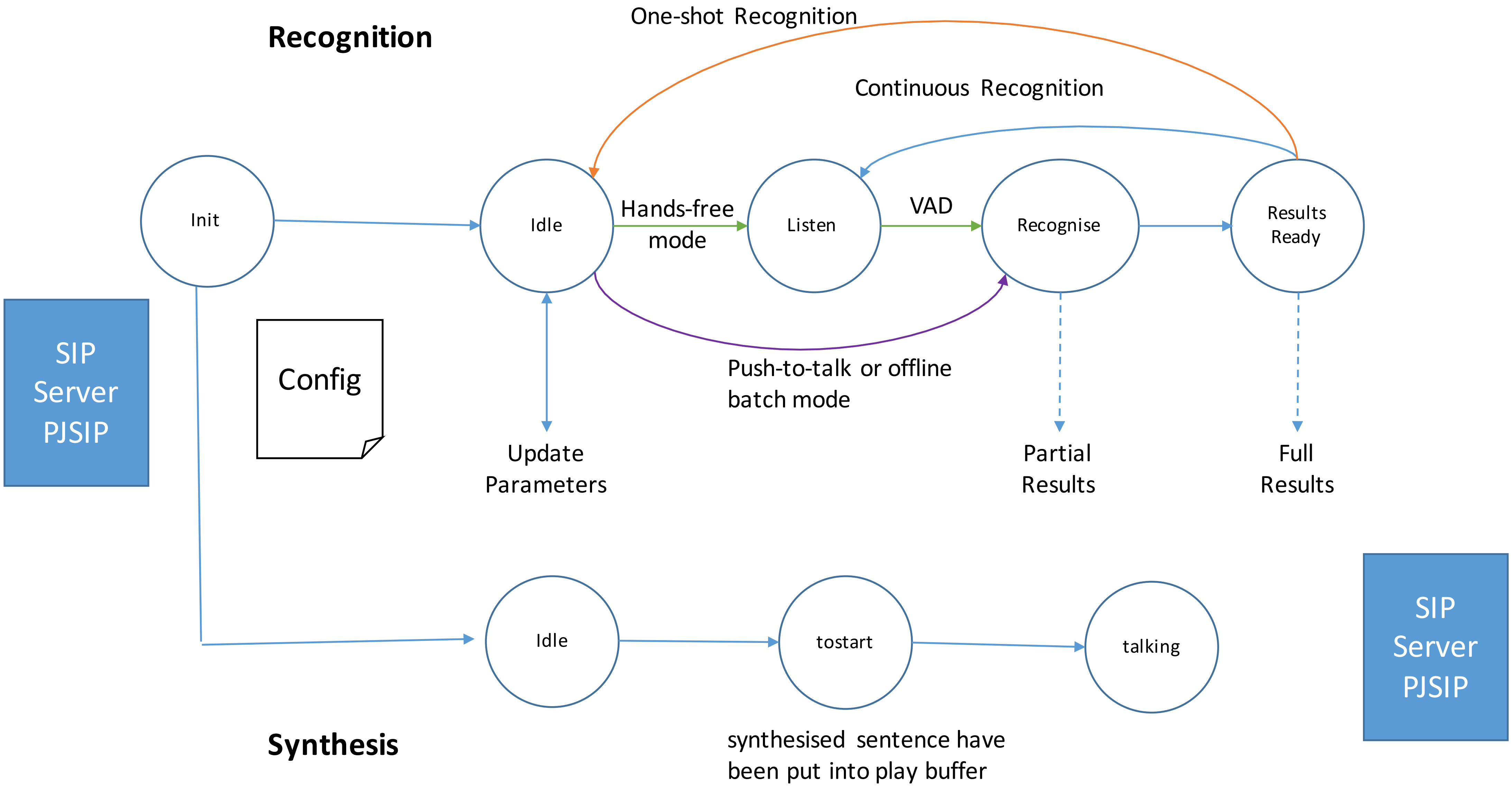
VoiceServer
Conceptual State Machine



Recognition

One-shot Recognition

Continuous Recognition

Init

Idle

Hands-free mode

Listen

VAD

Recognise

Results Ready

SIP Server PJSIP

Config

Push-to-talk or offline batch mode

Update Parameters

Partial Results

Full Results

Idle

tostart

talking

SIP Server PJSIP

Synthesis

synthesised sentence have been put into play buffer

High Level API

Init

InitVoiceServer()	Load all voice resources (Recognition + Synthesis) as specified in config file (Init->Idle)
SetParameter(x=y)	change parameter x to y (eg. Adjust beam width, VAD threshold, etc) (Idle)
GetParameter(x)	return current value of x

Recognition

StartRecognition(start,end)	if start=immediate (Idle -> Recognise), if start=VAD (Idle -> Listen)
	if end=one-shot (ResultsReady-> Idle), if end=continuous (ResultsReady-> Listen)

