

VOICE BASED SELF HELP SYSTEM

USER EXPERIENCE VS ACCURACY

Dr. Sunil Kumar Kopparapu

**TCS Innovation Lab – Mumbai, Thane, Maharashtra
INDIA**

December 8, 2008

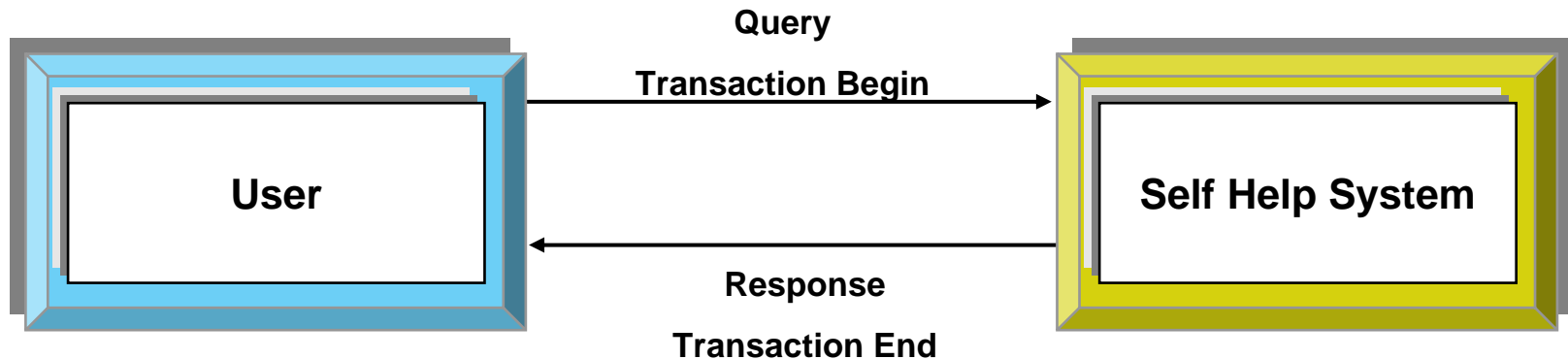
Location (Thane, Maharashtra, India)



What is a Voice Based Self Help System?

