Voice Commands to Control Recording Sessions

John “Marty” Goddard

Thesis in partial fulfillment of the degree
Master of Science in Computer Engineering

Syracuse University

Fall 2011

Introduction: Motivation

Goal

- Provide hands-free recording operation
- Allow musicians to record themselves

Historical

- "Put That There" gestures and voice 1981
- Voice Navigator for Macintosh Musicians 1989
- Guitar Guru music teaching software, slogan "Keep your hands on your axe"

Recording Workflow for Overdubs:

(adding instruments and voices to existing recording)

<u>Record Take</u>: Begin Recording, perform music, stop recording

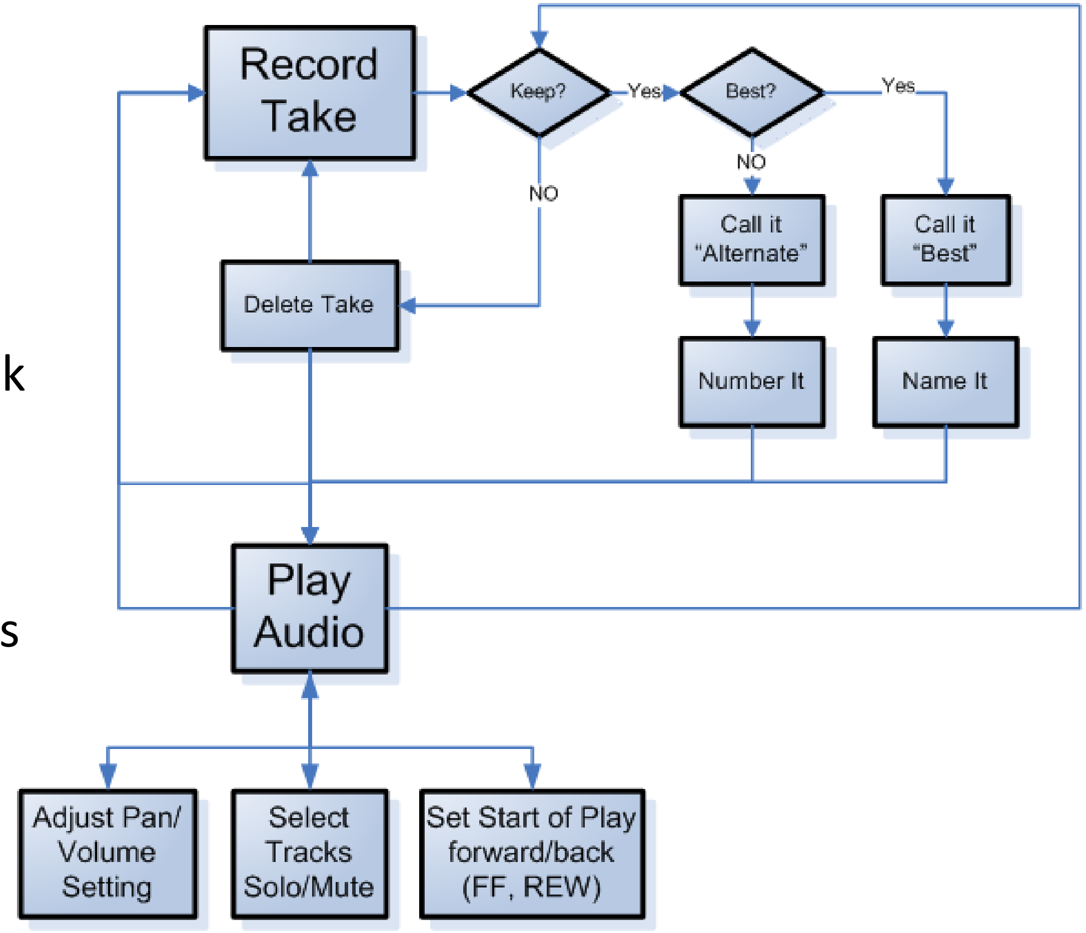
<u>Delete take</u> if not salvageable

<u>Play Audio</u>: Audition, or playback recorded tracks and listen for performance problems.

