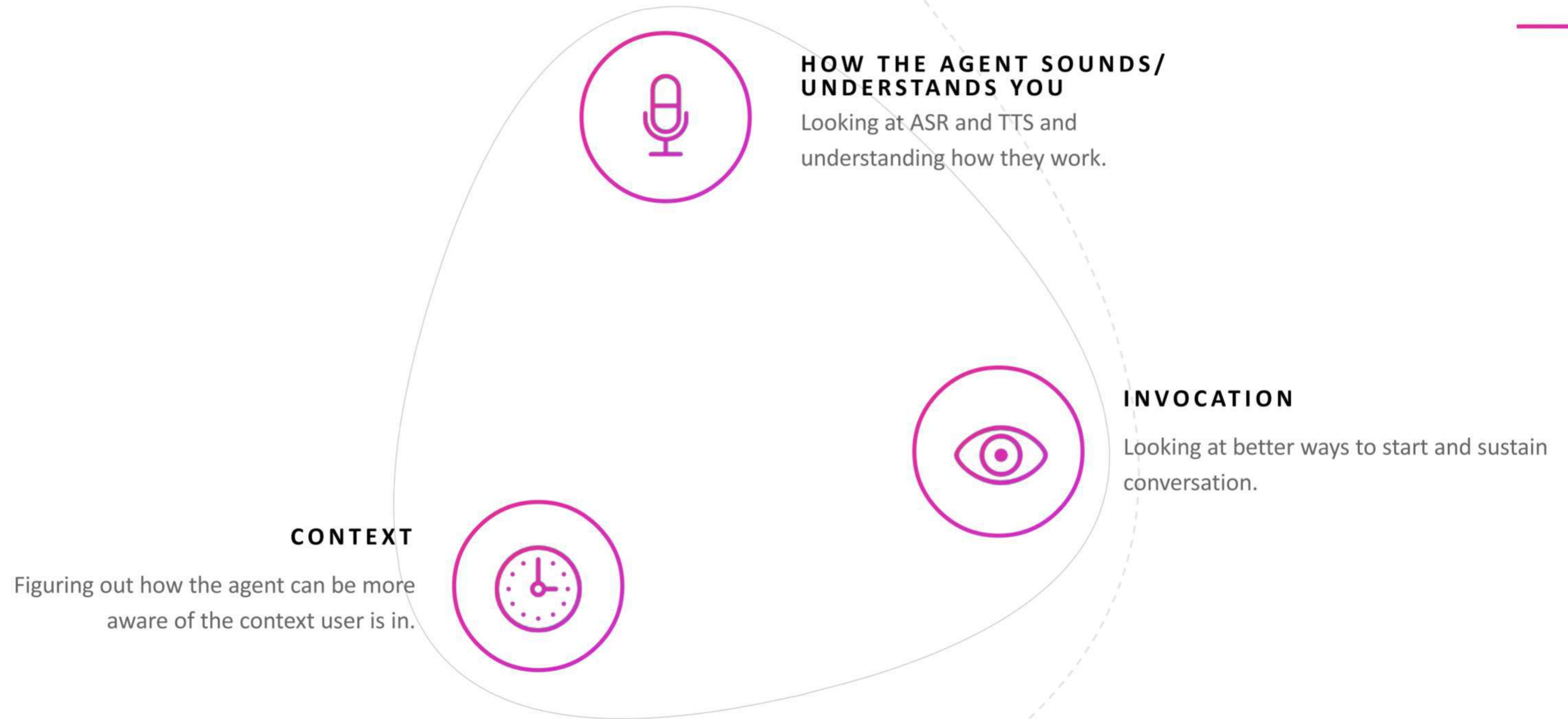
# What is more natural?

---



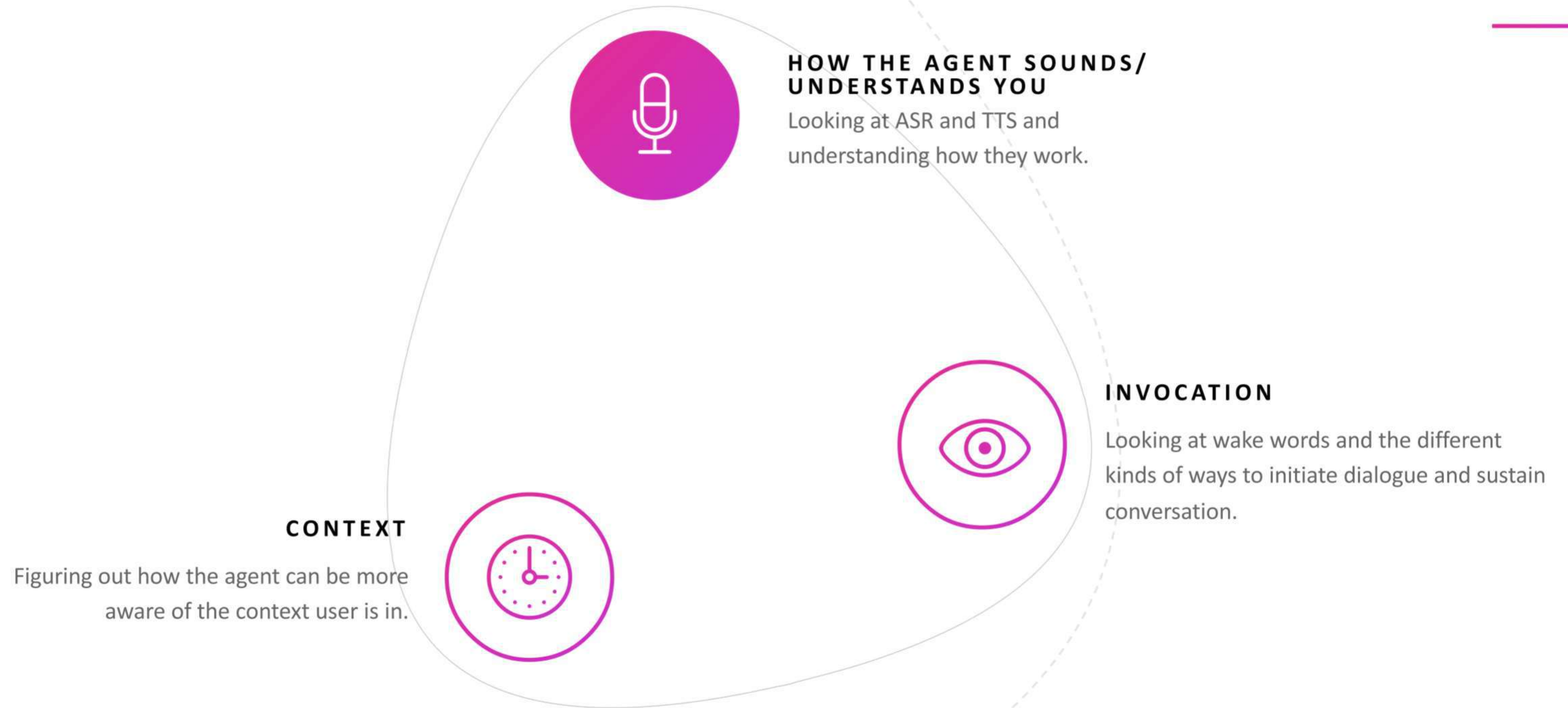
# What is more natural?

---

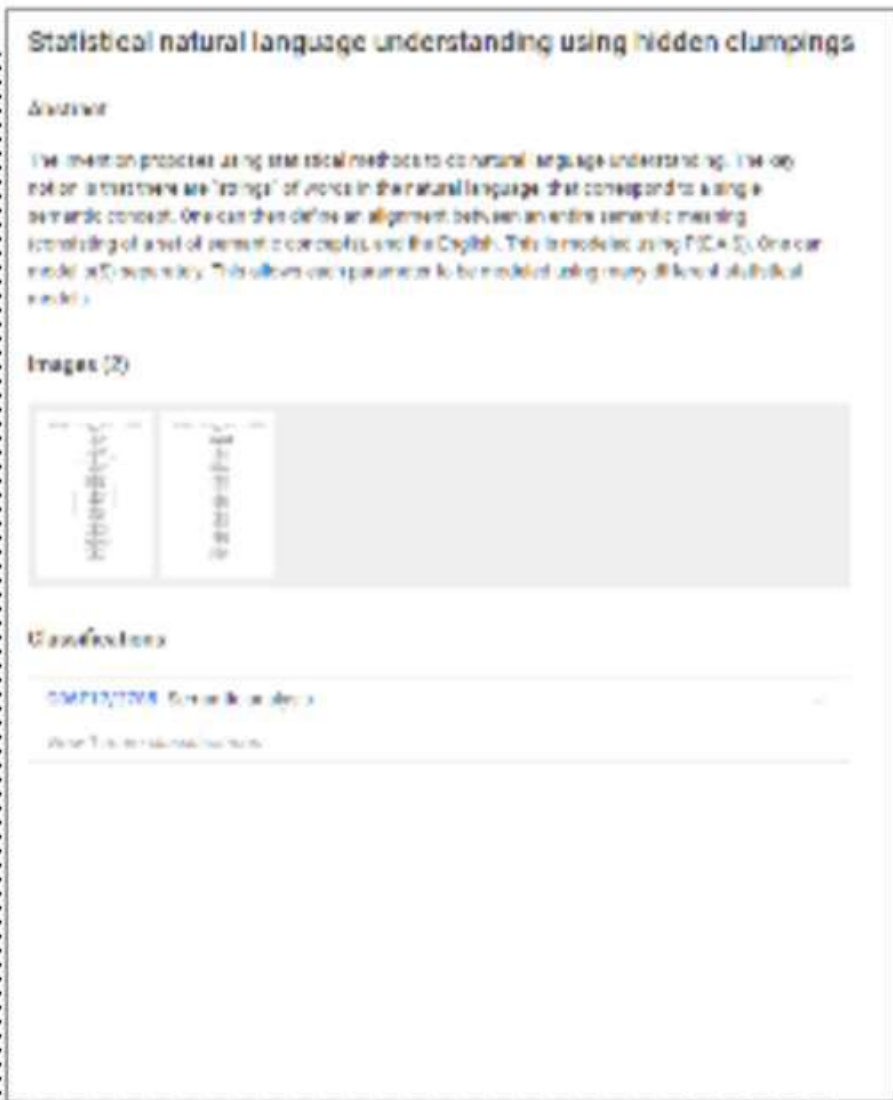


# What is more natural?

---



# How the Agent Understands



SAMUEL S. EUSTEN  
Sci. Communication Research

Categories and subject headings: D.2.7 Software Engineering; Database and Management systems; programming languages; F.0.2 Theoretical Foundations; Languages—query languages; H.0.1 Information Storage and Retrieval; Information Search and Retrieval—query languages (the vertical search); H.0.1 Information Systems Applications (the horizontal); I.2.4 Artificial Intelligence; Knowledge Representation Formalisms and Methods—semantic networks; L.1.1 Data Models (Research View of Language Processing)—language models and constraints.

