Spoken English Grading: Machine Learning with Crowd Intelligence

Date : 2015/12/24
Author :Vinay Shashidhar, Nishant Pandey and Varun Aggarwal
Source : ACM KDD'15
Advisor : Jia-ling Koh
Speaker : Yi-hui Lee

Outline

- **Introduction**
- Approach
- Experiment
- Conclusion

Introduction

- Goal :

Address the problem of grading spontaneous speech using a combination of machine learning and crowdsourcing

- Insight :

Right set of features can be easily derived by (or with help of) humans, even if they are non-experts

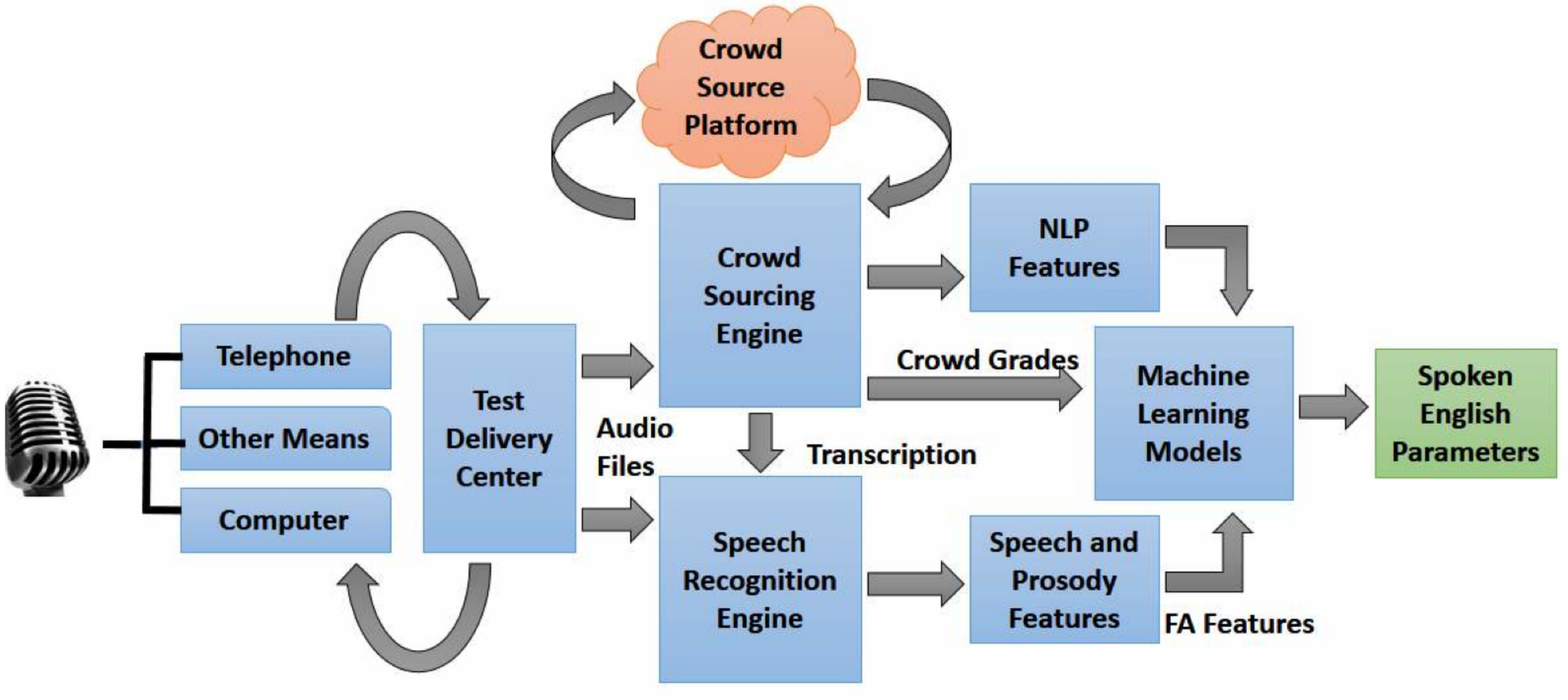
Introduction

- Previous work :

use the crowd to label samples

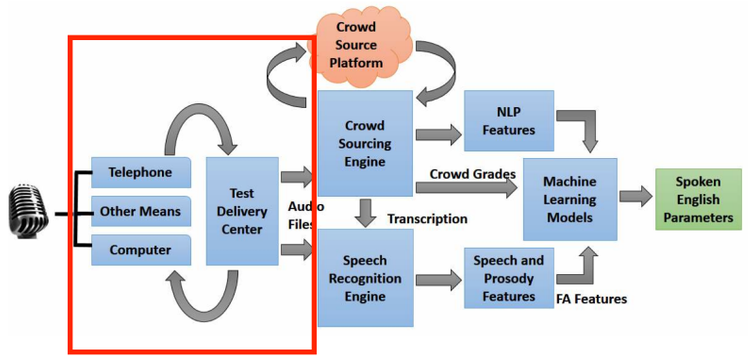
Introduction

- Framework :



