COSSAT

CODE-SWITCHED SPEECH ANNOTATION TOOL

Sanket Shah, Pratik Joshi, Sebastin Santy, Sunayana Sitaram

Microsoft Research
Bangalore, India

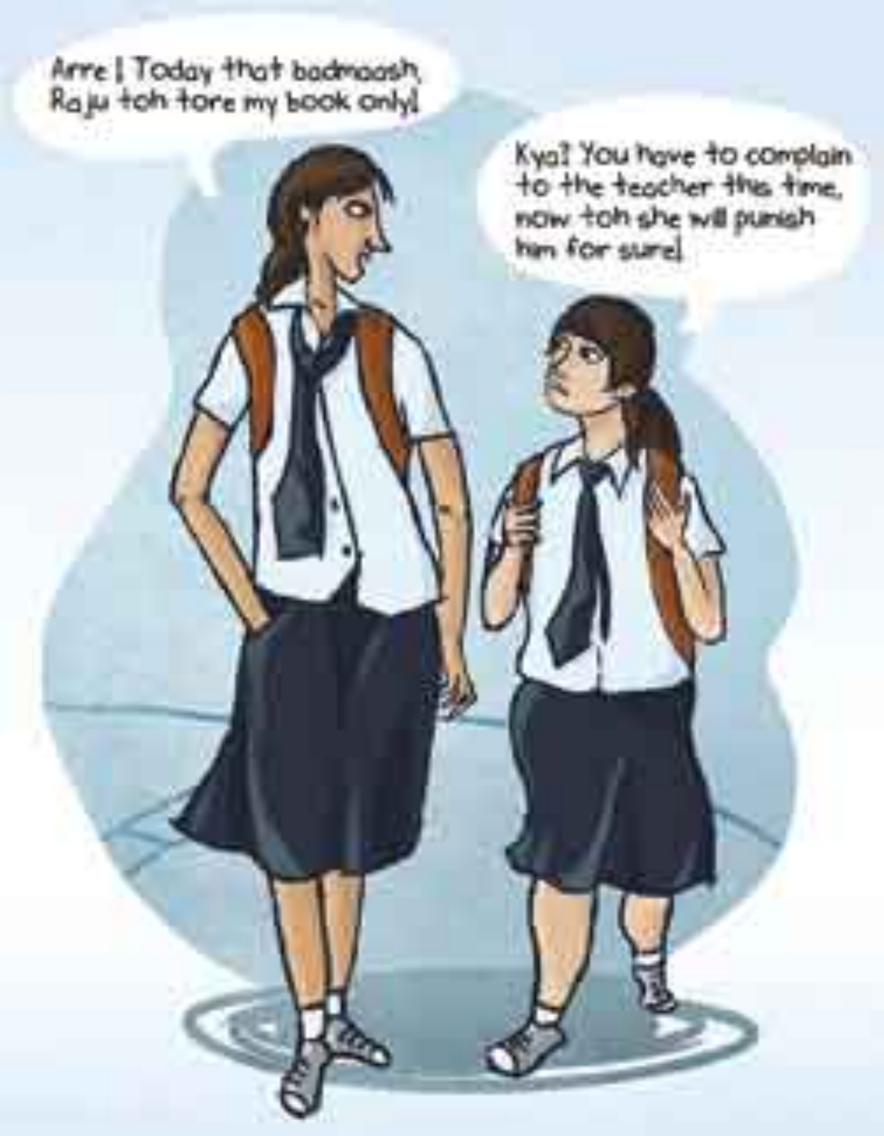
AnnoNLP '19

Microsoft

CODE-SWITCHING

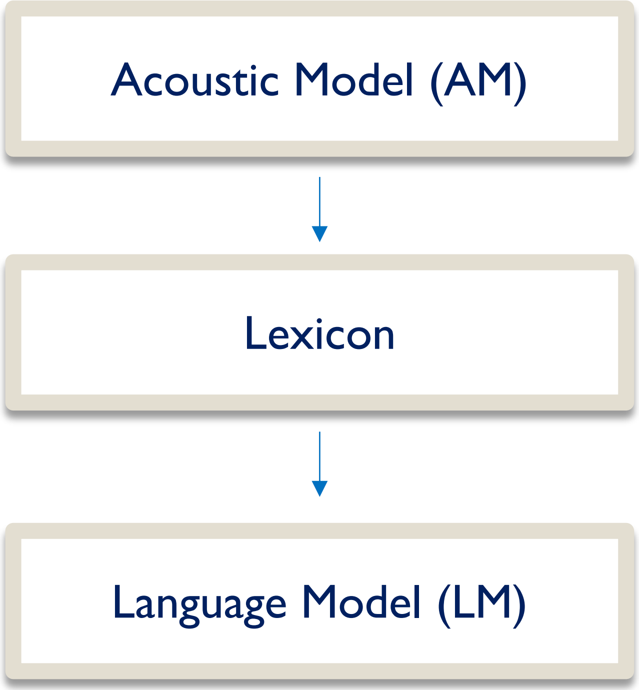
- Alternation of two languages by a bilingual speaker, usually in informal speech.

- Example:
  - Tume nahi pata, she is the daughter of the CEO, yaha do char din ke liye ayi hai. Maine socha, I should introduce myself to her.
  - Don’t you know, she is the daughter of the CEO, she’s here for a couple of days. I thought, I should introduce myself to her.


Arre! Today that badmaash, Raju toh tore my book only!
Kya? You have to complain to the teacher this time, now toh she will punish him for sure!

DATA REQUIREMENT FOR TRADITIONAL SPEECH RECOGNITION SYSTEMS?

- Not enough Annotated Data Available for training ASRs