<u>Solo</u>: Selects track to hear, others are temporarily turned off

<u>Mute</u>: Temporarily turn off track

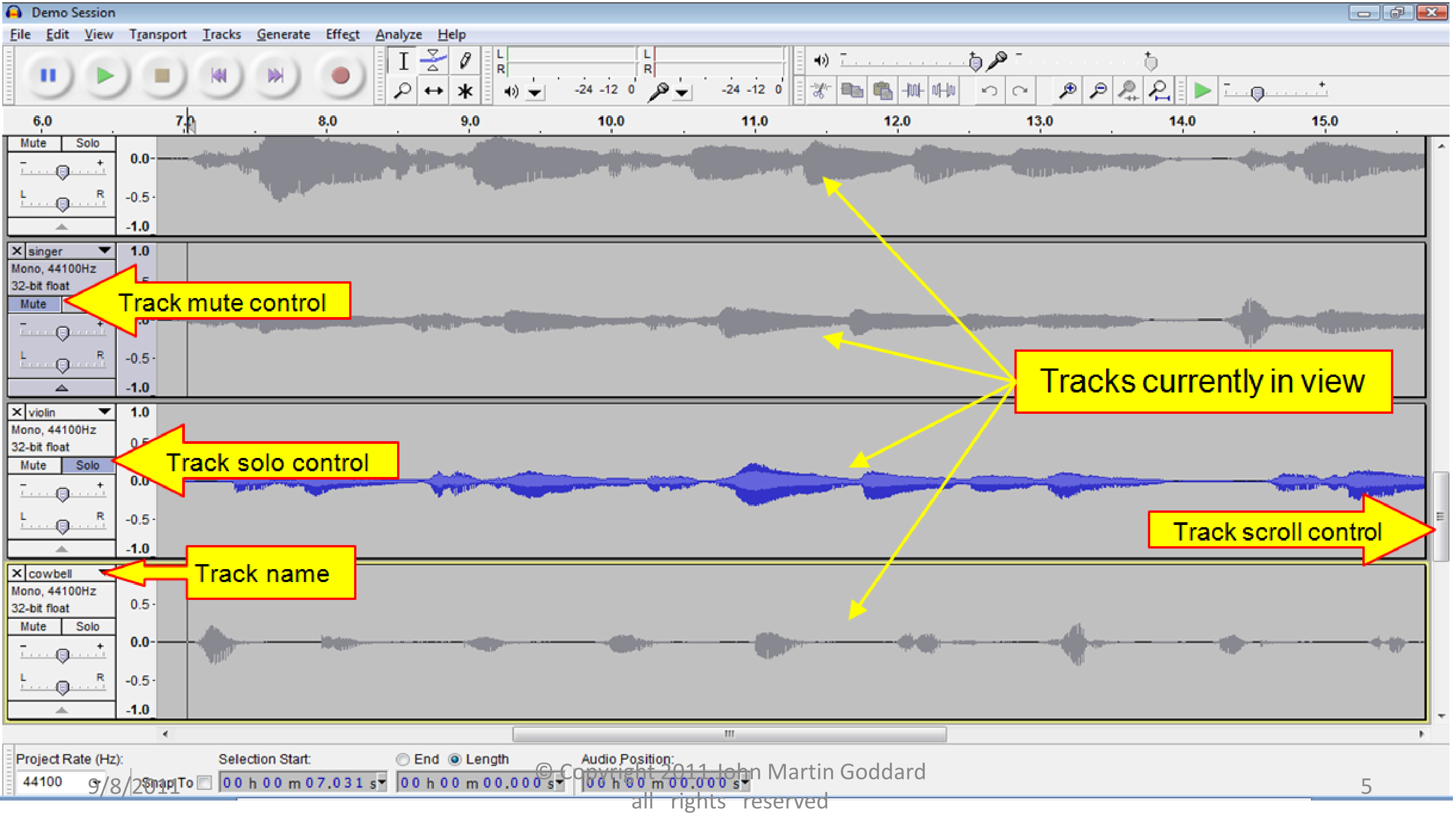
<u>Pan</u>: shift left or right in stereo mix



Experimental Apparatus

- Audacity (audio recorder/editor)
  - Open-source: No cost, can modify
  - Scripting command interface
    - Allows control from other applications/programs
- SayPlay (created by me for this research)
  - Accepts voice commands, sends to Audacity
  - Utilizes Windows Speech Recognition
    - Built into Windows (.NET 3.5 and later)
    - Supports custom grammars, Documented API

Audacity audio recorder/editor



Track mute control

Tracks currently in view

Track solo control

Track scroll control

Track name

Project Rate (Hz): 44100

Snap To □

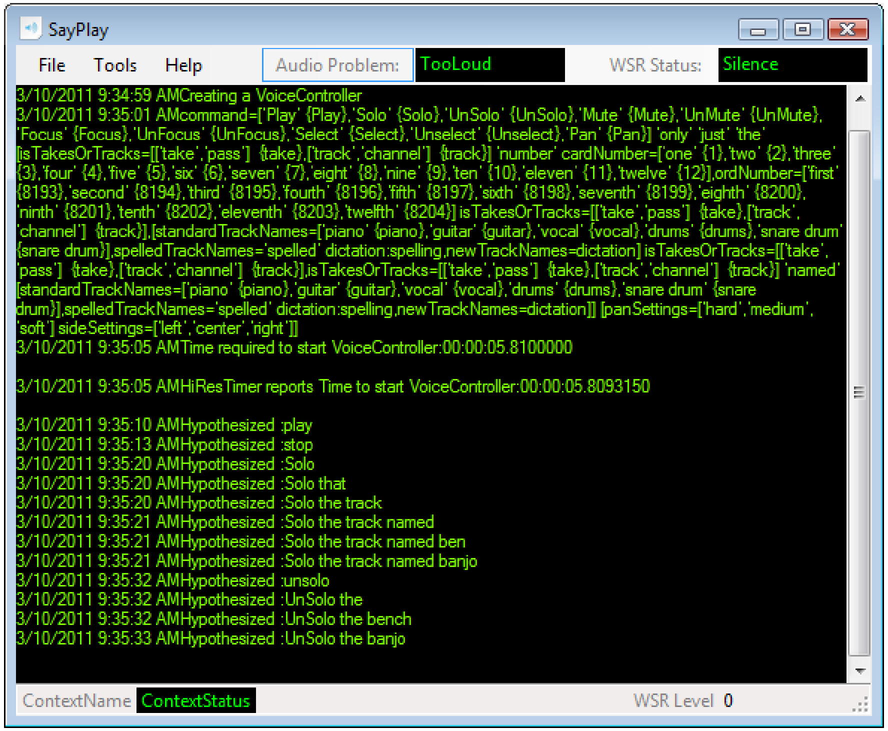
Selection Start: 00 h 00 m 07.031 s

○ End ● Length: 00 h 00 m 00.000 s

Audio Position: 00 h 00 m 00.000 s

SayPlay: handles voice command events, sends formatted commands to Audacity

- Accepts Voice input
- Interacts with Speech Recognition Engine
  - Creates Grammar structures
  - Handles recognition events
- Sends commands to Audacity via Script Interface