Outline

- Introduction
- **Approach**
- Experiment
- Conclusion

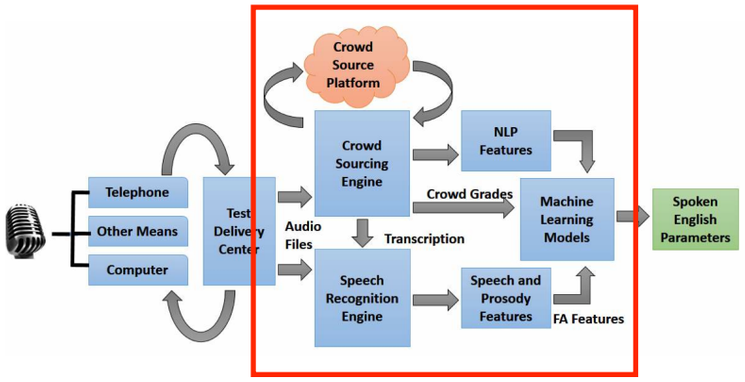


Approach

- Grading Task :

Aspiring Minds' automated speech assessment tool– SVAR

- read sentences aloud
- listen and repeat sentences
- listen to a passage or conversation and answer multiple choice questions
- spontaneously speak on a given topic

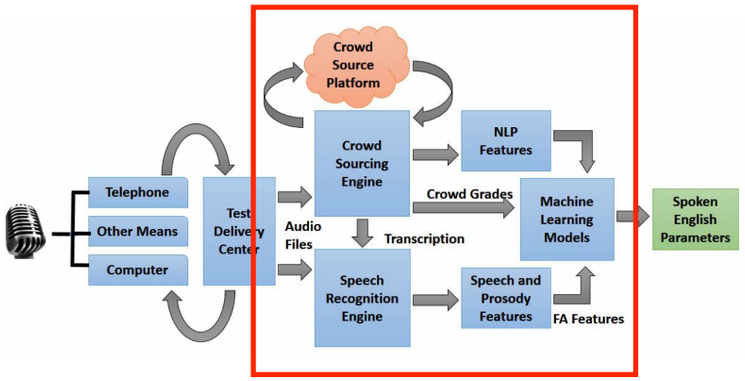


Approach

- Grading Task :