  - Audio-transcript pairs. [Not Available]
  - Words and their phonetic sequence/g2p [Not Available]
  - Text data for training language model. [May be Available]
- Case is worse for Deep Neural Networks.



Acoustic Model (AM)

Need 1000s of hours of speech + corresponding transcripts

Lexicon

Need list of all words with corresponding phoneme* expansions/g2p

Language Model (LM)

Needs millions of lines of text data.

RELATED WORK

Sperber et al., 2016 uses hypothesis produced by ASR system to reduce human effort in transcribing speech.

Shan-Ruei You et al., 2004 combines scores from monolingual Chinese and English ASRs to determine the most probable output.

OUR WORK

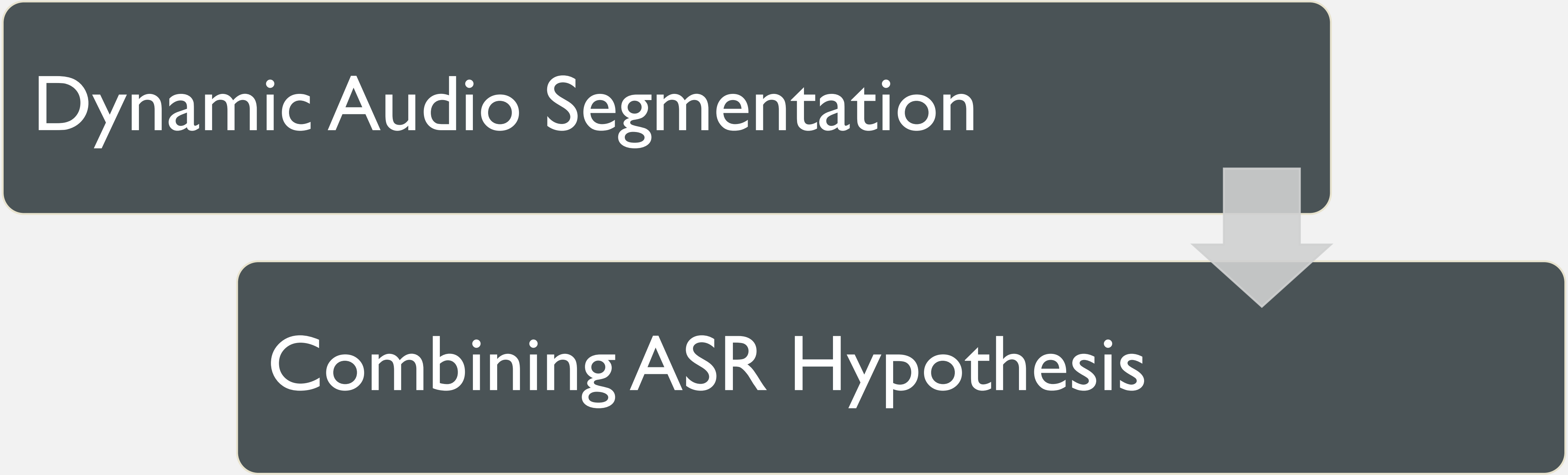
Inspired from Shah and Sitaram, 2019, where the authors use post-processing techniques on the recognition from monolingual ASRs to improve task of Spoken Term Detection in Code-Switched Speech.