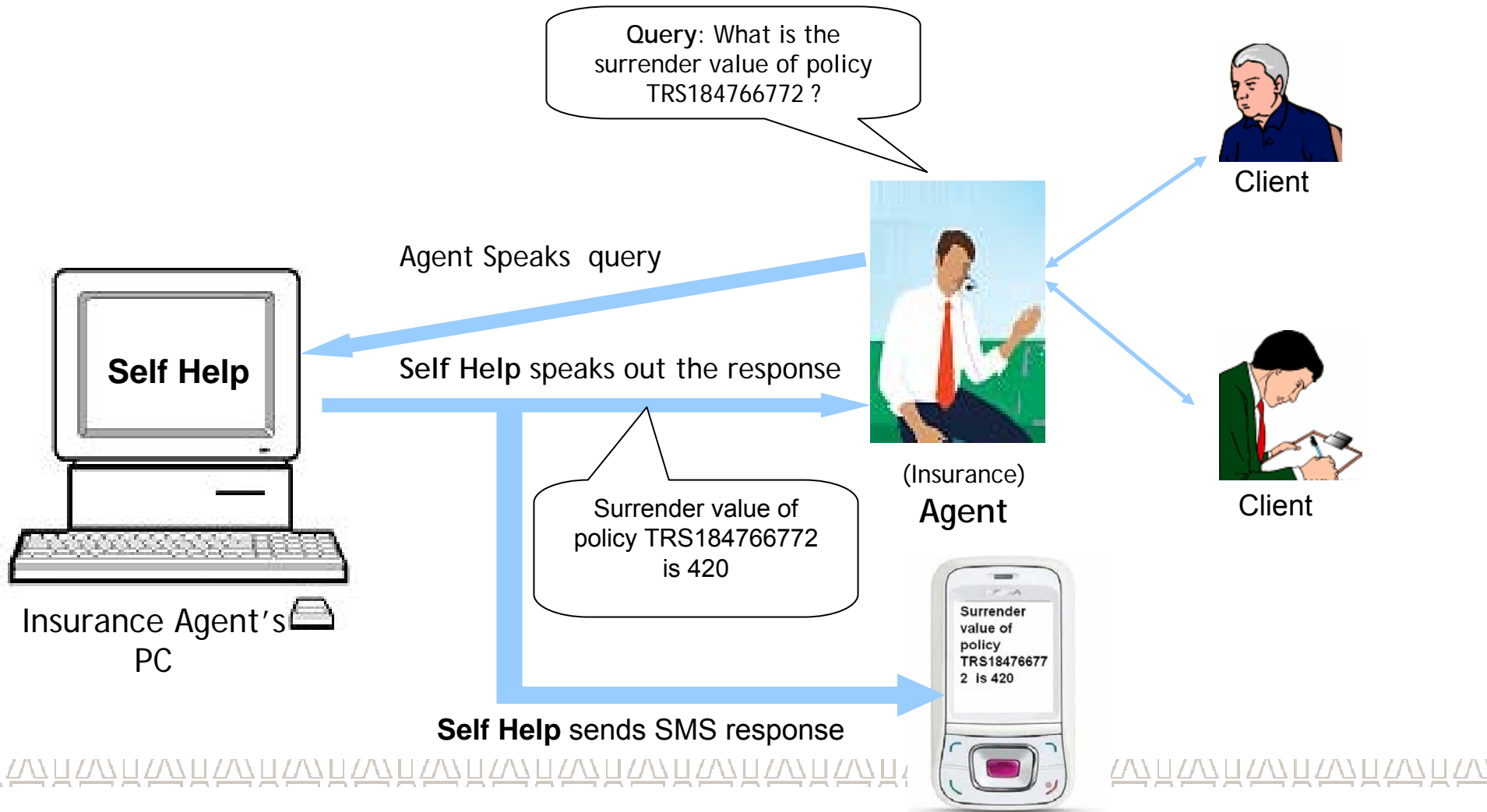
Self Help Systems, typically are automated systems that enable users to help themselves.

When the medium of transaction is voice or speech, it is called a **Voice Based Self Help System**.



An Example

Voice Enabled Self Help System: An Example



What is a Successful Speech Solution?

- For a speech solution to work in *field* there are two important parameters that needs to be satisfied
 - the **accuracy of the speech recognition engine** and
 - the overall **user experience**
- **Best solution**
 - better speech recognition accuracy
 - better user experience
- **Feasibility of the best solution**
 - **Speech Recognition Performance**
 - Better if user **restricted** to speak in a limited way
 - Bad if the user has **no restriction** on what he can speak
 - **User Experience**
 - Better if the user has **no restriction** on what he can speak
 - Bad if user **restricted** to speak only in a certain way
- User Experience and Speech Recognition performance contradict