Audio Problem: TooLoud

WSR Status: Silence

```
3/10/2011 9:34:59 AMCreating a VoiceController
3/10/2011 9:35:01 AMcommand=['Play' {Play},'Solo' {Solo},'UnSolo' {UnSolo},'Mute' {Mute},'UnMute' {UnMute},
'Focus' {Focus},'UnFocus' {UnFocus},'Select' {Select},'Unselect' {Unselect},'Pan' {Pan}] 'only' 'just' 'the'
[isTakesOrTracks=[['take','pass'] {take},['track','channel'] {track}] 'number' cardNumber=['one' {1},'two' {2},'three'
{3},'four' {4},'five' {5},'six' {6},'seven' {7},'eight' {8},'nine' {9},'ten' {10},'eleven' {11},'twelve' {12}],ordNumber=['first'
{8193},'second' {8194},'third' {8195},'fourth' {8196},'fifth' {8197},'sixth' {8198},'seventh' {8199},'eighth' {8200},
'ninth' {8201},'tenth' {8202},'eleventh' {8203},'twelfth' {8204}] isTakesOrTracks=[['take','pass'] {take},['track',
'channel'] {track}],[standardTrackNames=['piano' {piano},'guitar' {guitar},'vocal' {vocal},'drums' {drums},'snare drum'
{snare drum}],spelledTrackNames='spelled' dictation:spelling,newTrackNames=dictation] isTakesOrTracks=[['take',
'pass'] {take},['track','channel'] {track}],isTakesOrTracks=[['take','pass'] {take},['track','channel'] {track}] 'named'
[standardTrackNames=['piano' {piano},'guitar' {guitar},'vocal' {vocal},'drums' {drums},'snare drum' {snare
drum}],spelledTrackNames='spelled' dictation:spelling,newTrackNames=dictation]] [panSettings=['hard','medium',
'soft'] sideSettings=['left','center','right']]
3/10/2011 9:35:05 AMTime required to start VoiceController:00:00:05.8100000

3/10/2011 9:35:05 AMHiResTimer reports Time to start VoiceController:00:00:05.8093150

3/10/2011 9:35:10 AMHypothesized :play
3/10/2011 9:35:13 AMHypothesized :stop
3/10/2011 9:35:20 AMHypothesized :Solo
3/10/2011 9:35:20 AMHypothesized :Solo that
3/10/2011 9:35:20 AMHypothesized :Solo the track
3/10/2011 9:35:21 AMHypothesized :Solo the track named
3/10/2011 9:35:21 AMHypothesized :Solo the track named ben
3/10/2011 9:35:21 AMHypothesized :Solo the track named banjo
3/10/2011 9:35:32 AMHypothesized :unsolo
3/10/2011 9:35:32 AMHypothesized :UnSolo the
3/10/2011 9:35:32 AMHypothesized :UnSolo the bench
3/10/2011 9:35:33 AMHypothesized :UnSolo the banjo
```

ContextName ContextStatus

WSR Level 0

- Logs command events and results to a text file

Voice Commands

(categorized by function)

- Single word or phrase:
  - Play, Stop, Record, Save, “Skip Start”, “Play One Second”, etc.
  - There are ~100 such commands, hard to remember them.

- Naming things: (tracks; in future: takes, sections)

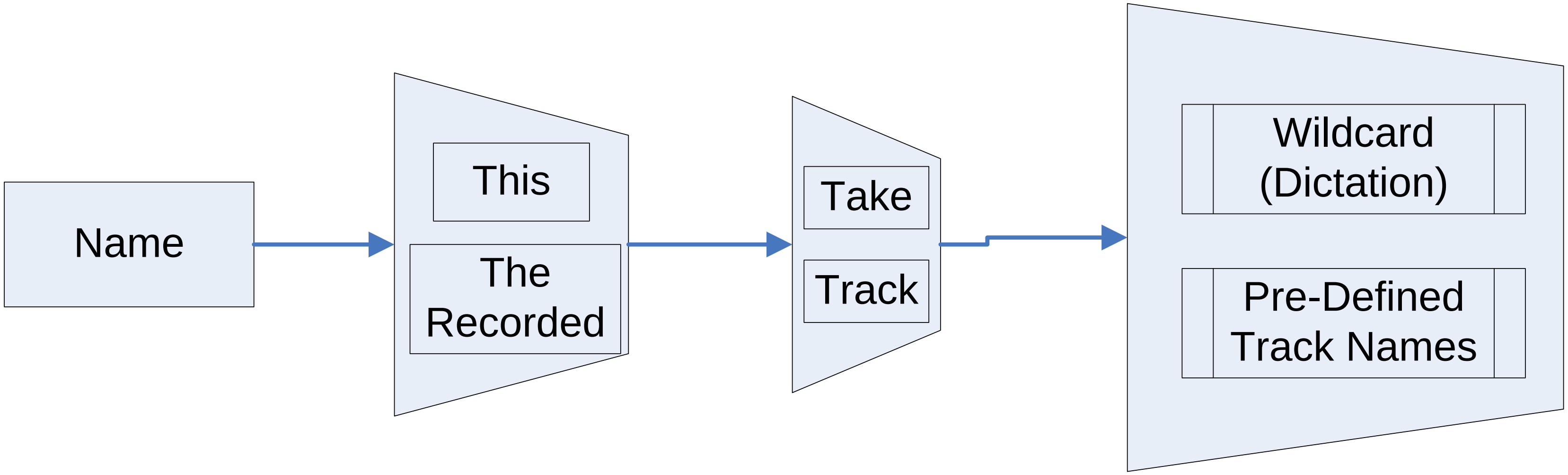
  Ex: “Name this track ‘kazoo’ “

- Adjusting things: Refer to tracks by *assigned name*, to control mute, solo, pan, gain, selection settings (etc.)