**Abstract:** This design-based research project investigated the effects of a computer-based, self-paced, self-directed, and self-paced learning environment on the learning of basic algebraic concepts. The study was conducted in a middle school classroom. The results of the study indicated that the self-paced, self-directed, and self-paced learning environment had a positive effect on the learning of basic algebraic concepts. The study also indicated that the self-paced, self-directed, and self-paced learning environment had a positive effect on the learning of basic algebraic concepts. The study also indicated that the self-paced, self-directed, and self-paced learning environment had a positive effect on the learning of basic algebraic concepts.

PRR (Proprietary Restricted Right) is a restricted English database query language whose implementation has achieved engineering goals, namely, scalability, interoperability, transparency, performance, and ease with a reliable database management system that supports large numbers of concurrent users and large databases. Reliability has not been demonstrated, but initial indications are encouraging. The other goals have clearly been achieved.

This work was partially funded by an NSF DUE LUMS grant. The authors would like to express appreciation to the American Society of Mechanical Engineers for the award of a Graduate Research Fellowship to the first author. © 1999 John Wiley & Sons, Inc. <http://www.interscience.wiley.com>

© 2011 Transnational & Global Communication Systems, Vol. 3, No. 2, April 2011, Pages 113-120

How does the Agent know if  
'foot' is a measurement or a  
body part?



## TEXT TO SPEECH (TTS)

# What the Agent Sounds like

How does an agent know how  
to vocalize words/letters?

## End-to-End Neural Speech Synthesis

Alex Barron  
Stanford University  
adb@cs.stanford.edu

### Abstract

In recent years, end-to-end neural networks have become the state of the art in speech recognition tasks and they are widely deployed in industry (Amodei et al., 2016). Naturally, this has led to a proliferation of systems to do the opposite: text-to-speech synthesis from raw audio. Very recently, neural TTS systems have become highly competitive with their traditional counterparts, showing high naturalness scores in a variety of incarnations. The flexibility and relative simplicity of these models is compelling and their use for transfer learning (Arik et al.,

2017) is making them difficult to implement. In this paper, we release Tacotron architecture (Wang et al., 2017) which makes significant strides to overcome these obstacles and it is this system we present in depth.

### 2 Related Work

There currently exist four main approaches to neural networks for all or most of the speech pipeline. The first is DeepMind (van den Oord et al., 2016), a generative model for audio generation which uses a sequence of convolutions (with residual skip connections) to form a conditional probability for each audio sample. For TTS tasks, Wavenet

## Characterization and Incorporation of Emotions in Speech

T. V. Sagar, K. S. Rao, S. R. M. Prasanna, and S. Dandapani

**Abstract**—Emotional speech plays an important role for conveying the desired message. Emotions are manifested in both signal and text levels. In particular they are significant at prosodic level (i.e., prosodic level). In this paper four emotions are characterized (anger, compassion, happy and neutral) using the prosodic features such as duration, intonation, stress of pitch, and energy. The synthesis is performed on TTS Speech Under Simulated Emotion database. The learned knowledge with respect to each emotion is used to generate the prosodic features to feed the speech transformation from one emotion to another. In this paper an attempt is made for transforming speech from neutral to other emotions. The transformed speech after incorporating the emotion is evaluated by using the listening tests.

**Index Terms**— Anger, Compassion, Emotion, Happy, LP Neural Network, Prosodic features, Incorporation of emotion.

### 1. INTRODUCTION

Emotion is a subcategory of stress, and stress refers to speech spoken based on emotional or state induced or situational condition [1]. Emotions play an important role in human-human communication. Speech is one of the dominant modes that can simultaneously convey linguistic information along with the speaker's emotional state [1, 2]. The realization of a Text-to-Speech (TTS) synthesis system can express emotions and be an effective communication tool for people with disabilities and persons that are unable to talk and expressing their emotions. Towards this, many emotional speech synthesis systems were developed by creating the emotional speech corpus for the basic emotions such as anger, compassion and happy [3]. The paradigm of automatically recognizing the emotion from raw speech has grown with the increasing role of spoken language interfaces in human-computer interaction scenarios [2].

Integration of emotions in human-human speech communication and decision making made it desirable that communication technology interfaces should be able to

be able to produce speech with emotions. To achieve this, synthetic systems should acquire the knowledge specific to each of the basic emotions and then use that knowledge either for recognition or for synthesizing the emotion in speech [2-4].

Several studies have been made to identify different six conditions (emotions) that are present in speech [5-6]. Further, such knowledge has been used for processing speech stress, improve the naturalness of speech coding algorithms, increase the robustness of speech recognition systems [8, 9]. Conventional speech features (spectral and cepstral) were being used in neural networks for emotion categorization [9]. The results of quantitative prosodic analysis of a large corpus reveal emotion patterns indicate that there is a very influence of emotion on F0 contour (intonation pattern) [10]. Hidden Markov Model (HMM) is also used as an emotion classifier with short-term log-frequency power coefficients represent the speech signals [11]. Several pattern recognition methods using semantic signal information have been applied for automatic emotion recognition from speech. The semantic correlates include the following: (1) pitch contour features, (2) energy, (3) speech rate, (4) duration frequency, (5) voice quality parameters and (6) articulation parameter [12, 13]. A statistical model has been proposed for characterization and classification of different stress states (emotional) in the speech signal [14].

In this paper the basic emotions are characterized using prosodic features such as duration, intonation (F0 contour) and energy. The acquired characterization of emotions are used for transforming the speech from one emotion to the other. The prosodic information corresponding to each emotion incorporated in the Linear Prediction (LP) residual using knowledge of instances of significant variations. Towards significant evaluation refers to features of global domain (one of varied speech and onset of events like burst in one emotion speech). The reason for choosing the residual does for incorporating the specific prosodic information is that

# *Sounding* more natural

---

## **EMOTION**

Expressed through the timbre of voice and the mean pitch value



## **PROSODY**

Collection of things, such as intonation, stress, and rhythm.



## **PAUSES**

*Hmms, aahs* and fake pauses are added when processing takes too long.



# *Sounding more natural*

—

All the things I read about human communication etc how an agent can be made to sound even more like a human being.

## PIVOT

### INTONATION

There are many variations of passages of Lorem Ipsum available, but the majority.



### PROSODY

There are many variations of passages of Lorem Ipsum available, but the majority.

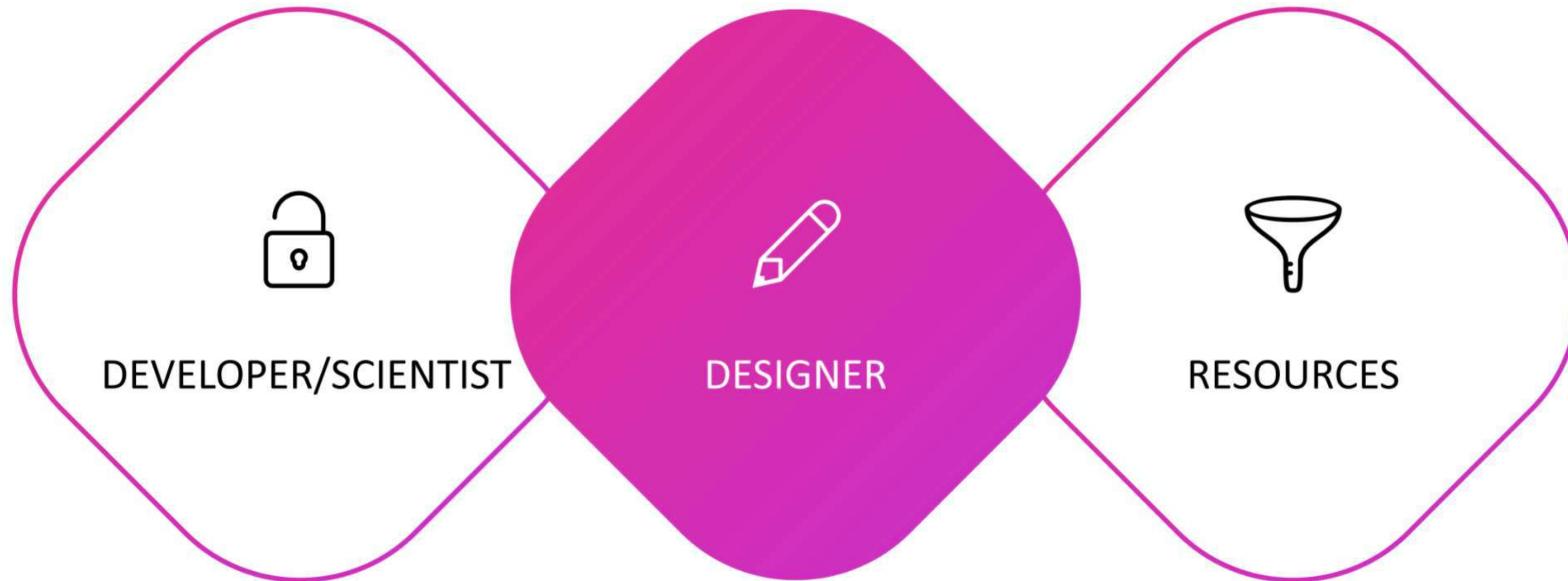


### PAUSES

There are many variations of passages of Lorem Ipsum available, but the majority.