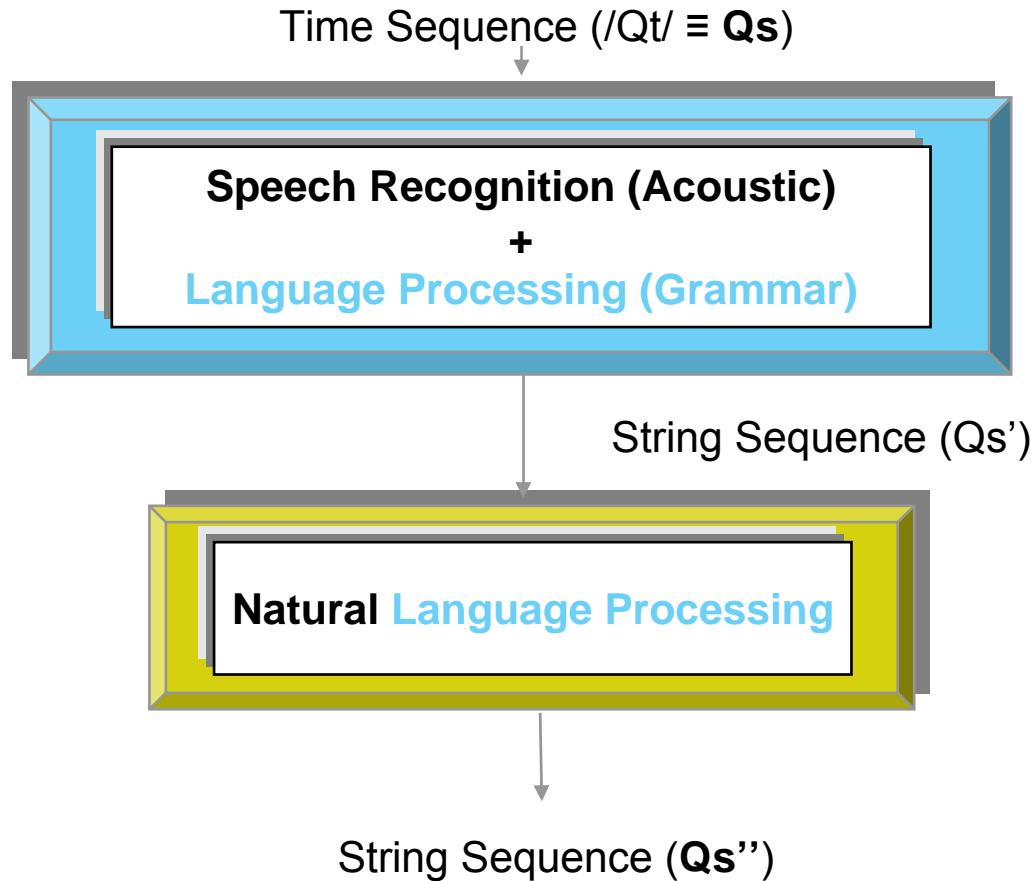
Optimal solution lies in between!

Measure: User Experience, Speech Recognition

- **User experience** is measured by the freedom the system gives the user in terms of
 - who can speak (speaker independent),
 - what can be spoken (large vocabulary) and
 - how to speak (restricted or free speech)
- **Speech recognition** accuracies are measured as
 - the ability of the speech engine to convert the spoken speech into exact spoken text.
 - Word Error Rate



Overall Performance of Speech Solution



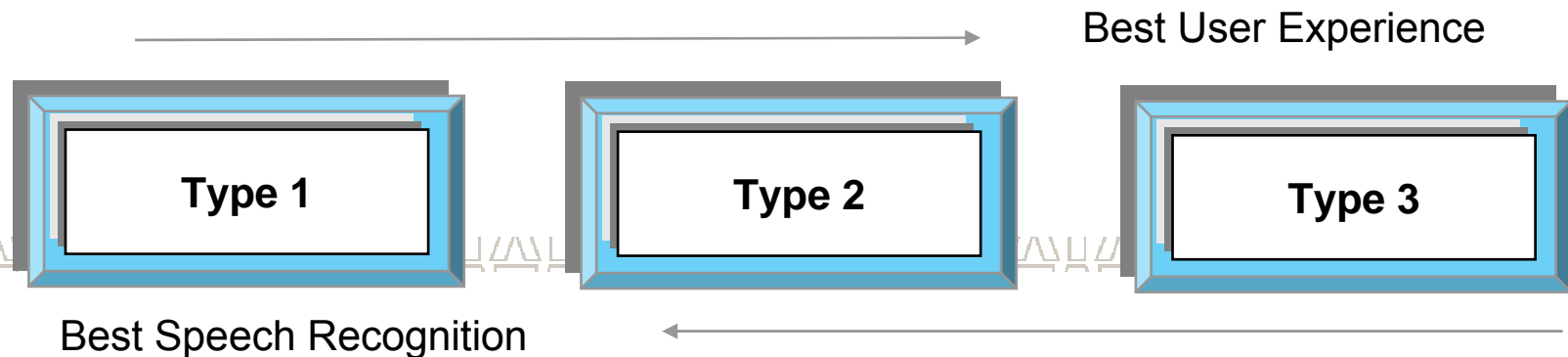
- Ideally Qs'' should be equal to Qs
- Overall performance of the system depends on how close Qs'' is to Qs in the sense of *intent*