We use monolingual ASRs to produce a set of candidate words being spoken in the code-switched speech.

The words are displayed to the annotator and serve as reference for the annotation task.

We provide a simple user interface to enable our method to be used for annotation.

METHODOLOGY

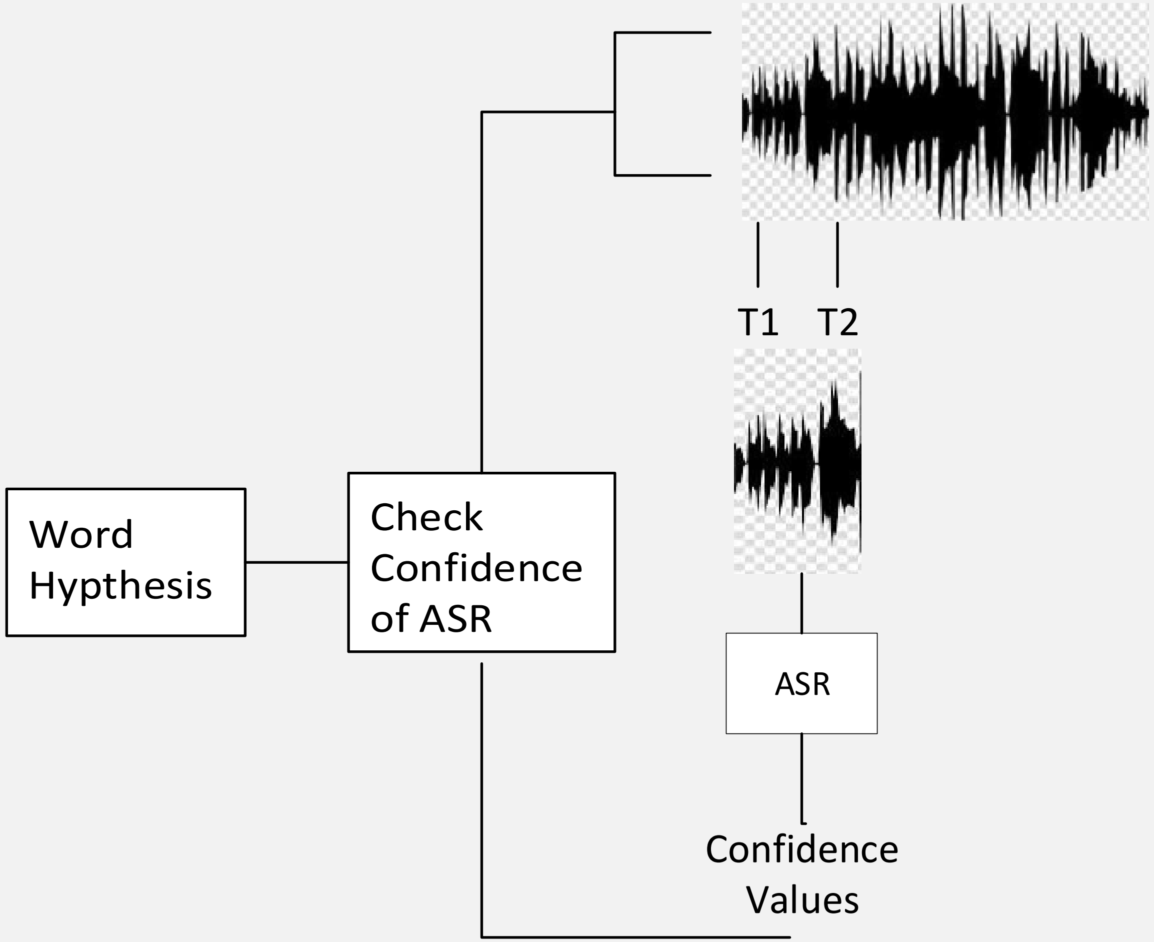


Dynamic Audio Segmentation

Combining ASR Hypothesis

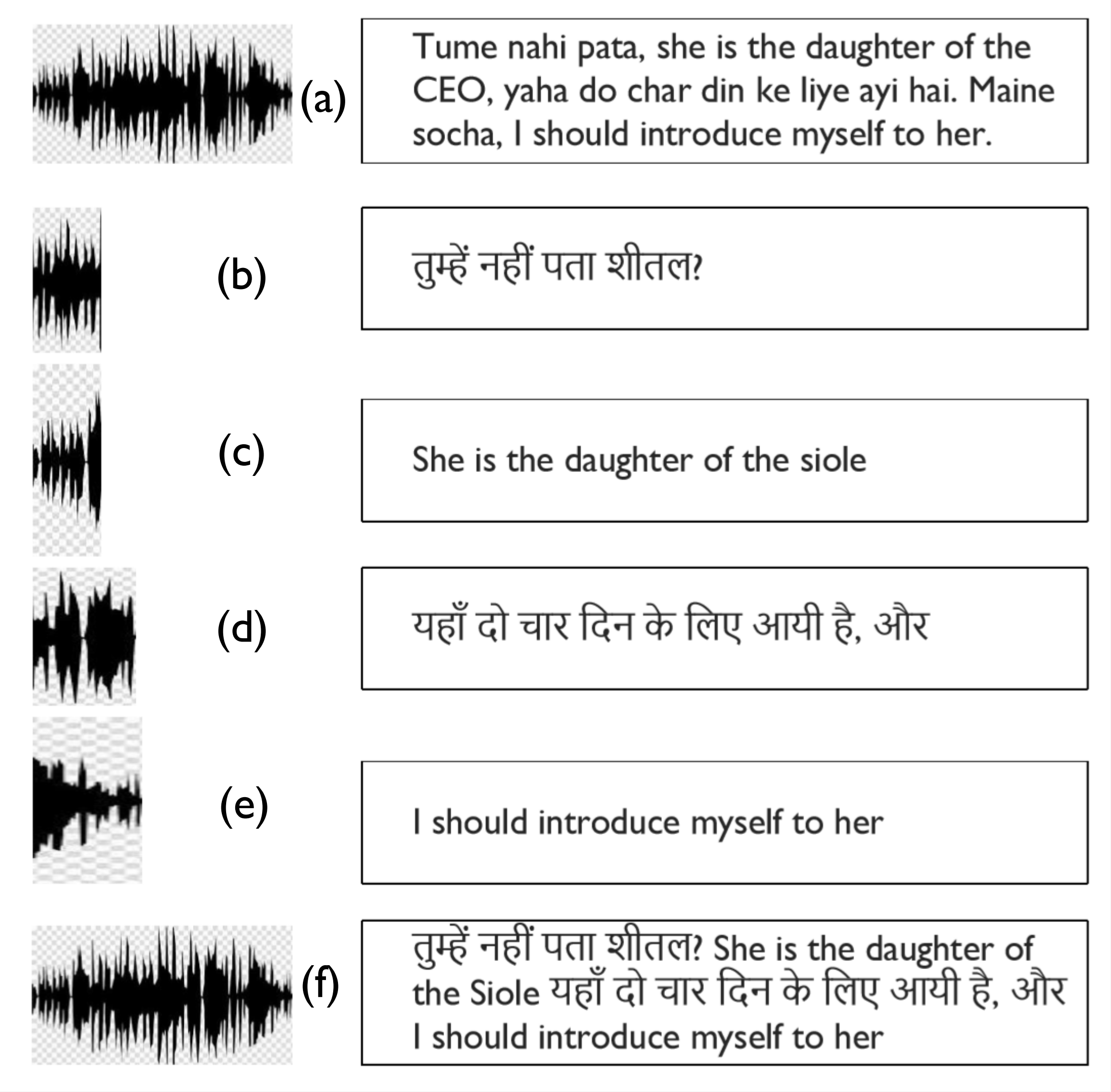
DYNAMIC AUDIO SEGMENTATION

- Monolingual ASRs have low accuracy on recognizing words in code-mixed speech.
- We dynamically chunk long audios into smaller segments based on ASR confidence.
- We initially start with chunk size of 0.5ms, but keep expanding the chunk size by 0.25ms based on ASR confidence.
- We empirically select a threshold of 0.3 for determining if the ASR system is confident about the selected chunk size.