# Why Pivot





**A DESIGNER'S  
GUIDE TO  
VOICE TECH**

Blog/Research

# Did anything productive come out of this?

A simple 101 guide for how agents speak and understand  
a language

# Features



## Understanding

- Automatic Speech Recognition
- Indexing

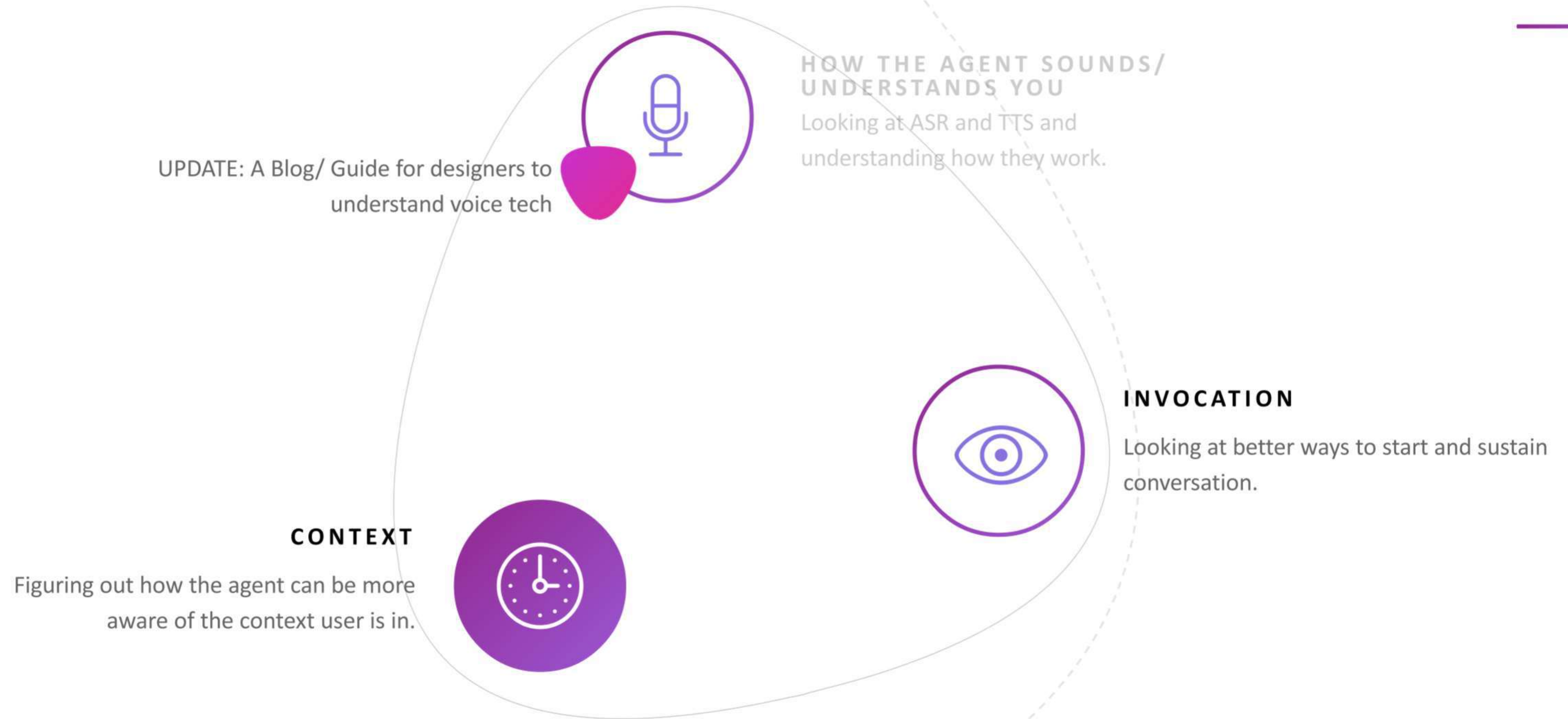


## Speaking

- Machine Learning
- Neural Networks
- Recurrent Neural Networks
- Long Short Term Memory
- Text-To-Speech (TTS)
- Concatenative TTS
- End-to-End Neural TTS

# What is more natural?

---



## PROACTIVE INTERRUPTION (CONTEXTUAL AWARENESS)

# When should the Agent interrupt?

**Hang on a Sec! Effects of Proactive Mediation of Phone Conversations while Driving**

Shamir T. Iqbal<sup>1</sup>, Eric Horvitz<sup>1</sup>, Yan-Cheng Ju<sup>1</sup> and Ella Mathews<sup>2</sup>

<sup>1</sup>Microsoft Research  
Microsoft Way, Redmond, WA 98052  
shamir, horvitz, yanjung}@microsoft.com

<sup>2</sup>California Institute of Technology  
1200 E. California Boulevard, Pasadena, CA  
mathews@caltech.edu

**ABSTRACT**

It is on cell phones while driving is a risky, yet increasingly common activity. State legislatures in the U.S. have laws that limit handheld phone conversations while driving, but the effects of hands-free conversations, which have been widely adopted, are less clear. Studies have demonstrated that the cognitive load of driving is a significant source of distraction, and that the likelihood of accidents. We explore in a study with a driving simulator the effectiveness of driving and mediation of conversations on conversations while driving. We study the use of messages, including speaking, context, and hand phone calls on field. We found that such hand driving errors and that alerts during driving errors were more effective than general alerts, and that a system capable of mediating conversations while driving could play a role in reducing drivers' attention on the road while conversations.

**KEYWORDS:** driving, context-sensitive systems, cell phones.

**ACM KEYWORDS:** interaction interfaces and presentation (e.g., HCI); tasks.

**CCS:** systems, human factors, measurement.

**INTRODUCTION**

It is on cell phone conversations while driving is a very common practice. Recognizing the danger to safety that simultaneous driving and conversing poses and others, several U.S. states have enacted laws that limit handheld phone conversations [1]. Yet studies show that using hands-free usage may not reduce the risk [22, 23], and may tell drivers into a false sense of security that half of all U.S. states forbid novice at school bus operators from any use of cell

phones while driving [1]. However, general bans on conversations while driving are unlikely given the ubiquity of modern mobile communication. Thus, we study that enable safer driving while communicating, by providing for automated mediation of conversations. Automated mediators of human attention in real situations has been a focus of the research community in the desktop domain [11, 12]. Our goal is to explore the effectiveness of similar mediated driving situations, where a context-sensitive system might justifiably share information with the people they are speaking with to provide the participants with useful guidance about the situation [27]. Establishing such mutual understanding with joint regulation of conversations, enabled to better focus on maintaining driving safety.

As a first step, we conducted a controlled study with a driving simulator. Different sets of mediated conversations were presented to drivers while they were driving. We presented the conversations with context-sensitive mediation of the effectiveness and acceptability of long conversation messages, pushing the cell phone conversation of the intervention, exploring outcomes of that through conversations of critical time benchmarks versus other times.

Our results showed that interventions positively affected driving, reducing driving errors compared to the baseline. The positive effect on driving safety was more pronounced for drivers in the study of the first experiment. Subjective feedback showed that also interventions to be useful for more situations. The preliminary results towards having a shared sense of a conversational break followed by the cell phone. We did not drive participants to conversations of more than about the time of the intervention, suggesting

## Stanford researcher examines how people perceive interruptions in conversation

Stanford doctoral candidate Katherine Hilton found that people perceive interruptions in conversation differently, and those perceptions differ depending on the listener's own conversational style as well as gender.



### BY ALLIE HARRINGTON

We all know that unpleasant feeling when we're talking about something interesting and halfway through our sentence we're interrupted. "Wait, what's for dinner?" "Good."

But does that really go into your mind? Whether or not, one person interrupted another depends on what you said, according to new research from Stanford's Katherine Hilton, a doctoral candidate in linguistics.

"What people perceive as an interruption varies systematically across different speakers and speech acts," said Hilton, who is also a graduate fellow at the Stanford Center for Language Acquisition. "Listeners' own conversational style influences whether they interpret simultaneous, overlapping talk as interruptive or conversational. We all have different opinions about how a good conversation is supposed to go."

Using a set of carefully controlled scripted audio clips, Hilton surveyed 6,000 American English speakers to better understand what affects people's perceptions of interruptions. She had participants listen to audio clips and then answer questions about whether the speakers seemed to be friendly and engaged, listening to one another, or trying to interrupt.



Stanford scholar Katherine Hilton surveyed 6,000 American English speakers to better understand what affects people's perceptions of conversational interruptions. Image credit: J.A. Govea

## Extensive Literature Review

How can an agent know when the best time to interrupt is?

## Evaluating Proactive and Reactive User Interface Assistants

by Richard Catrambone & John Stasko

Computing / School of Psychology & GVU Center  
Georgia Institute of Technology  
Atlanta, GA 30332

rc7@prism.gatech.edu, stasko@cc.gatech.edu

We study the ability of an anthropomorphic interface assistant to help people learn with a specific focus on assessing proactive assistant behavior. Participants used a text editing system that used keypress combinations for invoking the assistant. Participants were directed to make a set of prescribed changes to a document with the assistant that would hear and respond to questions orally, or an assistant that would additionally make proactive suggestions. Anecdotal evidence suggested that the assistant did not enhance performance and would be viewed as intrusive. Our study was performed similarly on objective editing performance (completion time, recall), while the participants in the latter two conditions strongly felt that the assistant was intrusive.