Language Processing in Speech Systems

- Language processing happens both in **speech recognition module** and the **natural language processing module**.
- The language processing in speech module is **necessary** to achieve reasonable recognition performance
 - Language processing (or grammar) used in speech module is tightly coupled with the acoustic models (degree of configurability very limited)
- Language processing in speech recognition module is **not sufficient**
 - Because of the tight coupling; need separate language processing
 - Why? A relatively high degree of configurability possible
- Question: Do we need language processing in both the modules?
 - isolate speech processing and language and have language processing only in language processing module or
 - combine all language processing into speech module and do away with separate language processing module completely.
 - Probably there is an optimal combination possible which produces a usable speech based solution.
- **Usable** speech based solution
 - A combination of speech module and language processing module such that their combined effort enables Q_s be as close as possible to the desired Q_s

Language Processing in Speech Recognition Module

- User experience can be controlled by changing the grammar (language processing) of the speech recognition engine
- Type 1 (High User experience)
 - Allow the user to speak anything (free speech; dictation like)
 - Allow any user to speak (speaker independent)
- Type 2 (Moderate User Experience)
 - Allow the user to speak a large vocabulary – but not everything
 - Configure for a particular user (speaker dependent)
- Type 3 (Low User Experience)
 - Allow the user to speak only what is expected (in terms of grammatical correctness)
 - Ask to repeat if the user does not adhere to the preset format
 - Configured to a particular speaker (Speaker dependent)



Need for Separate Language Processing Module

- Ideal System
 - Open Speech (free speech - speak without constraints),
 - Speaker independent (different accents, dialects, age, gender)
 - Environment independent (office, public telephone)
 - Accurate speech recognition
- For best User experience
 - Type 3 speech recognition module
 - ‘poor’ performance of the speech module
- For best speech recognition accuracies
 - Type 1 speech recognition module
 - ‘poor’ user experience
- Best of both; demands
 - Need for a separate Language Processing Module to take care of poor performance of the speech engine while retaining the high user experience