StopRecognition()	change end mode to one-shot (ie complete current recognition processing and stop)
AbortRecognition()	abandon current recognition and move immediately to idle state
GetRecognitionState()	Return current state of conceptual state machine

Synthesis

StartSynthesis(text=‘hello world’)

AbortSynthesis()

GetSynthesisState()

Protocols

-HTTP

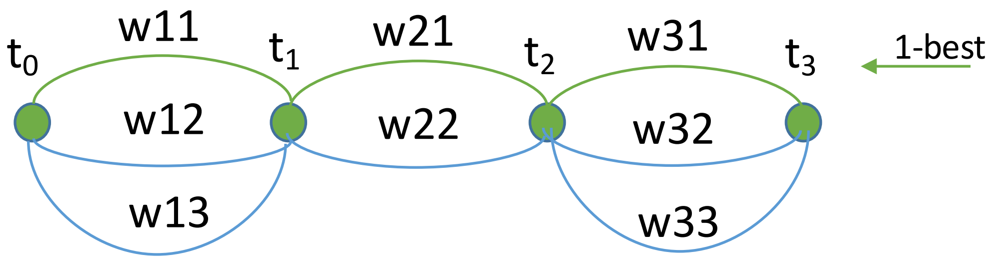
-TCP/IP

-Docker? Distributed applications

Speech Server
Recognition Results Format

All results are returned as python dict objects (or json format messages)
Key underlying data structure is a “confusion network”

```
{
   'resultType':'Final',   # alt is 'Partial')
   'nSpans' : 4,           # number of spans in network
   'spans' : [
      {
       'word' : 'the',     # most likely word in first span
       'prob' : 0.8,       # posterior prob of this word
       'id' : 834,         # unique vocab id
       'pron' : 2,         # pronunciation variant (optional)
       'alts' : [
         {'word' : 'this', # 1st alternative for this span
          'prob' : 0.15,   # posterior prob of this word
          'id' : 905},     # unique vocab id
         {'word' : 'three’,# 2nd alternative for this span
          'prob' : 0.05,   # posterior prob of this word
          'id' : 83}       # unique vocab id
          ]
      },
      {
       'word' : 'time',    # most likely word in 2nd span
       ....
      },
      ....
   ],
   'times' : [             # nSpans+1 boundary times
      47.612, 48.022, 49.764, 50.012, 51.021
      ]
}
```



Word boundary times
relate to 1-best

Recognition Result types:

-Simple transcription

-NBest Lists

-Confusion Network

Supporting VoIp: using pjsip as in Vocaliq.
We do not have to worry about sending speech data.

When is the Voice Server initialized?

- The Voice Server is initialized when it is started with a basic configuration file (Cornelia C++ style), is not initialized by the client.
- The Client can however change some parameters, such able/disable barge-in

OR

- Is The Voice Server initialized by the client at the when launching a client hub?

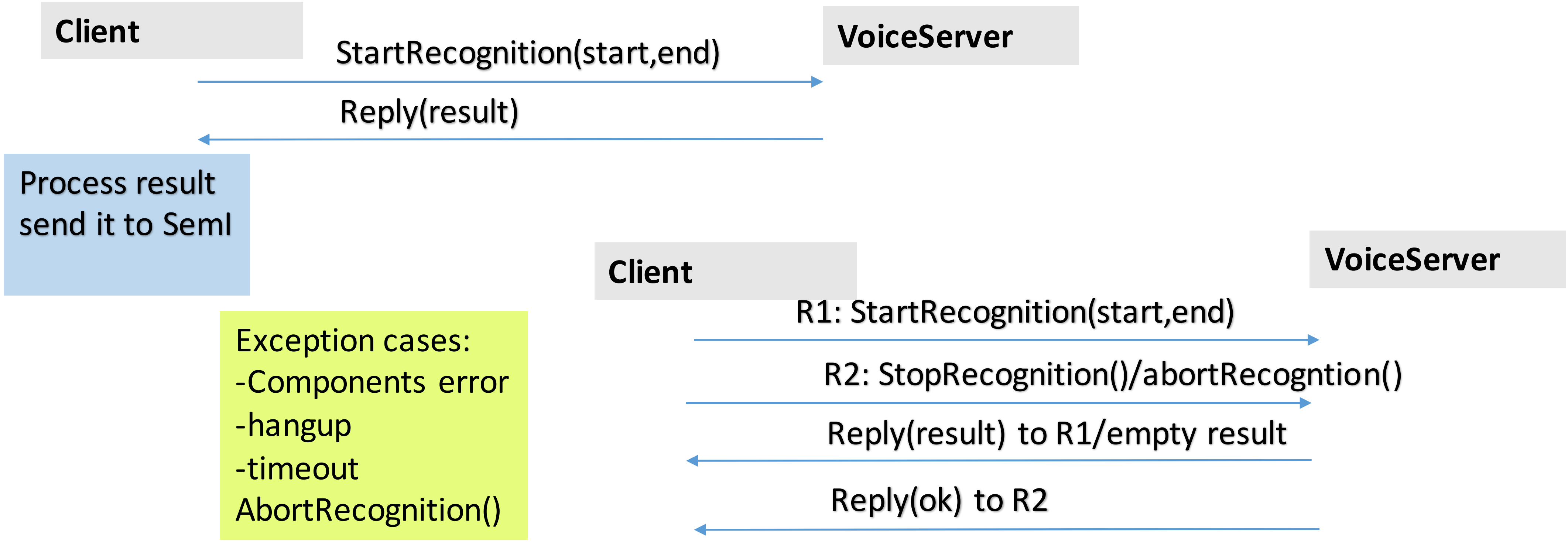
VoiceClient

Turn Manager?