COMBINING ASR HYPOTHESIS

- We hypothesize that each of the monolingual ASR will recognize a set of words in the given audio segment and the collection of the words of all the segments will comprise of all the words spoken in the given code-mixed speech.
- We try to evaluate our spoken term retrieval methodology using recall.
- We find that our system has a recall of 0.84.
- Thus, we were able to retrive 84% of the spoken terms using the method of chunking and combination.



(a) Tume nahi pata, she is the daughter of the CEO, yaha do char din ke liye ayi hai. Maine socha, I should introduce myself to her.

(b) तुम्हें नहीं पता शीतल?

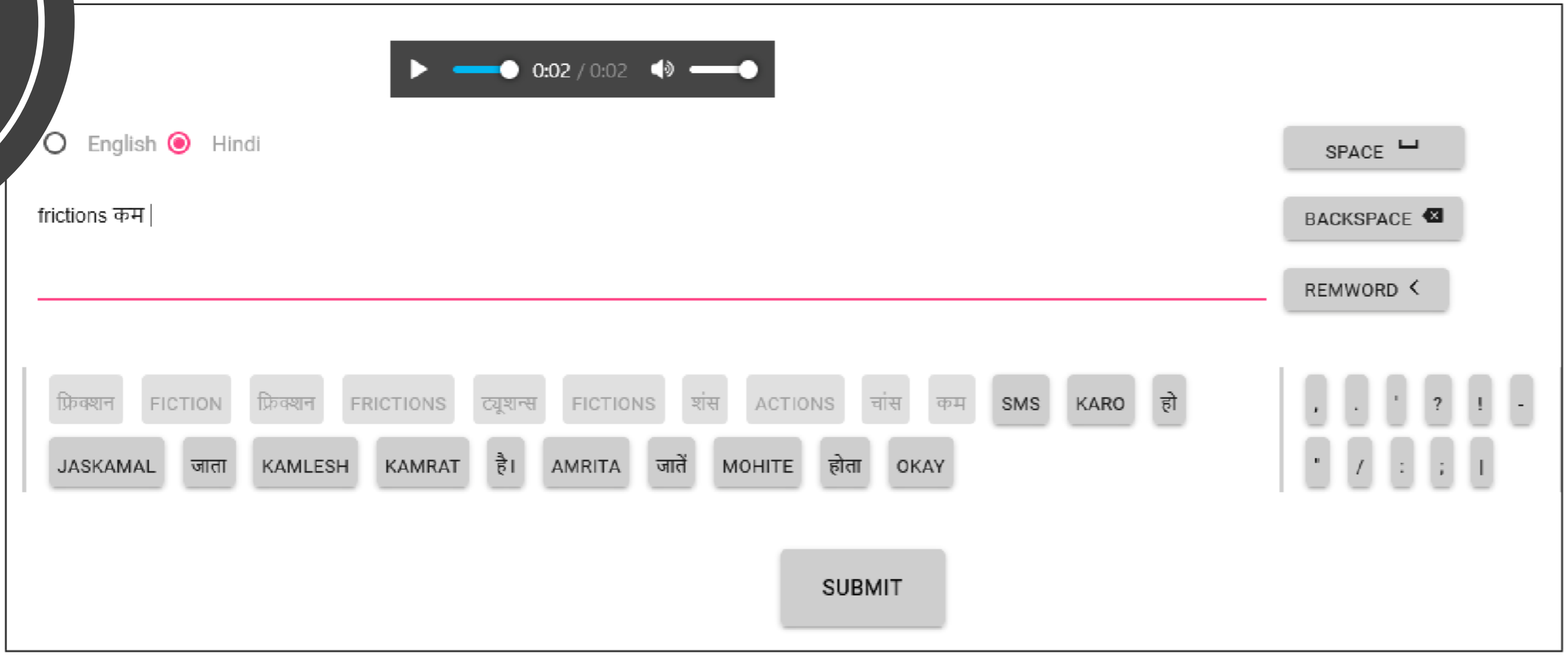
(c) She is the daughter of the siole

(d) यहाँ दो चार दिन के लिए आयी है, और

(e) I should introduce myself to her

(f) तुम्हें नहीं पता शीतल? She is the daughter of the Siole यहाँ दो चार दिन के लिए आयी है, और I should introduce myself to her

INTERFACE OVERVIEW



○ English ● Hindi

frictions कम

Figure 1: CoSSAT (Code-Switched Speech Annotation Tool)

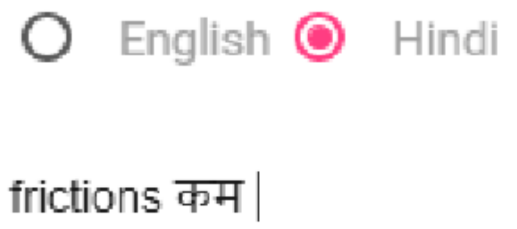
INTERFACE FEATURES

(a) The user can easily switch between Devanagari and Roman script using keyboard shortcuts. This enables generating transcriptions in the correct scripts.

(b) User can play, pause, forward and rewind the audio being played during transcription.