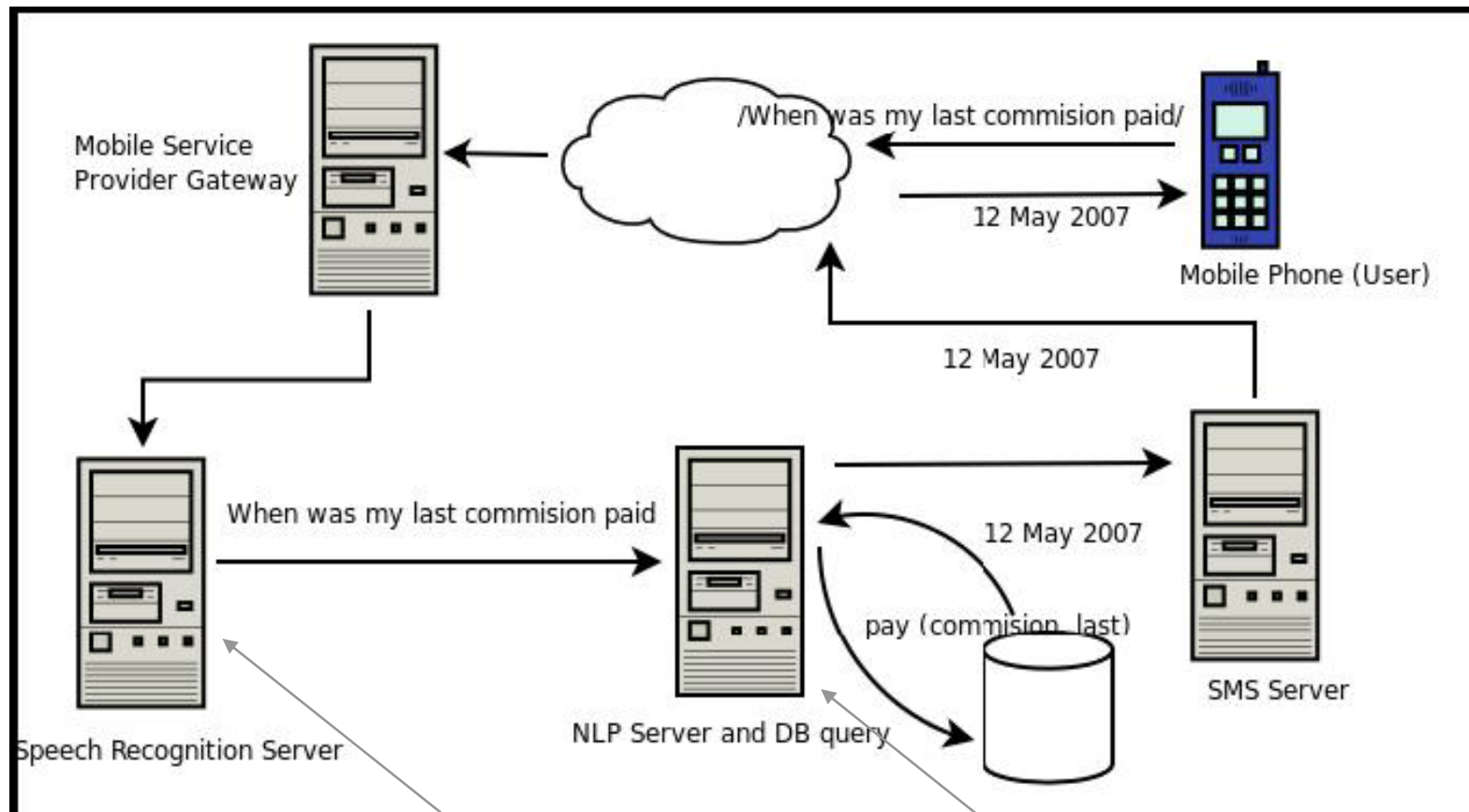


Experimental Setup

- A self help system was built for
 - Insurance agents (act as intermediaries between insurance company and clients)
 - To keep track and seek information for their clients (policy status, maturity status, change of address request)
 - Information related to the insurance agent (When was my last commission paid?)
- Why? build a self help system
 - to enable the insurance company to lower the use of human call center
 - providing up-to-date and dynamic information needed by agents
 - both this together provide better customer service.
- Using
 - Speech Recognition engine of Microsoft using Microsoft SAPI SDK
 - Language Processing module was developed in-house



Voice Based Self Help System: Block Diagram



Speech Recognition (Acoustic)
+
Language Processing (Grammar)

Natural Language Processing

Type 1: Bad User Experience Best Speech Recognition

<GRAMMAR>

<RULE NAME="F_3" TOPLEVEL="ACTIVE">

<o> <RULEREf NAME="StartTag"/> </o>

<RULEREf NAME="KeyConcept"/>

<o> of <o> the </o>

<RULEREf NAME="Keyword"/>

<o> <RULEREf NAME="EndTag"/> </o>

</RULE>

<RULE NAME="StartTag">

<P> What is the </P>

<P> Can you please send me</P>

<P> Can you tell me </P> </RULE>

<RULE NAME="KeyConcept">

<P> Surrender Value </P>

<P> Address Change </P> </RULE>

<RULE NAME="Keyword">

<P> Policy Number </P> </RULE>

<RULE NAME="EndTag">

<P> Thank You </P> </RULE>

</GRAMMAR>

- User
 - Constrained Speech
- Example Queries
 - *What is the surrender value of Policy number*
 - *Can you please tell me the maturity value of Policy number*
 - ...
- Note that the constrained grammar gives a very accurate speech recognition because the speaker speaks what the speech engine expects

Type 2: Moderate User Experience and Speech Recognition

<GRAMMAR>

<RULE NAME="F_2" TOPLEVEL="ACTIVE">

<RULEREFF NAME="DonotCare"/>

<RULEREFF NAME="KeyConcept"/>

<RULEREFF NAME="DonotCare"/>

<RULEREFF NAME="KeyWord"/>

<RULEREFF NAME="DonotCare"/>

</RULE>

<RULE NAME="KeyConcept">

<P> Surrender Value </P>

<P> Maturity Value </P>

<P> ... </P>

<P> Address Change </P>

</RULE>

<RULE NAME="KeyWord">

<P> Policy Number </P>

<P> ... </P>

<P> ... </P>

</RULE>

</GRAMMAR>

- User
 - Can Speak anything but with in a constrained vocabulary
- Valid Queries
 - *<anything> Surrender value*
 - *<anything> policy number xyz*
 - *<anything> policy number xyz*
 - *<anything> maturity value*