**KEYWORDS:** assistant, anthropomorphism, empirical study, text editing, help, evaluation

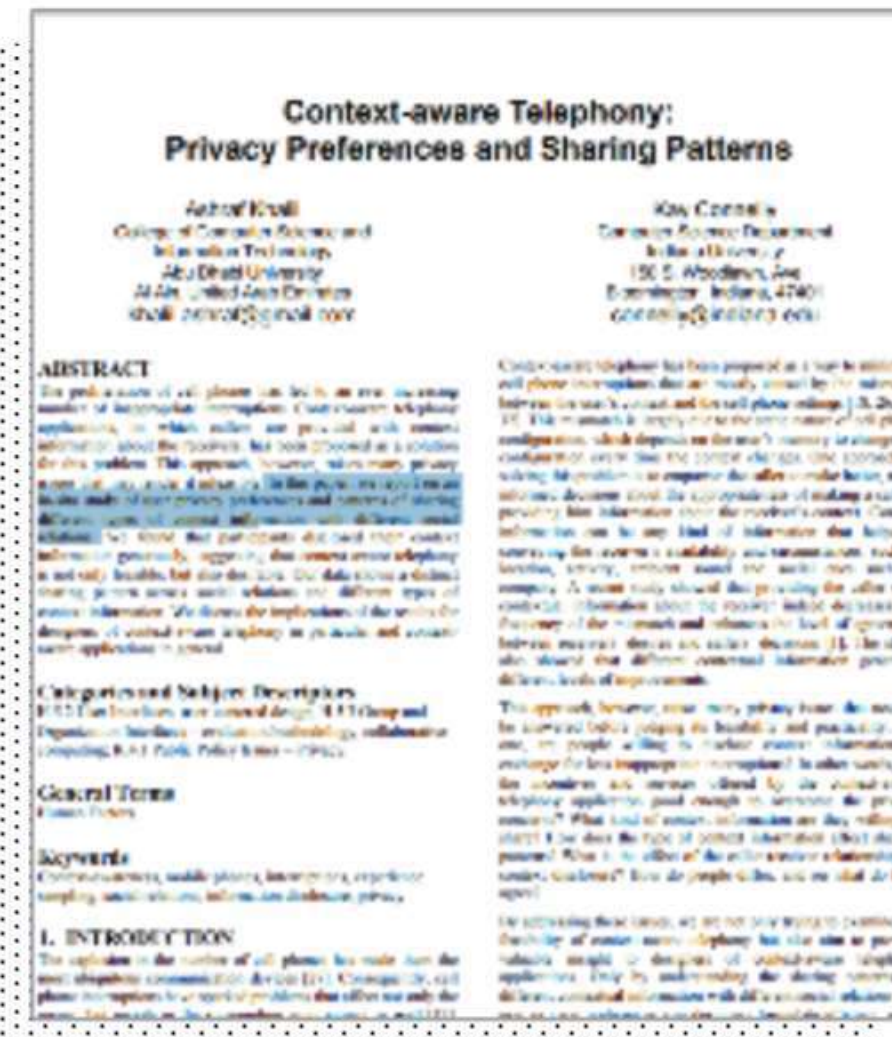
characters or assistants in user interfaces (Laurel, 1990). In this paper, we study the utility of an agent acting as a proactive assistant (Lanier, 1995; Cassell, 2000; or interface design

Our overall research program involves a systematic study of the factors influencing success. These factors are based on prior research and considerations about human performance. In this research, we study the utility of an agent acting as a proactive assistant, helping users perform tasks. The agent is presented through a 3D "talking head" face on a computer display. In the present study, we specifically

## PRIVACY CONCERNS (DATA RECORDED)

# What are the users fine with sharing?

Would a user be okay with giving their location data if it meant getting more relevant notifications?



## ABSTRACT

The proliferation of cell phones has led to an ever increasing number of inappropriate interruptions. Context-aware telephony applications, in which callers are provided with context information about the receivers, has been proposed as a solution for this problem. This approach, however, raises many privacy issues that may render it infeasible. In this paper, we report on an in-situ study of user privacy preferences and patterns of sharing different types of context information with different social relations. We found that participants disclosed their context information generously, suggesting that context-aware telephony is not only feasible, but also desirable. Our data shows a distinct sharing pattern across social relations and different types of context information. We discuss the implications of the results for designers of context-aware telephony in particular and context-aware applications in general.

## Categories and Subject Descriptors

H.5.2 User Interfaces- user-centered design, H.5.3 Group and Organization Interfaces – evaluation/methodology, collaborative computing; K.4.1 Public Policy Issues – Privacy.

## General Terms

Human Factors.

## Keywords

Context-awareness, mobile phones, interruptions, experience sampling, social relations, information disclosure, privacy.

# Being *aware* of context

---

## **PROACTIVE ASSISTANCE**

Proactively offering help to the user at the right point can be a huge help



## **PRIVACY PREFERENCES**

Depends of the context of use.



## **INTERRUPTIONS**

Badly timed interruptions can be extremely harmful.





Concept

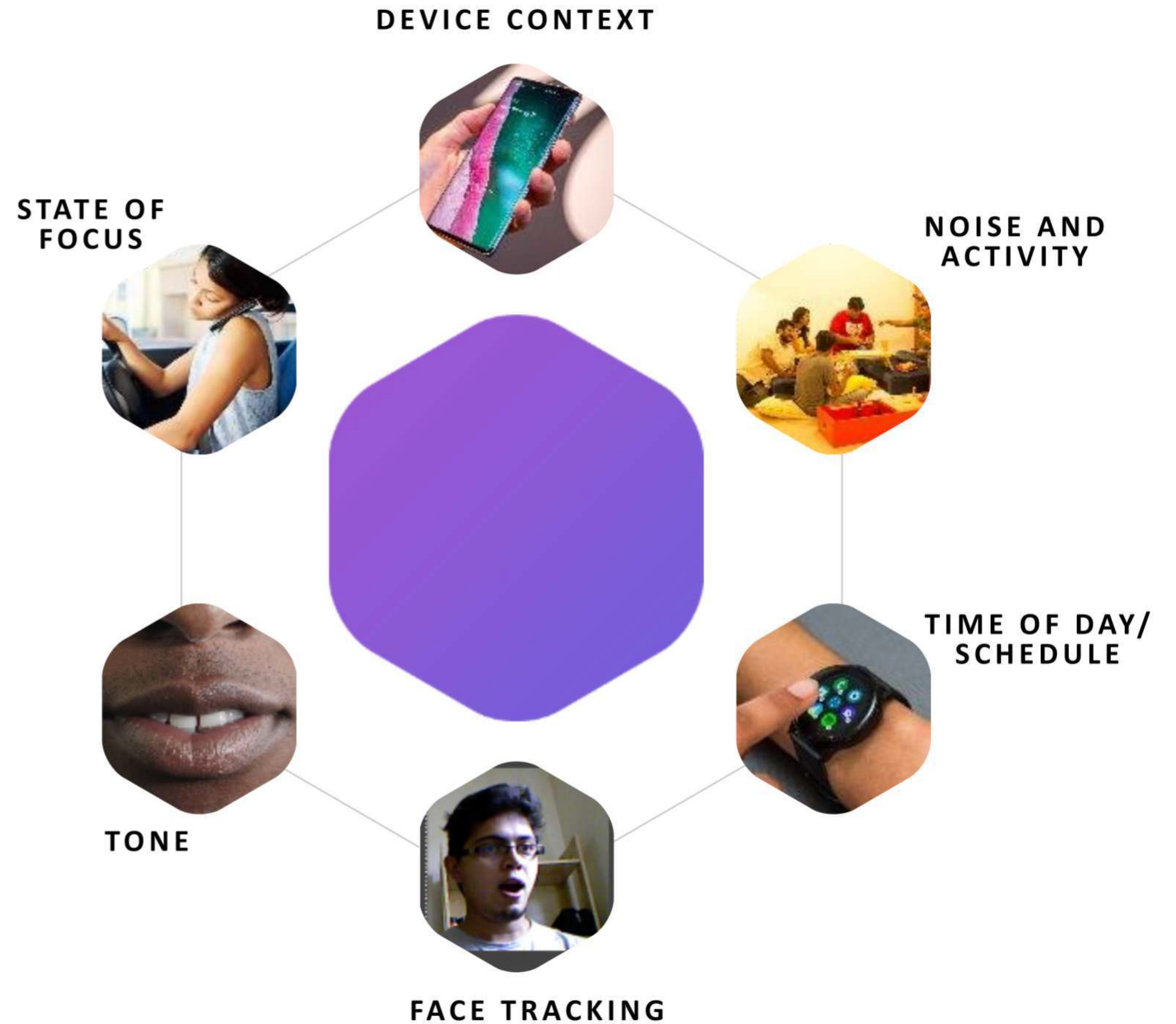
# A new kind of Agent

---

An assistant in a Multi-Device Environment (MDE) which understands the context you are in.

# Context-Aware Agent

In an MDE., a combination of these parameters will be taken by the agent.

