**Hands-free Text Editing using Voice and Gaze**

**Motivation**

Speech to Text through different voice recognizing platforms like (i) CMU Sphinx, (ii) Dragon’s Nuance, (iii) Microsoft’s Windows Speech Recognition Tool (iv) DeepSpeech2 (v) Google’s Speech to Text helps in effectively transcribing words spoken into a textual format for further analysis. However, the inherent nature of voice fails to provide a spatial context in case of error in transcription. We investigate a hands-free multimodal approach where gaze provides spatial context and voice helps in quantifying the need if an error is detected.

Speech to text-based applications are evaluated as unconstrained text entry evaluation paradigm. However, we intend to investigate our multimodal approach in understanding if different text entry requirements can be sufficed by the current voice recognizing platforms and if not, how can we improve our design and language model to cater to them.

Prior multimodal text entry process showed sequential usage of two different modalities. Our unified-interaction approach would reduce the number of steps for correction of incorrect transcribed words. Prior research has focused on controlled experimental scenarios. In order to understand the robustness of speech and gaze inefficient hands-free text entry process, evaluation can be done in different noise and light conditions. This would shed light on when one modality fails and the other modality can be used for efficient text entry.

**Points to Investigate**

1. Does multimodal selection approach improve the performance of hands-free text editing in comparison to unimodal approach?
2. Does inclusion of phonemes in text prediction improve the selection of text predictions?
3. Does hands-free text entry/edit improve with semantic information?

**Approach**

We develop an unconstrained text entry evaluation scenario where the participant inserts a text that is displayed via speech input. In case of an error, the participant can point at the incorrect word via gaze input and select for a plausible suggestion from the list. If it is not on the list, the participant can pronounce the word again or go in for a spell out mode where every letter is spelled out.

Another step towards this evaluation is to introduce an “improved” suggestion that is not only based on context but also on phonemes. A comparative analysis with the prior experiment might show if the improved language model helps in rectification of errors better.

**Tentative Keywords to define**

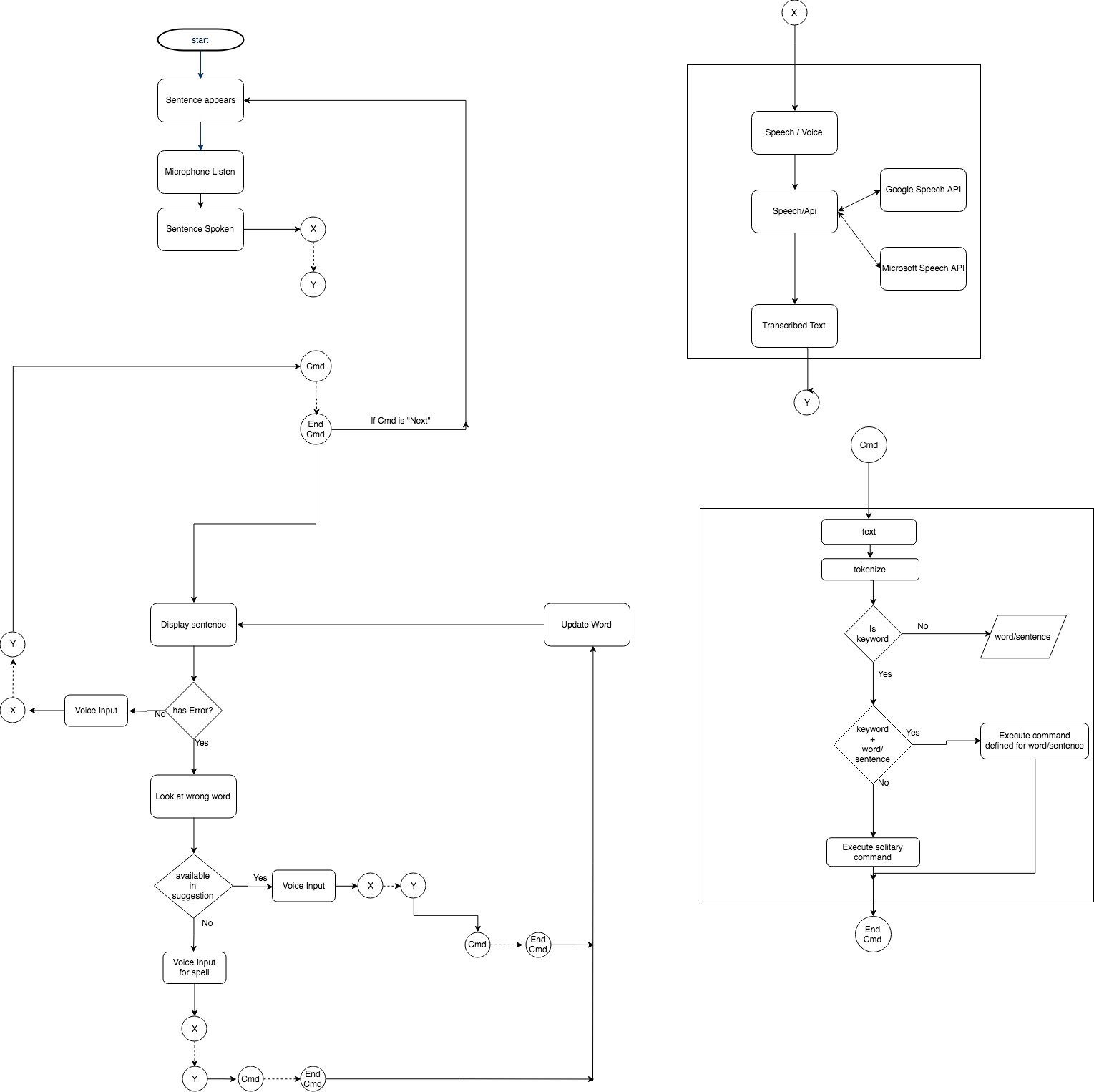
1. Select (select a suggestion from the list)
2. Delete (applicable for deleting single word)
3. Done (To confirm Editing of incorrect word to correct order)
4. Next (To give the indication that sentence is “correctly” inserted and Next sentence is anticipated)
5. Spell (in case of an issue where the participant needs to spell the word )