  Ex: “Play the kazoo track”, “Mute the ‘violin’ track”

- System Administration: Comments, context switching commands, refresh the list of names.

Grammar Structure for Naming Tracks

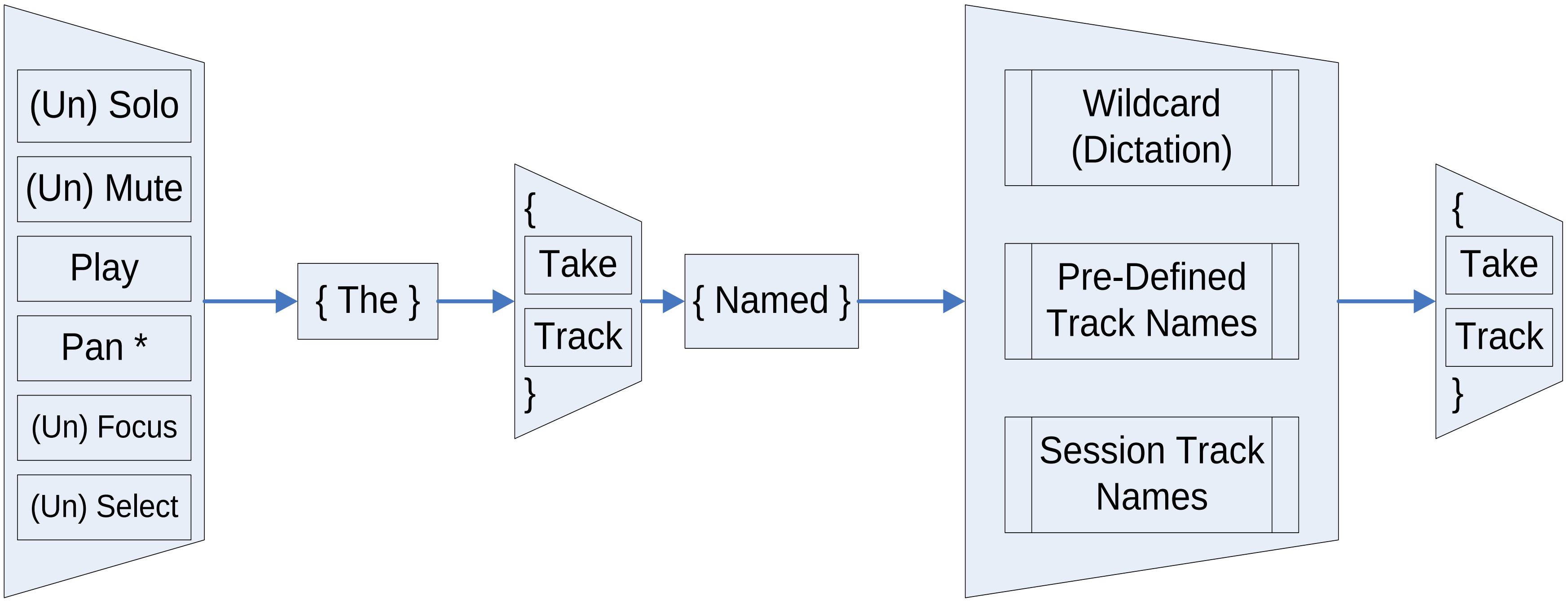


Creating the Name

- Trapezoid represents choices
- { } Indicates optional word(s)