Type 3: High User Experience Low Speech Recognition

<GRAMMAR>

<RULE NAME="F_1"
TOPLEVEL="ACTIVE">

<RULEREF NAME="DonotCare"/>

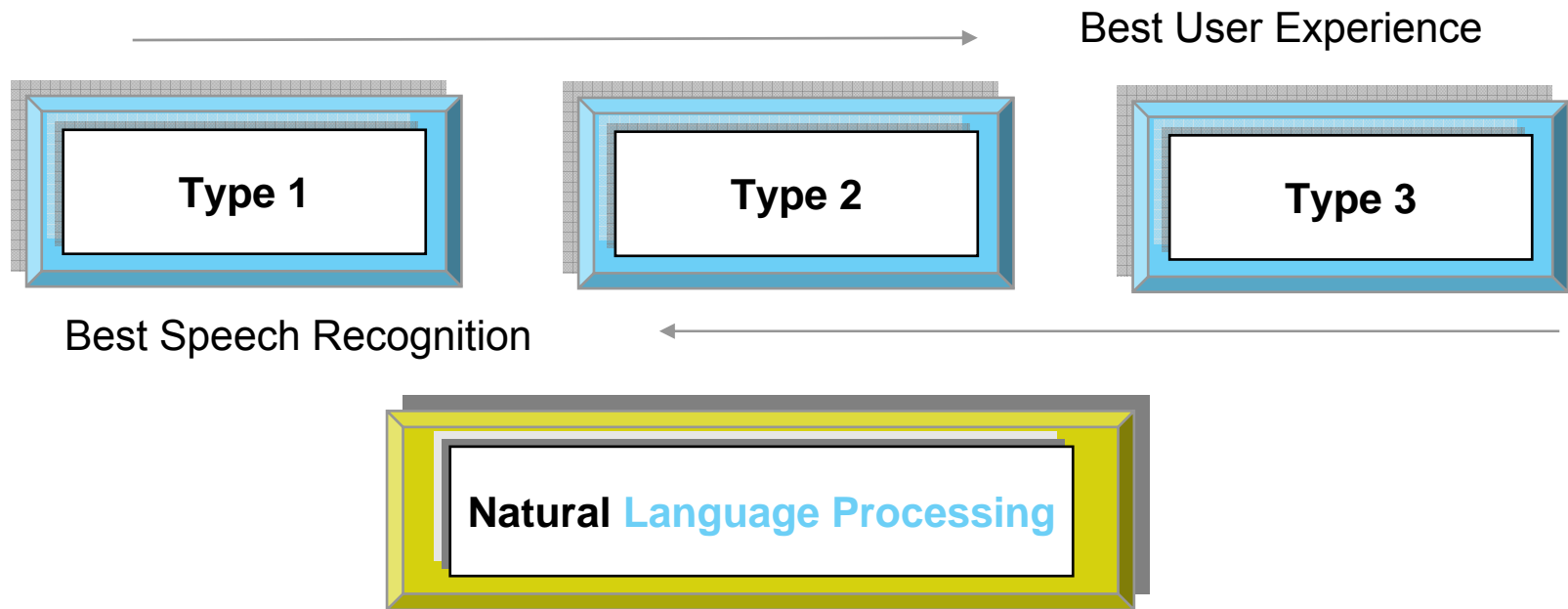
</RULE>

</GRAMMAR>

- User
 - Can Speak anything
- Valid queries
 - even out of domain query like *What does this system do?*
 - an incorrect query like *What is last paid commission address change?*
- Note that the liberal grammar gives a very poor speech recognition because the speaker can speak anything under the sun!