Responsabilities:

- Timing (user and system)
- Hungup
- Mantaining User and System status status (use getRecognitionState()/getSynthesisState()
- Backchannel cues?

It is visible to all the dialogue components (SemI, Dialogue Manager, SemO)



Client

VoiceServer

StartRecognition(start,end)

Reply(result)

Process result
send it to SemI

Exception cases:
-Components error
-hangup
-timeout
AbortRecognition()

Client

VoiceServer

R1: StartRecognition(start,end)

R2: StopRecognition()/abortRecogntion()

Reply(result) to R1/empty result

Reply(ok) to R2

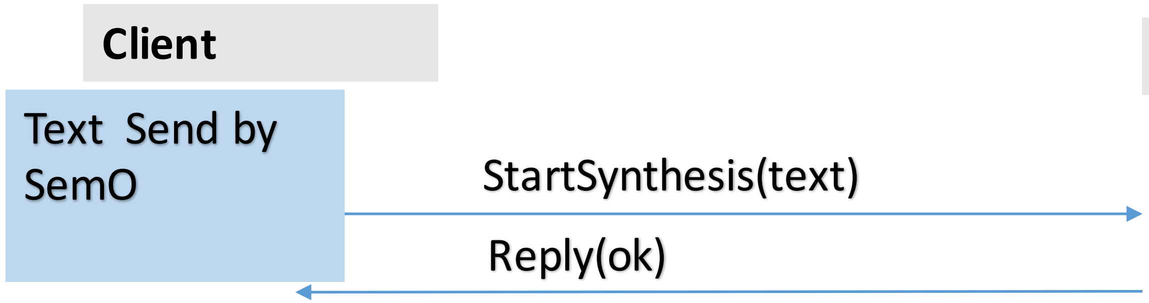
VoiceClient

Turn Manager?

Responsabilities:

- Timing (user and system)
- Hungup
- Mantaining User and System status (use getRecognitionState()/getSynthesisState()
- Backchannel cues?

It is visible to all the dialogue components (SemI, Dialogue Manager, SemO)



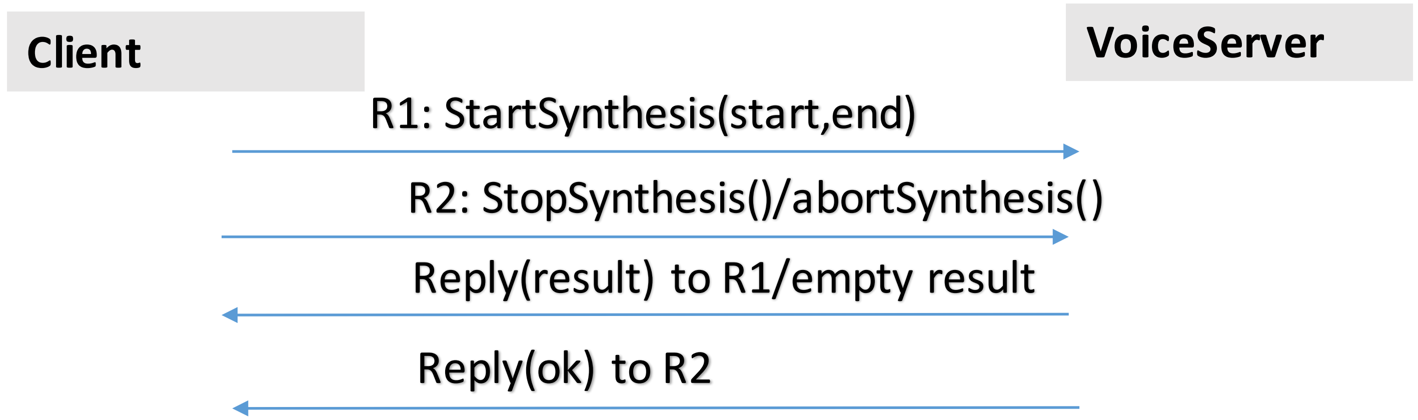
Client

VoiceServer

Text Send by SemO

StartSynthesis(text)

Reply(ok)



Client

VoiceServer

R1: StartSynthesis(start,end)

R2: StopSynthesis()/abortSynthesis()

Reply(result) to R1/empty result

Reply(ok) to R2

Exception cases:
-Barge-in
Components error
-hangup
-timeout
AbortSynthesis()