Grammar Structure for Named Tracks

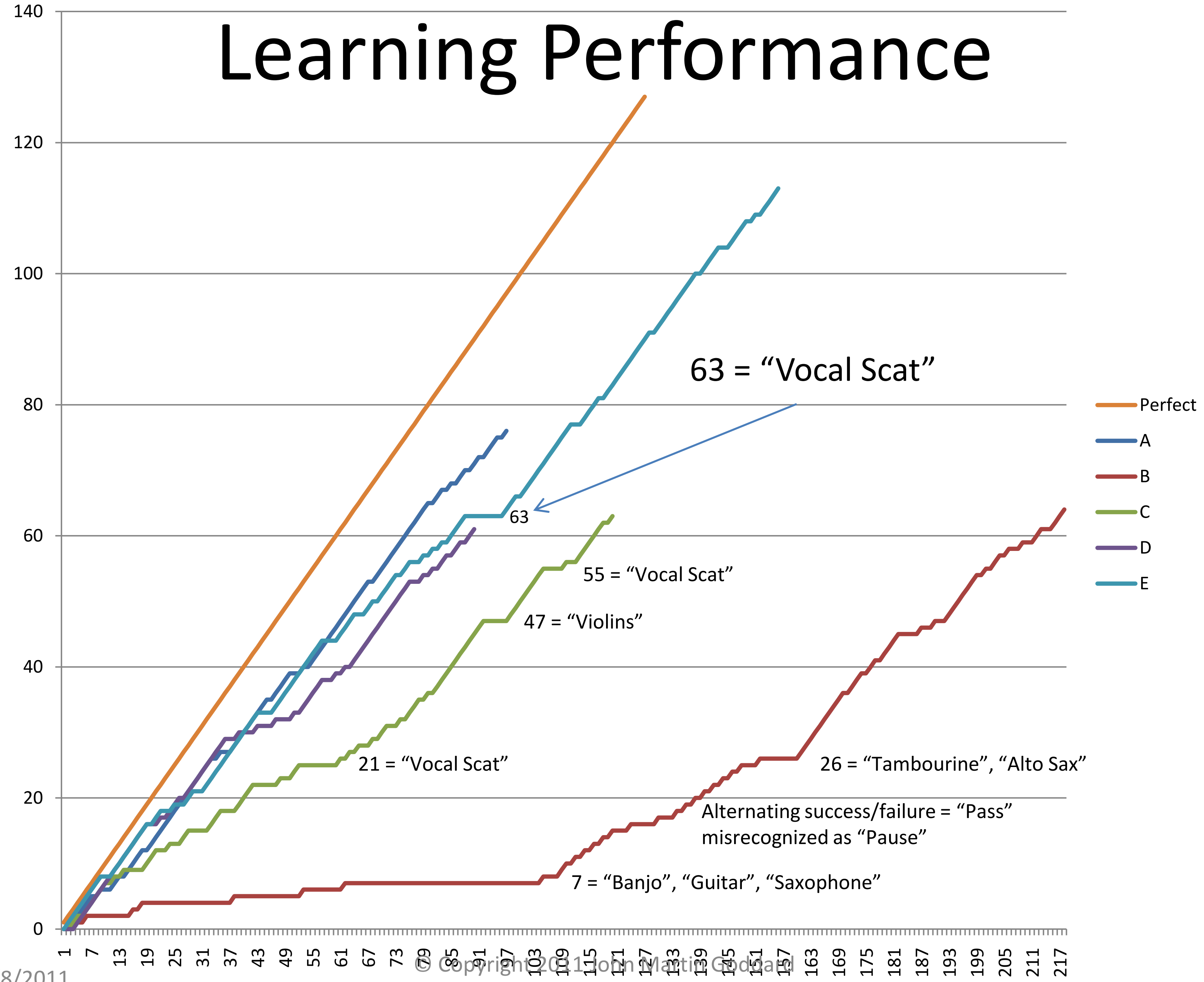
Using the Name



- Trapezoid represents choices
- { } Indicates optional word(s)

* Pan command appended with "Medium Left", "Hard Right" Etc.

Learning Performance



63 = “Vocal Scat”

55 = “Vocal Scat”

47 = “Violins”

21 = “Vocal Scat”

26 = “Tambourine”, “Alto Sax”

Alternating success/failure = “Pass”
misrecognized as “Pause”

7 = “Banjo”, “Guitar”, “Saxophone”

Types of Recognition Failures

- **Wrong Name:** track name is misrecognized
- **Wrong phrasing:** User error, incorrect wording
- **Low Confidence:** Correctly recognized but value of confidence (returned by WSR) is below threshold
- **Timeout:** Long pause in speaking truncates phrase
- **False positive:** Utterance misinterpreted as a command, with confidence above threshold
- **True rejection:** A non-command is misrecognized, but confidence is below threshold, so it is justly rejected
- **"Breath After":** WSR misrecognizes a dysfluency following a recognized word. Possible when an *optional additional* word is allowed. Ex: "Wrong *and*"

Problem With Dictation Speech Recognition: Names from "out of the blue"

- Random names (even made up ones), require large vocabulary dictation speech recognition

- Even after a name is correctly assigned, it can still be misrecognized when using the name

- Techniques were developed to improve accuracy assigning names and using names

3 Techniques were developed to Improve Assigning Names

- **1. Elaboration:** Use "Like" or "As In" to add words to a phrase about, or containing the desired name. Ex: "Name this track 'bass' as in bass guitar".
- **2. Quotation:** Enable a quoted phrase to be the name. Ex: "Name this track quote 'scream like a banshee' unquote".
- **3. Spell it:** "Name this track ***spelled*** W, O, W".

3 Techniques to Correctly Refer to a Named Item, Once the Name is Assigned

(in addition to methods for Assigning Names)