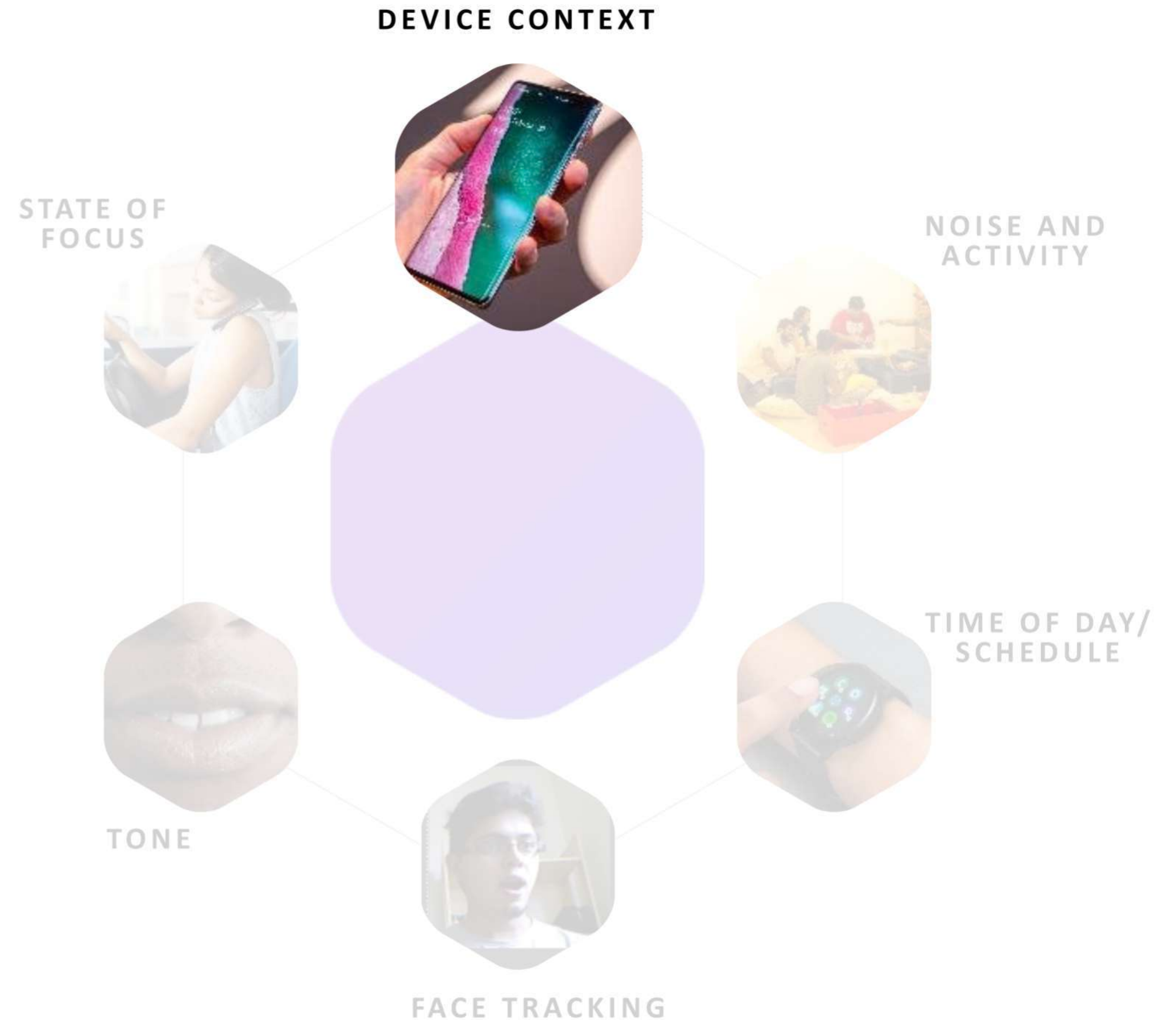


\*Multi Device Environment

*User was using the phone, then kept it on the table.*

*User asks agent to send them their flight tickets.*

*Agent recognizes that the last device used by user was phone and sends it there*

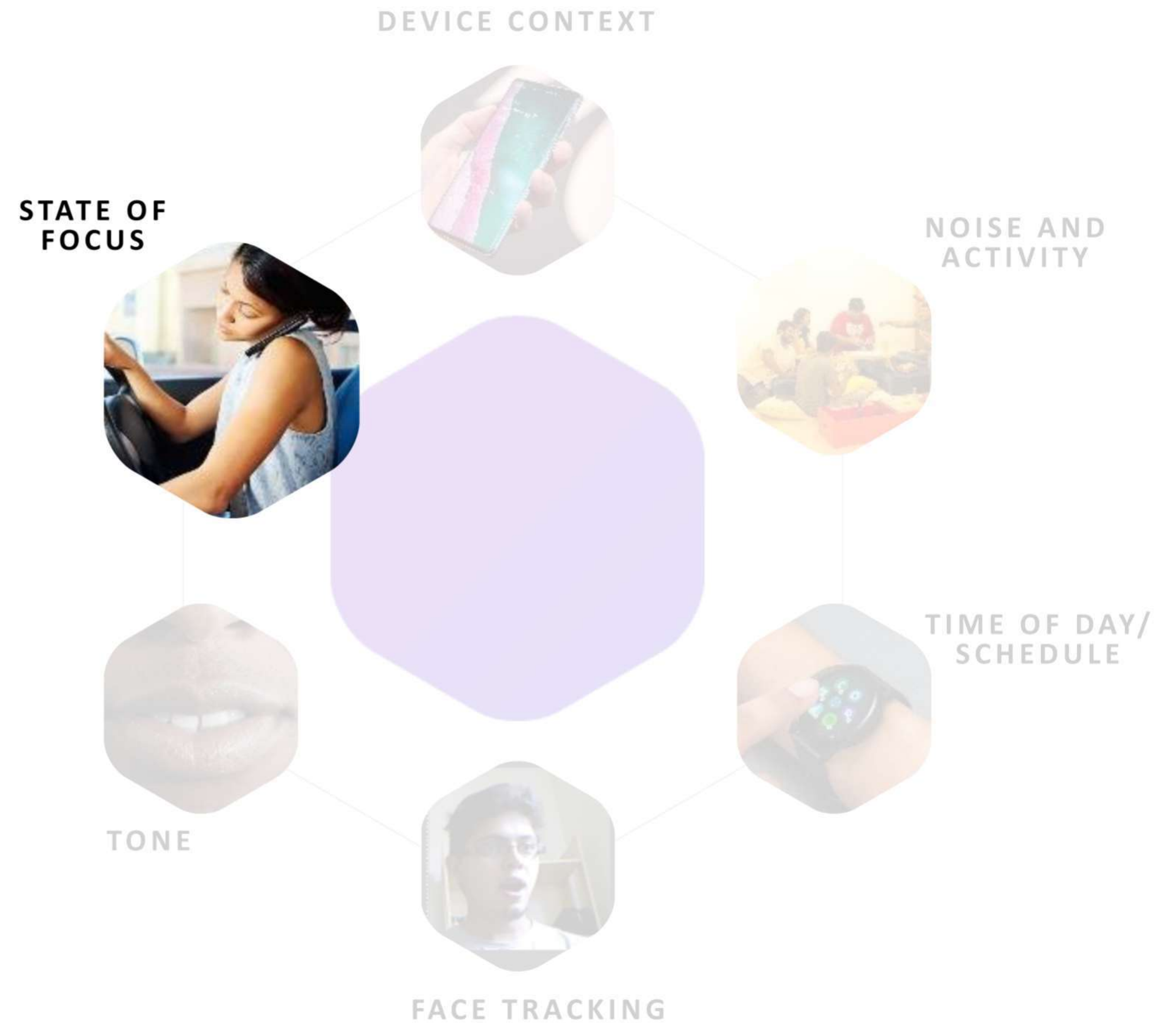


\*Multi Device Environment

*User is making a long presentation/report*

*User has previously asked agent to notify*

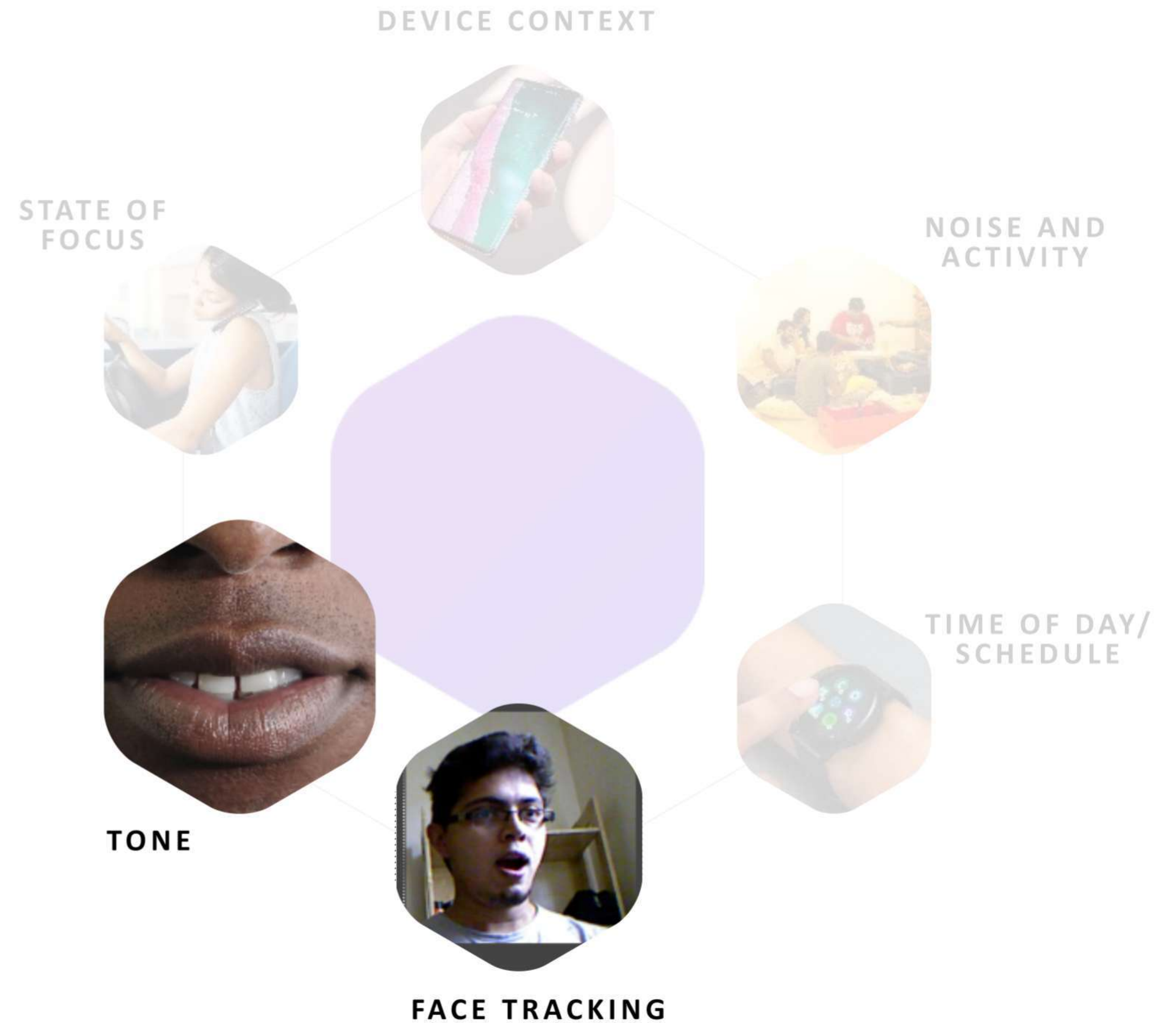
*Agent recognizes the optimum time to notify  
the user*