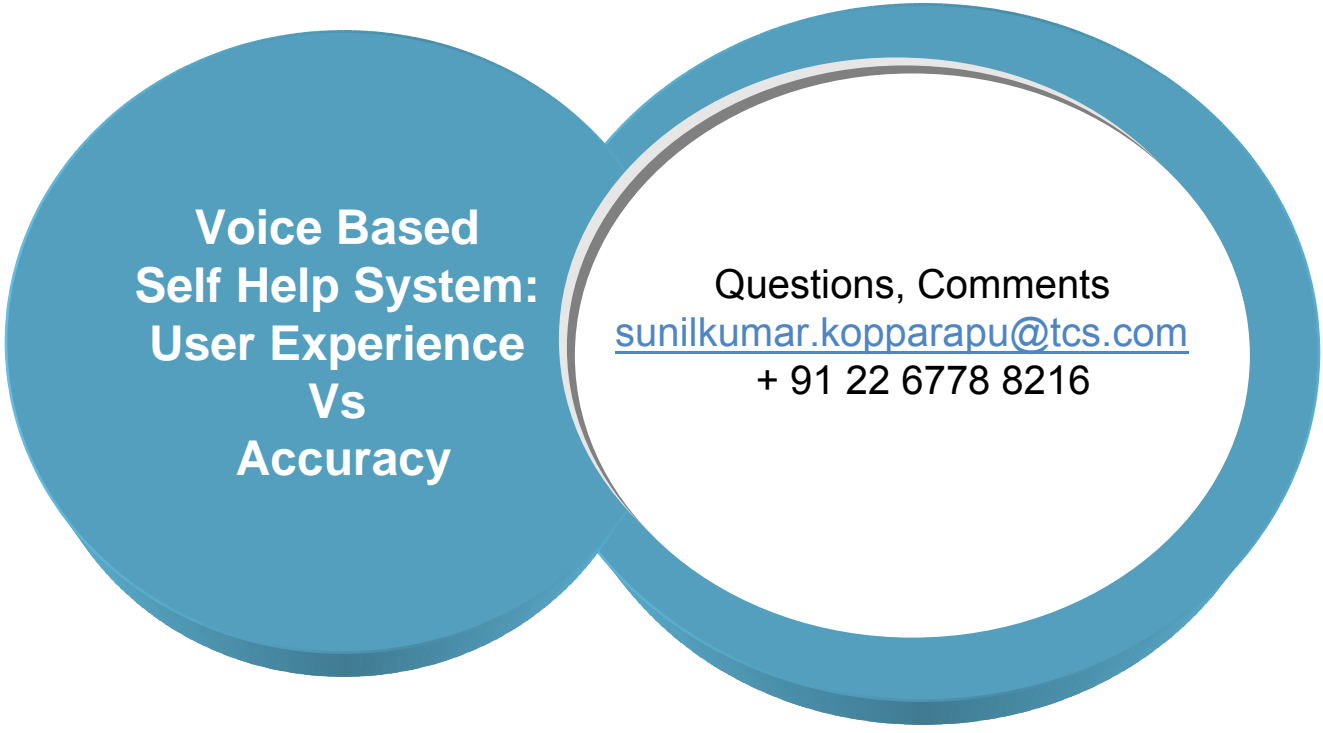
Evaluation



- 27 possible query types
 - 27 valid
 - 0 invalid
 - 3 not responded by Type 1 grammar + Language processing module
- 76 possible query types
 - 56 valid queries
 - 20 invalid but processed by Type 2 grammar + language processing module
- 357 possible query types
 - 212 valid (meaningful)
 - 145 invalid but processed by Type 3 grammar; handled by separate language processing module

Conclusions

- The performance of a voice based self help solution has two components
 - user experience and
 - the performance of the speech engine in converting the spoken speech into text
- We demonstrated that
 - speech and language module can be used jointly to come up with types of self help solutions which have varying effect on the user experience and performance of the speech engine.
 - by controlling the language grammar one could provide better user experience but the performance of the speech recognition became poor
 - on the other hand when the grammar was such that the performance of speech engine was good the user experience became poor.
- For designing usable voiced based self help systems
 - A balance between speech recognition accuracy and user experience is to be maintained so that
 - the speech recognition accuracy is good
 - without sacrificing user experience



**Voice Based
Self Help System:
User Experience
Vs
Accuracy**

Questions, Comments
sunilkumar.kopparapu@tcs.com
+ 91 22 6778 8216