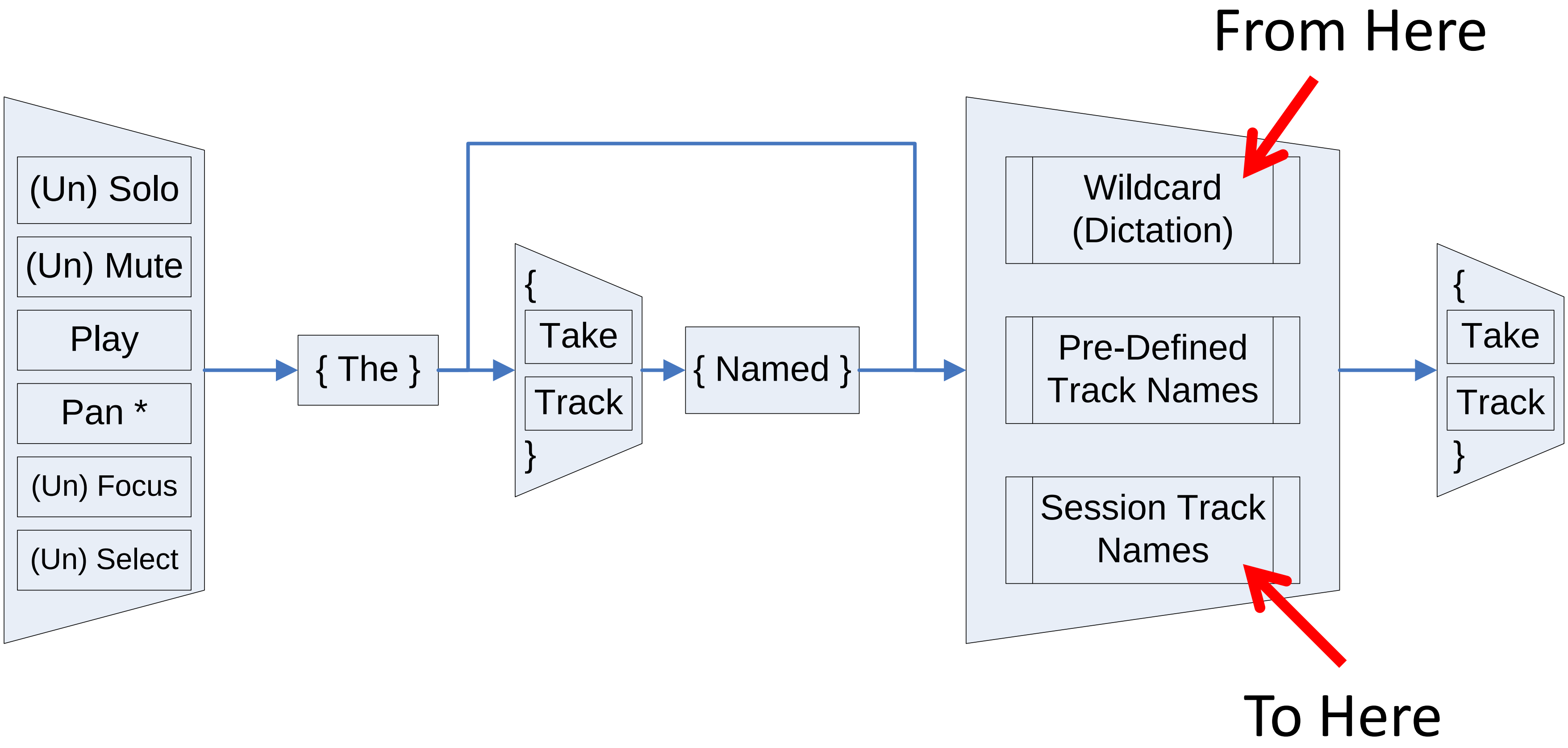
1. Load track names into the Grammar loaded into the Speech Recognition Engine (see next slide)

2. Add a name to the Windows Speech Dictionary

3. Prevent recognition of confused names

Loading Names into Grammar

Moves Recognition Duty…


From Here

(Un) Solo

(Un) Mute

Play

Pan *

(Un) Focus

(Un) Select

{ The }

{ Take Track }

{ Named }

Wildcard (Dictation)