*User gets some bad news*

*Agent recognizes the user's emotional state and  
assumes the appropriate tone*

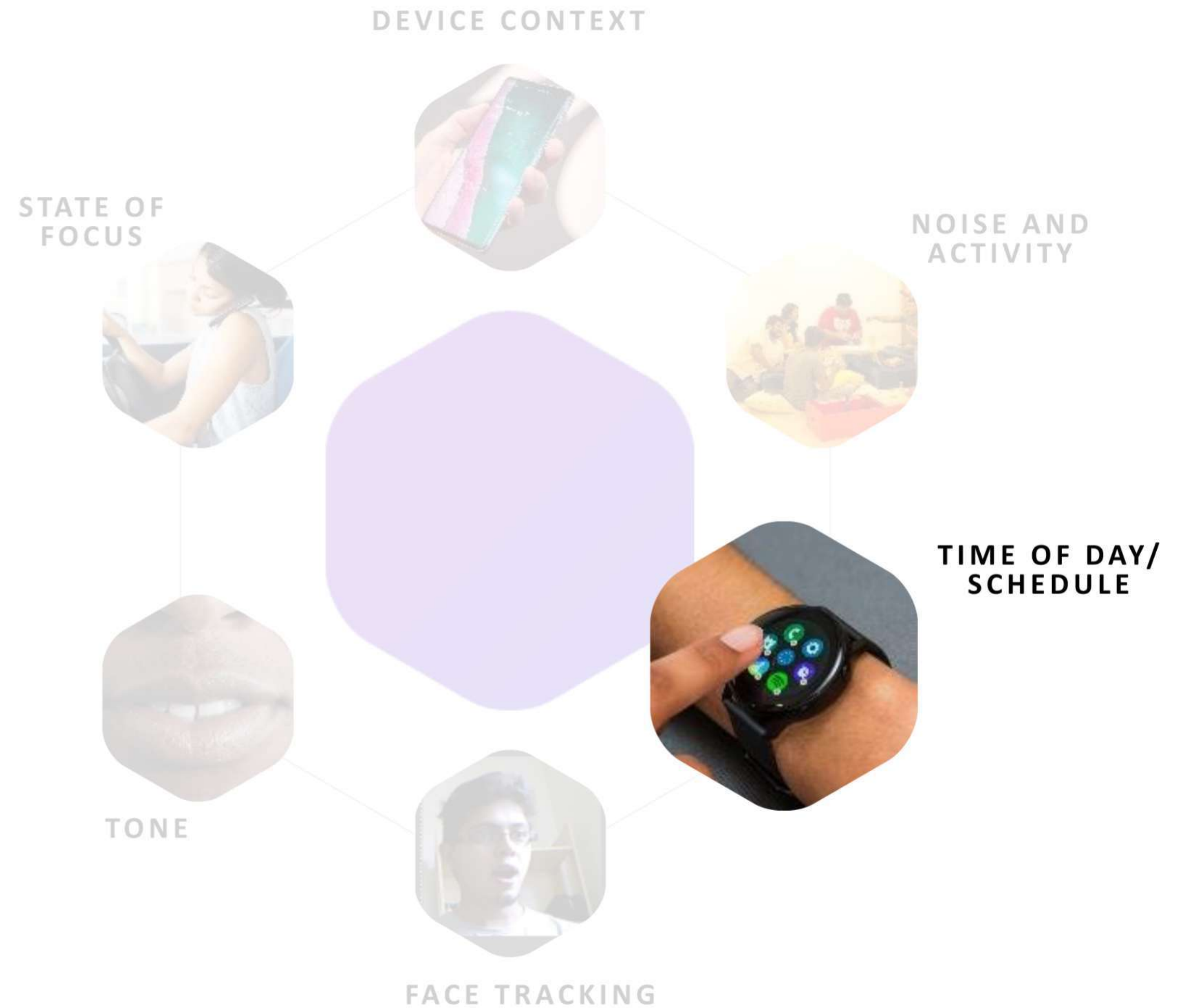
*Agent acts sensitive to the situation*



*User is in a meeting*

*Agent recognizes this based on schedule and  
diverts calls*

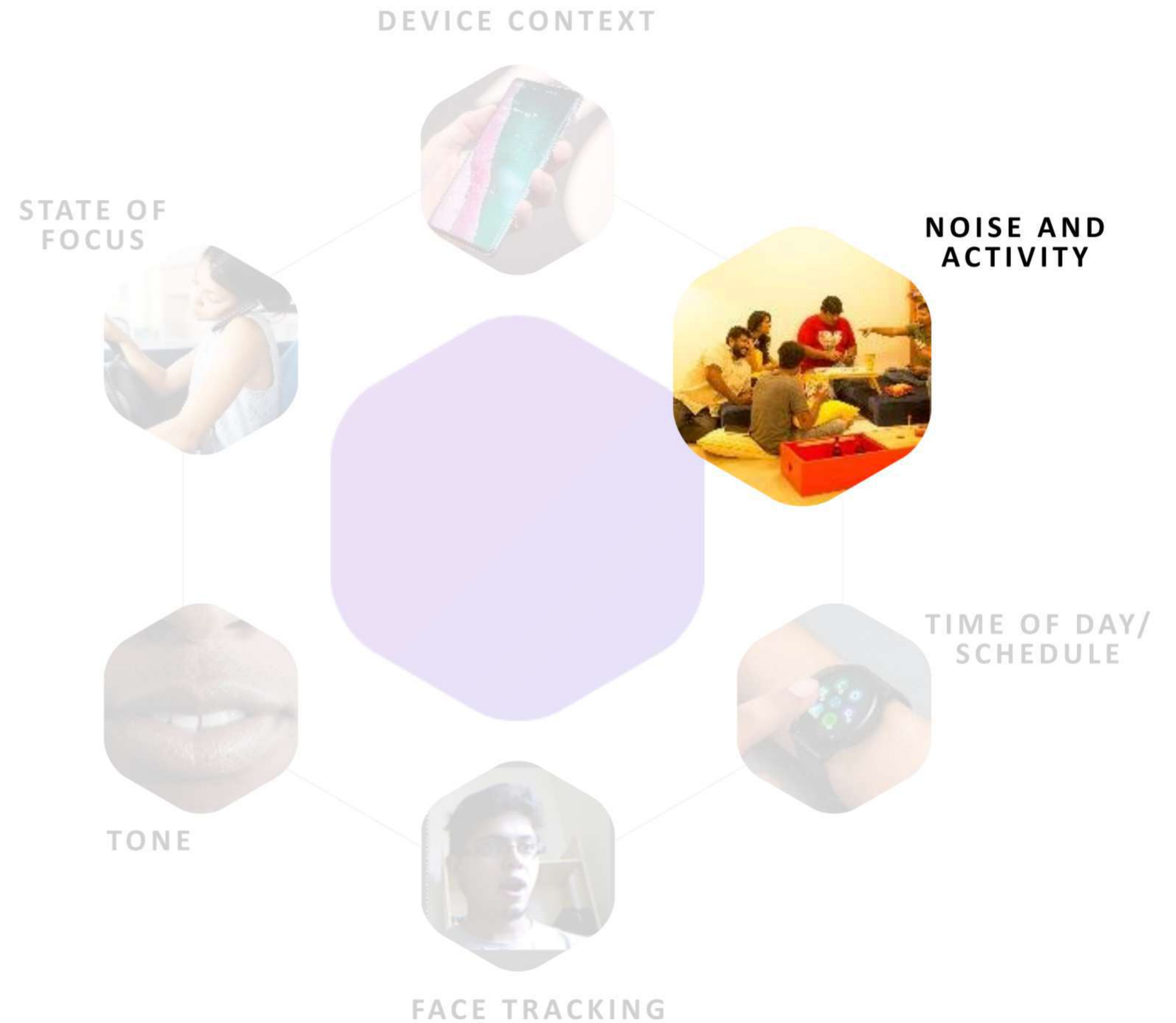
*Agent alerts user on missed calls from  
important contacts*