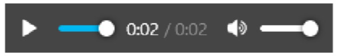
(c) As the user clicks on the corresponding words being spoken in the audio, we make the words prior to the current selection, out of focus. This enables the user to focus on the words being spoken at that moment.

(d) User can click on SPACE, BACKSPACE, REMWORD buttons to type space, delete one character or delete one word respectively.

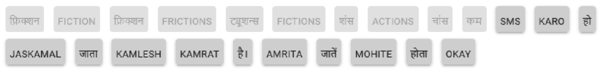

○ English ● Hindi

frictions कम

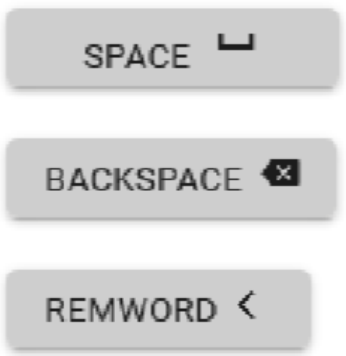
(a)



(b)



(c)



(d)

Baseline: No ASR hypothesis is shown and the annotators are expected to type out the entire transcription.

CoSSAT: Annotators are shown probable word hypothesis in the form of clickable buttons which can be used to transcribe the audios.

METHODOLOGY

EXPERIMENTAL SETUP

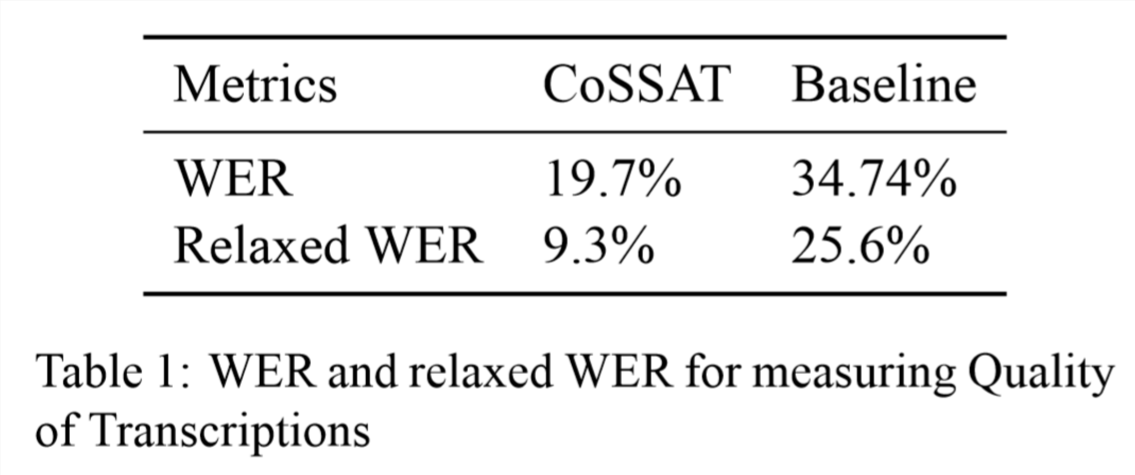
- Total no. of users is equal to 10.
  - HSetA contains 5 users.
  - HSetB contains 5 users.
- Total no. of audios to be transcribed is equal to 14.
  - SetA contains 7 audios.
  - SetB contains 7 audios.
- Total no. of tasks is equal to 2.
  - TaskA: HSetA is asked to transcribe audios from SetA using baseline method and audios from SetB using CoSSAT.
  - TastB: HSetB is asked to transcribe audios from SetA using CoSSAT method and audios from SetB using baseline method.