Pre-Defined Track Names

Session Track Names

{ Take Track }

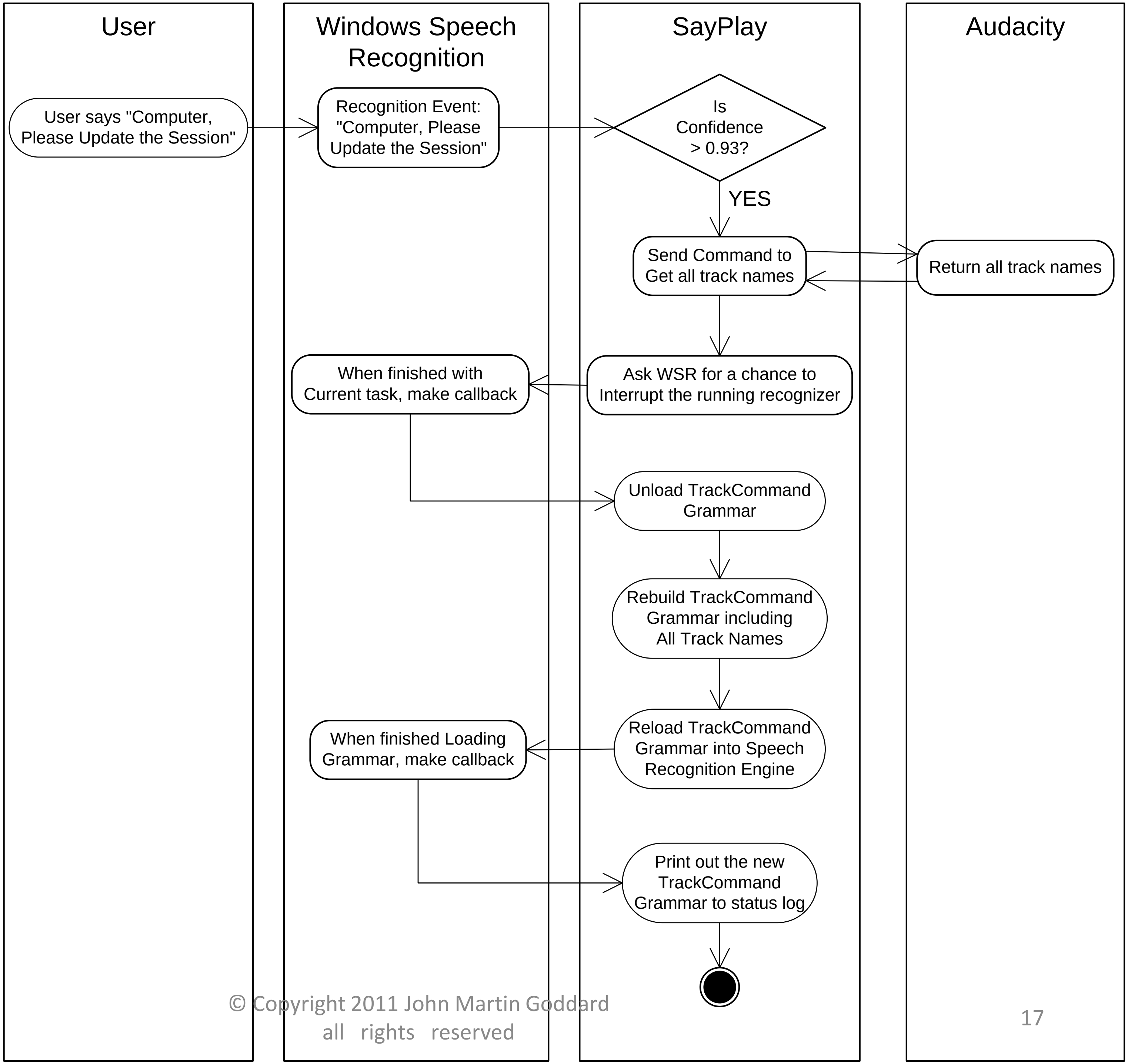
To Here

9/8/2011

Loading Names into Grammar

- Get list of track names from Audacity
- Unload TrackCommands Grammar from Speech Recognition engine
- Construct a new grammar with new track names
- Reload the new TrackCommands Grammar

Steps to Load Names into Gram-mar



User

Windows Speech Recognition

SayPlay

Audacity

User says "Computer, Please Update the Session"

Recognition Event: "Computer, Please Update the Session"

Is Confidence > 0.93?