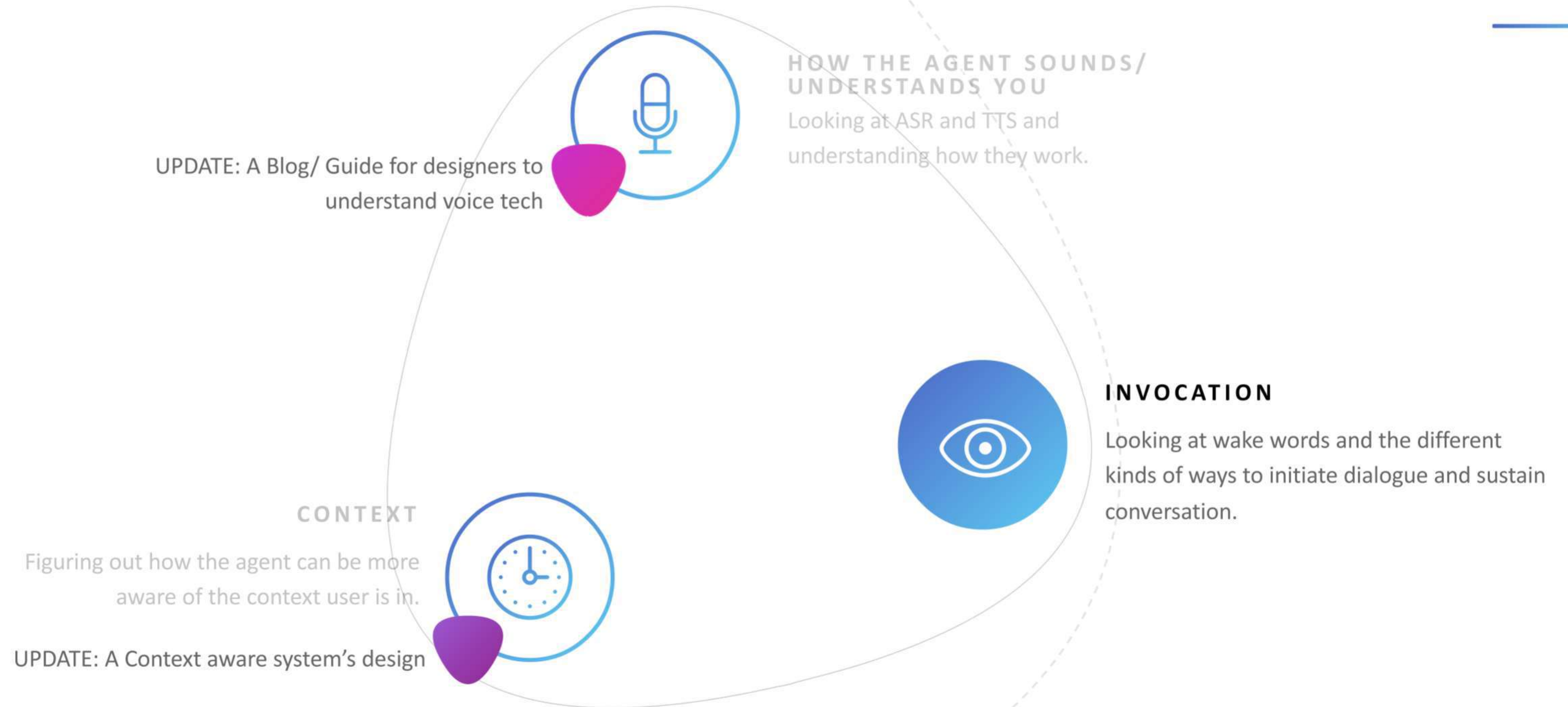
*User has friends over*

*User has earlier told Agent to read emails at  
8PM everyday*

*Agent recognizes context and doesn't do so*

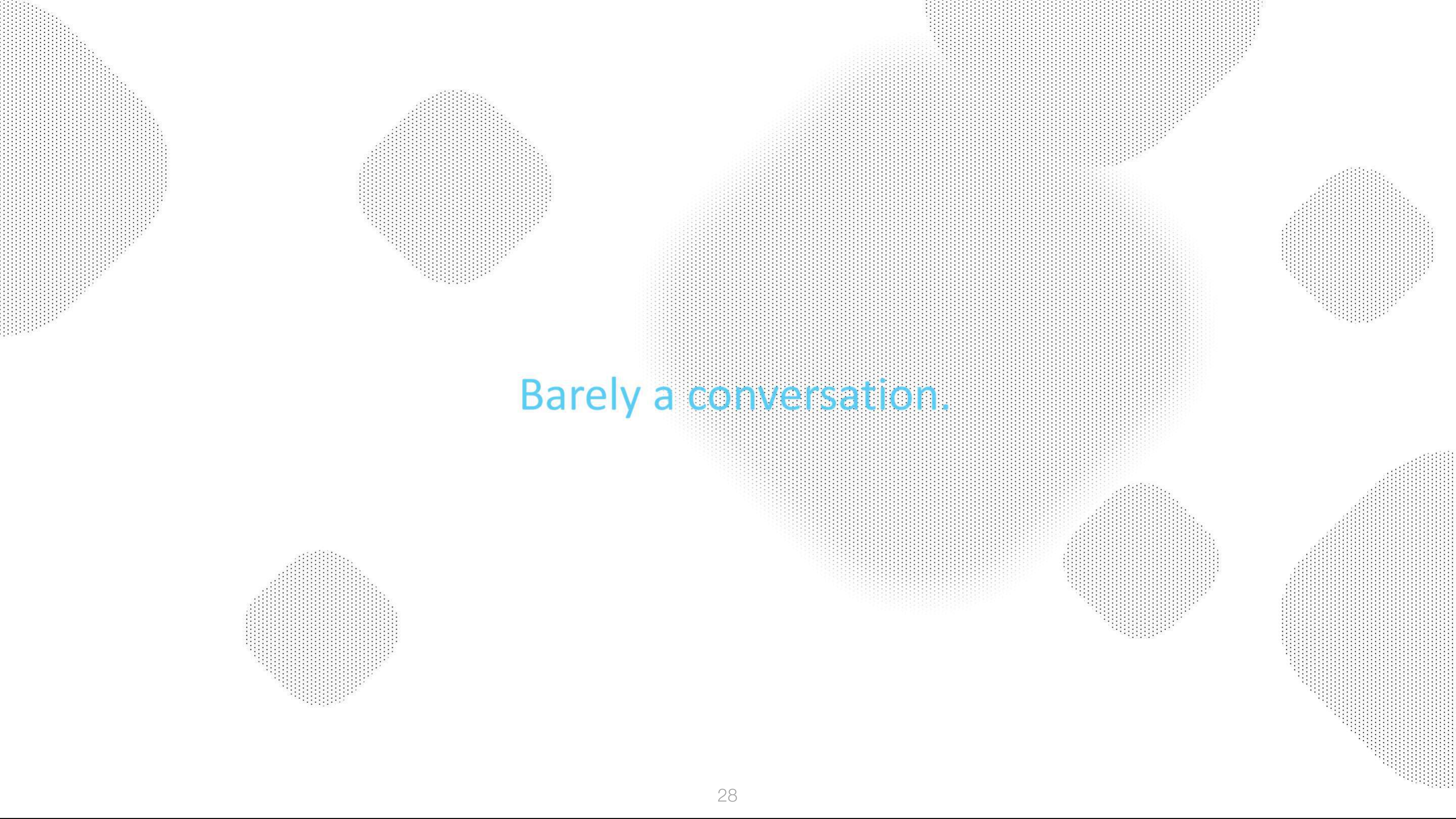


# What is more natural?



# Today's Conversational Assistants

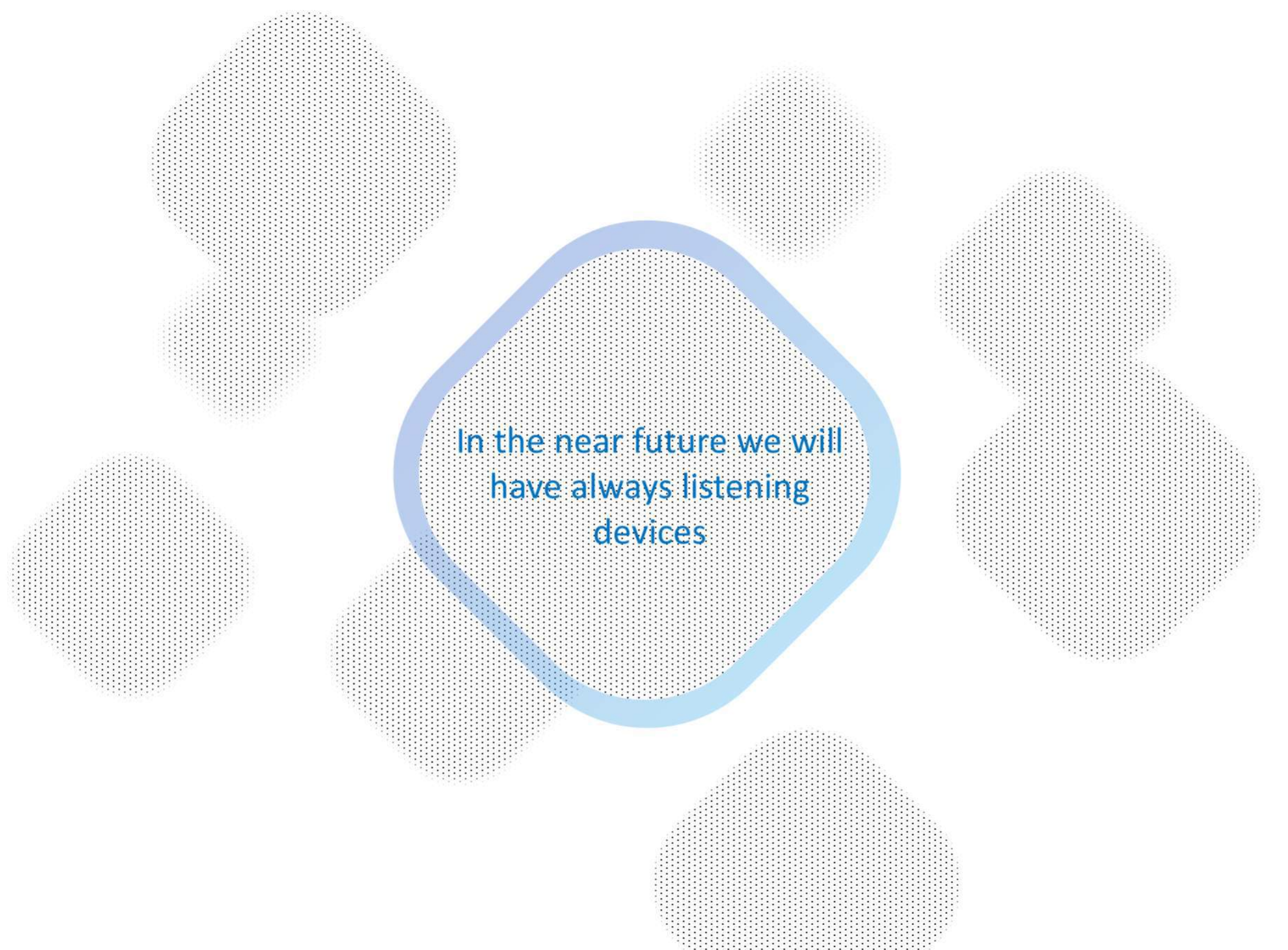
- Alexa, what's the weather in Seattle on Sunday?
- Hey Cortana, open Netflix.
- OK Google, how tall is Mount Everest?



Barely a conversation.