**Measurements**

1. Aggregate Measures:
   1. Entry Rates:
      1. Words per minute
      2. Sentence Completion Time
   2. Error Rates:
      1. Minimum String Distance
      2. Corrected, Uncorrected and Total Error Rates
   3. Efficiency Measures:
      1. Correction Efficiency
      2. Participant Conscientiousness
      3. Cost Per Correction
   4. Character Level Measures:
      1. Uncorrected Errors in Transcribed String
      2. Corrected Error in Input Stream
      3. Corrected-and-Wrong, Corrected-and-Right
   5. Initial Speech Transcription
   6. Speech Specific Measures:
      1. How responsive is the server
      2. How responsive is speech input
      3. Delay by user
      4. Delay by server
      5. % of time invested in correction or text input
      6. % of time invested in correction by speech.
2. Absolute/Relative Measures:
   1. Selection Savings

**Edge-case Scenarios in Web based text Entry**

1. Insertion of Email address.
2. Alpha-numeric combination of string
3. Insertion of Telephone numbers
4. Insertion of Dates
5. Spelling out letters to form words.

****

**Sentence Set**

1. MacKenzie, I. Scott, and R. William Soukoreff. "Phrase sets for evaluating text entry techniques." *CHI'03 extended abstracts on Human factors in computing systems*. ACM, 2003.
2. Garofolo, John S., et al. "TIMIT acoustic-phonetic continuous speech corpus." *Linguistic data consortium* 10.5 (1993): 0.

**Some Related Work**

1. Beelders, Tanya René, and Pieter J. Blignaut. "Using vision and voice to create a multimodal interface for Microsoft Word 2007." *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications*. ACM, 2010.
2. Beelders, Tanya René, and Pieter J. Blignaut. "Measuring the performance of gaze and speech for text input." *Proceedings of the Symposium on Eye Tracking Research and Applications*. ACM, 2012.
3. Hoste, Lode, Bruno Dumas, and Beat Signer. "SpeeG: a multimodal speech-and gesture-based text input solution." *Proceedings of the International working conference on advanced visual interfaces*. ACM, 2012.
4. Beelders, Tanya René, and Pieter J. Blignaut. "Using eye gaze and speech to simulate a pointing device." *Proceedings of the Symposium on Eye Tracking Research and Applications*. ACM, 2012.