YES

Send Command to Get all track names

Return all track names

Ask WSR for a chance to Interrupt the running recognizer

When finished with Current task, make callback

Unload TrackCommand Grammar

Rebuild TrackCommand Grammar including All Track Names

Reload TrackCommand Grammar into Speech Recognition Engine

When finished Loading Grammar, make callback

Print out the new TrackCommand Grammar to status log

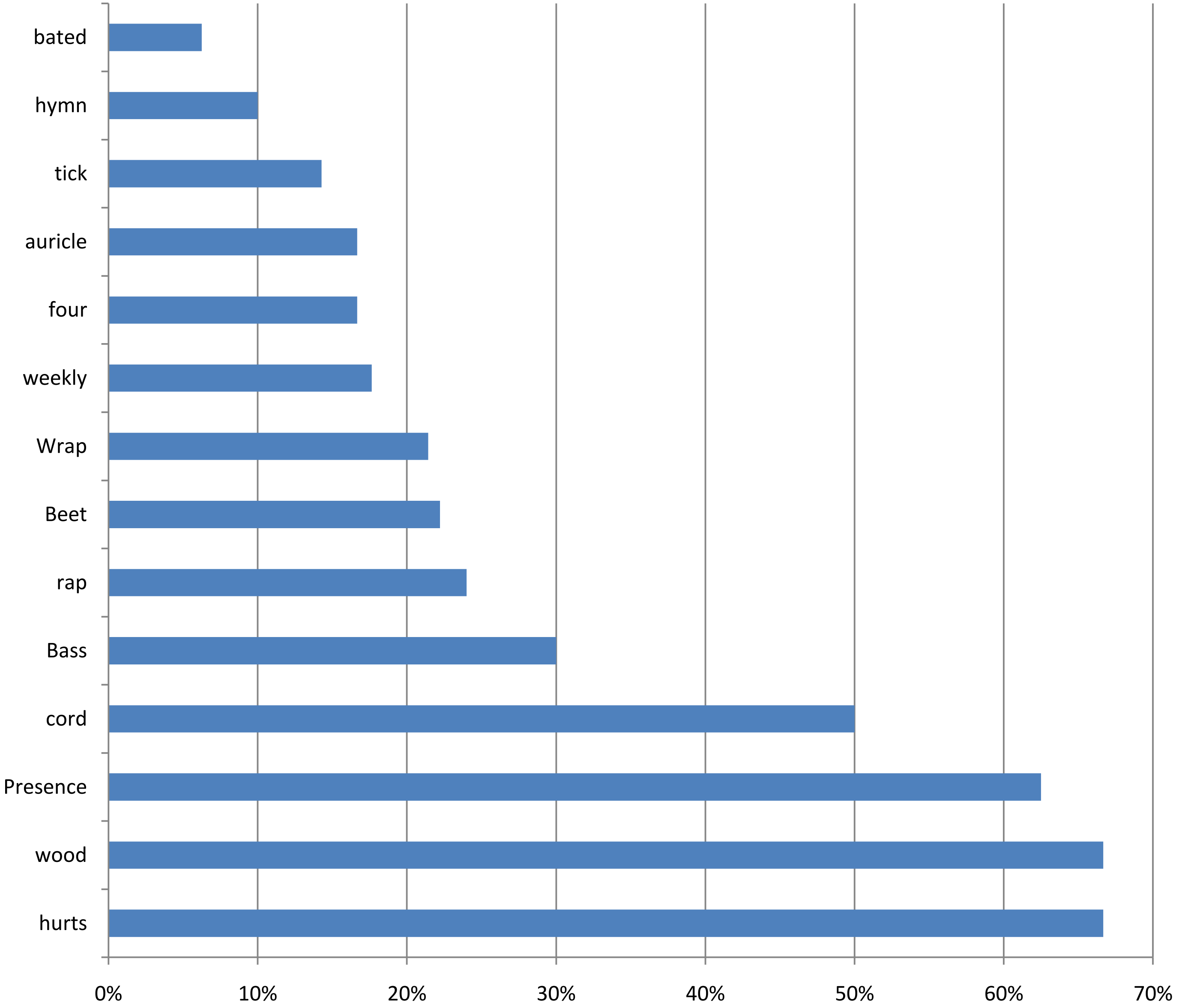
Experiments with Homophones

Pairs of words that sound the same …

… but are Spelled Differently

- Example: “Pairs” and “Pears”
- I selected from over 1000 homophone pairs only those having to do with sound or music
- Find which is usually recognized (the default)
- Then try elaboration to assign the non-default
- And measure the effectiveness of loading the non-default spelling into the loaded grammar

elaboration success rates

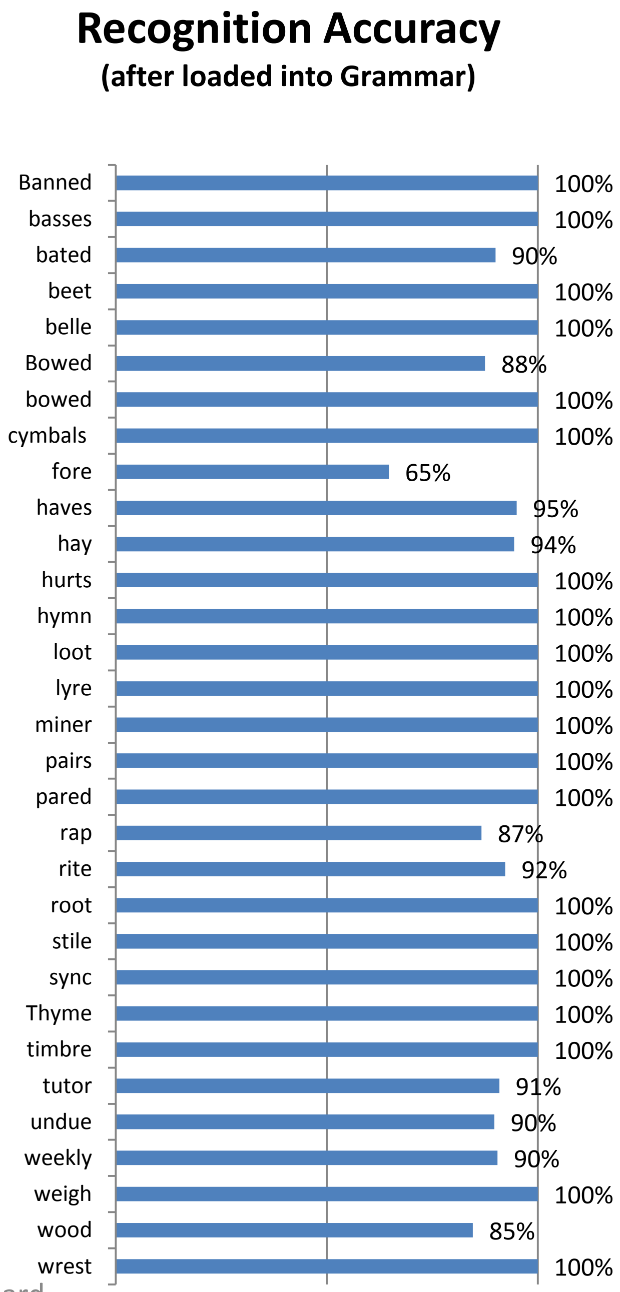


Interpretation of Elaboration Results

- Elaboration works only sometimes
- There seems to be a skew toward assigning Proper Names to named entities
  - Wei instead of way
  - Ryan or Orion, instead of rhyme
  - Kazue, instead of kazoo
  - Other confused names: Sarandon, Shanti, Peres,…
- Conclude that the command phrase “Name this track Theremin”, has greater influence over outcome than by adding “like the Russian Inventor”

After Loading into Grammar

- The non-default spelling of the Homophone pair was recognized with very high accuracy (even though it was rarely, if ever, recognized before loading the name into the grammar)



Interpretation

- Using dictation speech recognition for names already assigned, does perform reasonably well (>90%) on many ordinary track names
- Loading names into the loaded grammar works even better than relying on dictation recognition
- So, why not *always* load the names into the grammar? Because the names must first be acknowledged as correctly spelled.
- Why not always load session names into grammar *when first opening a session* (when all names are known good)? Future work.

Future Work on Voice Commands to Control Recording Sessions

- Name song sections for navigating in time
  - “Name this section ‘chorus’, or ‘verse’, or ‘hook’”
  - “Play the Chorus section” or “jump to the chorus section”
- Command Macros using “This Means That”
  - “Computer, please rename the ‘Play’ command ‘Audition’”
- Reference multiple objects in one command
  - “Mute the piano, bass and backing vocal tracks”

Other Possible Future Work

- Naming tasks, processes, or searches for voice commands to complete user-defined multi-step commands.

- Transpose this work to
  - Personal note taker/assistant
  - Video recording/playback
  - Other domains of process control

Questions?

- Thank you for attending