# Trends



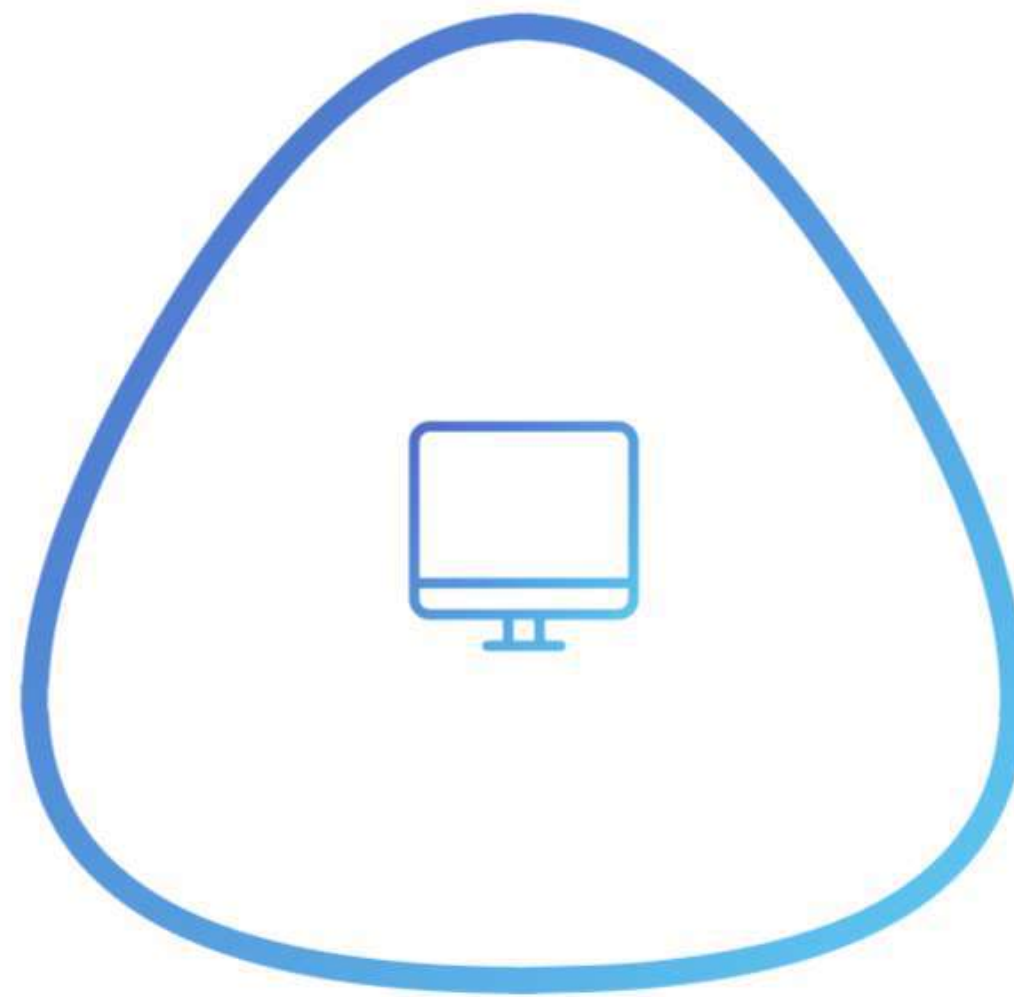


In the near future we will  
have always listening  
devices

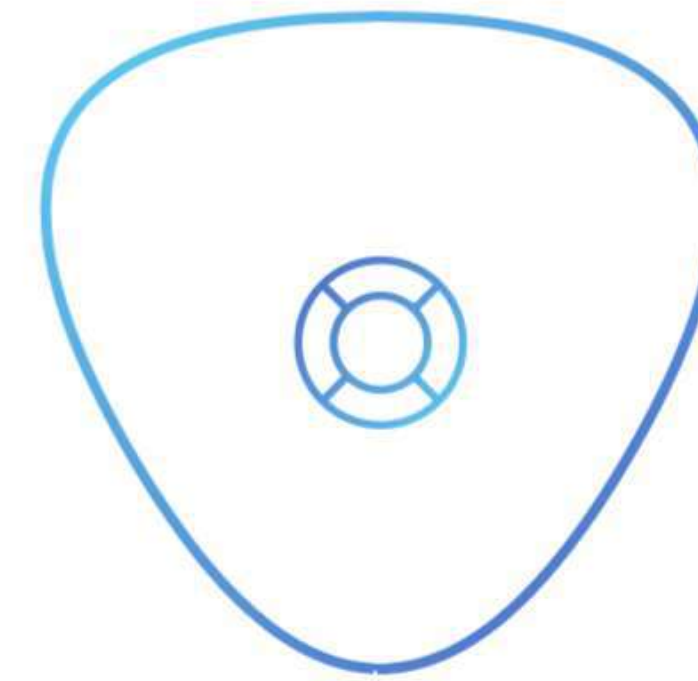
# UX Process



**FEEL IN CONTROL**



**TALK NATURALLY**



**MAKE TASKS EASIER**



## Conversation Guidelines

Concept

# Talking to an Always-On Agent

Conversation Design Guidelines for such an agent

# Design Principles

01

## **Maximize Efficiency**

Allow non-wake word  
related invocations

02

## **Accommodate Errors**

Foresee new kinds of  
errors

03

## **Maintain Transparency**

If user asks, tell them why  
certain decisions were made

04

## **Ensure Clarity**

User should be on the  
same page at all times

05

## **Provide Control**

Make the user feel  
comfortable and in-control

01

# Maximize Efficiency

## USER INPUT

*Yooooo! I'm in the mood for a movie. Are there any Kabir Singh screenings nearby?*

## SPOKEN PROMPT

*Yeah there are 3 screenings nearby, at 1:30, 4 and 6:30.*

## USER INPUT

*Cool, book the 6:30 one*

## SPOKEN PROMPT

*Alright cool! Booking the show for 6:30 PM at Metro Cinemas, today.*

## DOs

Allow the user to call out/invoke the agent in other "fuzzy" terms, than it's standard wake word.

## USER INPUT

*Hey Samsung! I'm in the mood for a movie. Are there any Kabir Singh screenings nearby?*

## SPOKEN PROMPT

*Yeah there are 3 screenings nearby, at 1:30, 4 and 6:30.*

## USER INPUT

*Hey Samsung! Book the 6:30 one*

## SPOKEN PROMPT

*Alright, booking the show for 6:30 PM at Metro Cinemas, today.*

## DONTs

Don't require repeated usage of the wake word as it can get annoying.

# Accommodate Errors

## USER INPUT

Hey Samsung! Are there any Kabir Singh screenings nearby?

## SPOKEN PROMPT

Cool, for what time?

## USER INPUT

<silence?>

## SPOKEN PROMPT

What show timings should I look for?

## USER INPUT

<silence?>

## SPOKEN PROMPT

Do you still want me to look for Kabir Singh movie screenings near you?

## DOs

Restate the question in a way that the user has one last chance to respond.

## USER INPUT

Hey Samsung! Are there any Kabir Singh screenings nearby?

## SPOKEN PROMPT

Cool, for what time?

## USER INPUT

<silence?>

## SPOKEN PROMPT

Sorry I didn't get it please repeat it or try again in a few minutes.

## USER INPUT

<silence?>

## SPOKEN PROMPT

Sorry I didn't get it please repeat it or try again in a few minutes.

## DONTs

Don't suggest trying again in a while if it's unlikely to fix the issue.

# Maintain Transparency

---

## USER INPUT

*Ummm... Samsung, why did the lights go out?*

## SPOKEN PROMPT

*Kabir used the Philips Hue app to set a timer for these lights yesterday. The timer turns your downstairs lights off at 9PM every day.*

### DOs

Explain to the user clearly why some event happened and what app used to execute it so if they want to change things they can easily do so.

## USER INPUT

*Ummm... Samsung, why did the lights go out?*

## SPOKEN PROMPT

*I'm not sure about that.*

## SPOKEN PROMPT

*Philips Hue's legal team doesn't allow me to. disclose that information.*

### DONTs

Not giving the users a clear answer may make them feel uncomfortable, scared and out of control of their own devices. Putting the blame on others can also backfire, offer them the answer and take responsibility.

04

# Ensure Clarity

## USER INPUT

Hey, I added that number for Jenny the other day right?

## SPOKEN PROMPT

Yep! You added +91 987654321

## USER INPUT

Send this to Saahil.

## SPOKEN PROMPT

Sending Jenny's new number to Saahil through text.

## DOs

Reiterating the points acknowledge that the agent heard the user correctly.

## USER INPUT

Hey, I added that number for Jenny the other day right?

## SPOKEN PROMPT

Yep!

## USER INPUT

Send this to Saahil.

## SPOKEN PROMPT

Done.

## DONTs

Without the reiteration the user might feel unsure about if the agent has heard them correctly.

05

# Provide Control

---

## **BARGE IN**

Allow the user to start speaking while the agent still is going on and on

## **RECOVERY**

Allow the user to feel undo any command they had given, easily

## **CUT OFF**

Make it easy for the user to make the agent stop listening

Domestic Abuse Relief

Solving for some  
hard questions

Stories

Sr. Citizens

Long term Memory

Loneliness

Possibilities?

AI / Judgement?

# Future Scope

What's next for the project?



# Key Learnings



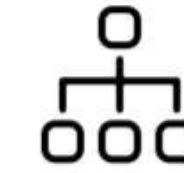
## **EMBRACE AMBIGUITY**

When exploring such a new field, almost nothing has been defined. And that's okay.



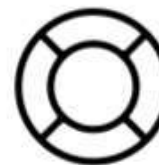
## **MISTAKES**

It's alright to make mistakes as long as you learn from them and move ahead.



## **HOW LARGE COMPANIES WORK**

How shipping some feature or a product happens at such a scale



## **MAKE TIME FOR FUN AS WELL**

Loving your work is fine, but if you overwork yourself... You'll get burnt out.



## **BE FLEXIBLE**

Move fast, break things but learn from your mistakes and the advice of others as well!



## **OPEN ENDED PROJECTS**

Doing such an open ended project from scratch... Is really hard

# Thank You

LOOKING FORWARD TO YOUR  
FEEDBACK