EVALUATION

- Quantitative Evaluation
  - Transcription Quality
  - Annotation Speed
  - Annotation Effort
- Qualitative Evaluation

TRANSCRIPTION QUALITY

- Transcription quality was determined by computing word error rate (WER) using a standard procedure, using the transcriptions present in our in-house dataset as the gold standard.
- We calculated WER for the transcriptions created by users using our system as well as for the transcriptions created using the baseline approach.
- We also calculate relaxed WER* for taking care of minor spelling variants, cross-transcriptions, minor variations such as long/short vowels and nasalization.



Metrics	CoSSAT	Baseline
WER	19.7%	34.74%
Relaxed WER	9.3%	25.6%

Table 1: WER and relaxed WER for measuring Quality of Transcriptions

ANNOTATION SPEED

- CoSSAT: We record time taken by the user to transcribe from the moment user clicks on the first word or clicks on the textbox.
- Baseline: We record time from the moment user clicks on the textbox to start typing.

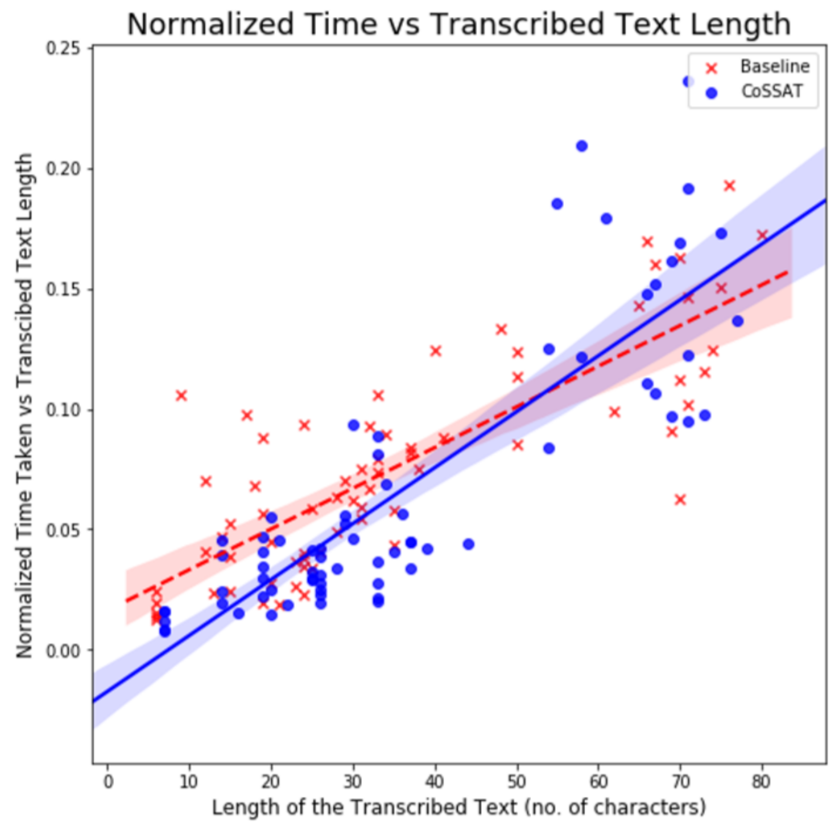


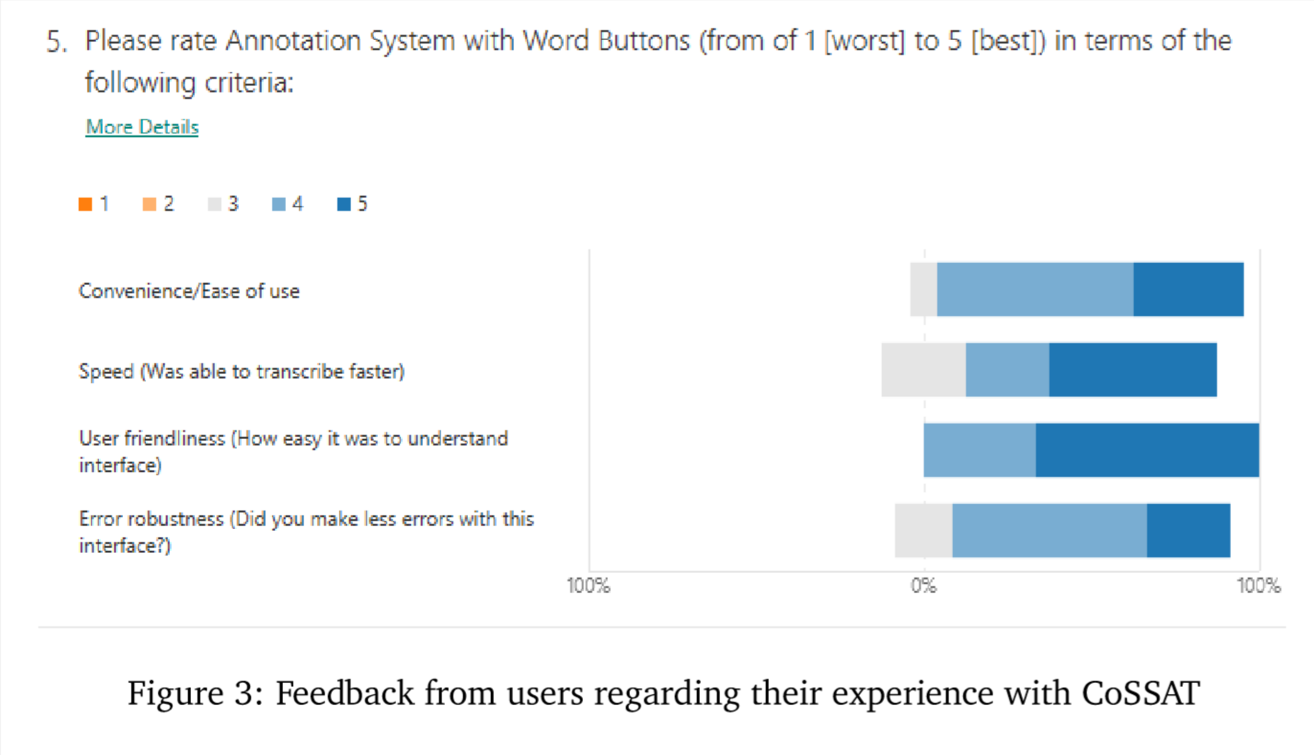
Figure 2: Annotation Speed Plot for each audio. Y axis is the Normalized Time Taken for the utterance. X axis is number of characters present in the utterance. Red colour (cross) is the Baseline system. Blue (dots) colour is CoSSAT.

ANNOTATION EFFORT

- We measure no. of keystrokes and mouse clicks to measure annotation effort.
- CoSSAT system resulted in 8 keystrokes and 8 mouse clicks, while baseline system had 57.1 keystrokes and 5.4 mouse clicks.
- Overall the numbers show that much less effort was required for annotation task while using the CoSSAT system.

QUALITATIVE EVALUATION

- Question: Please rate Annotation System with Word Buttons (from of 1 [worst] to 5 [best] in terms of following criteria)
- 1. Convenience / Ease of use
- 2. Speed (Was able to transcribe faster)
- 3. User friendliness
- 4. Error robustness



5. Please rate Annotation System with Word Buttons (from of 1 [worst] to 5 [best]) in terms of the following criteria:

Figure 3: Feedback from users regarding their experience with CoSSAT