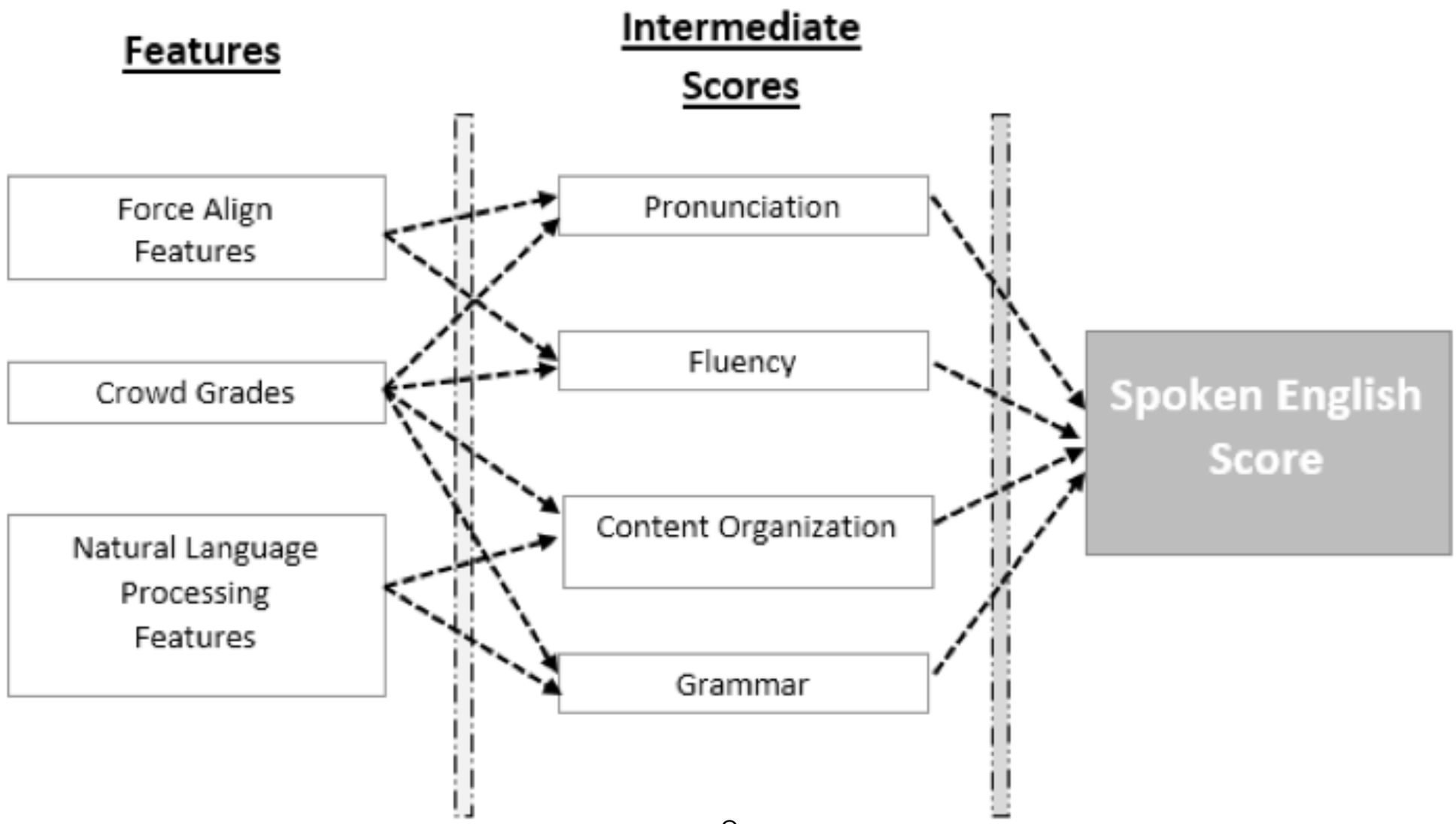
Composite score based on evaluate the spontaneous speech

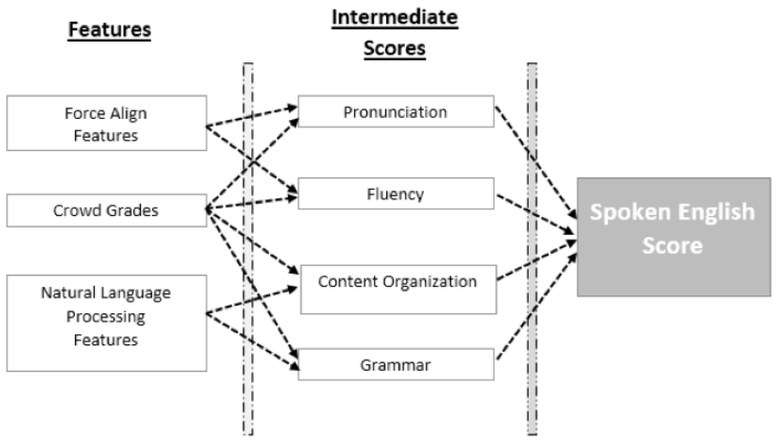
- pronunciation
- fluency
- content organization
- grammar quality



Approach

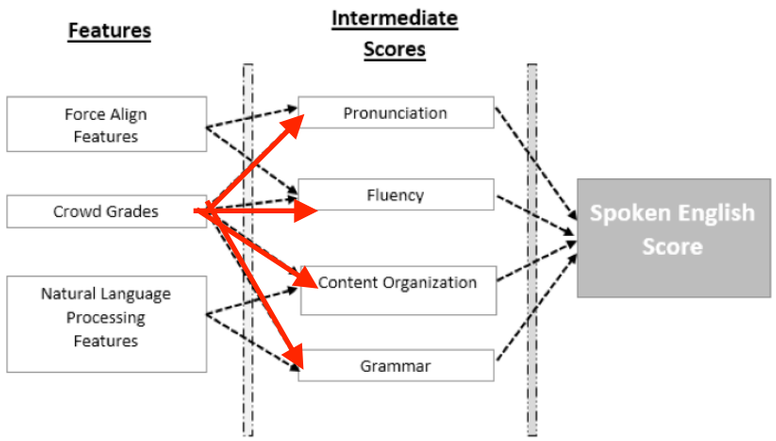
- Features :





Approach

- Features :
  - Crowd Grades(CG)
  - Force Alignment features(FA)
  - Natural Language Processing(NLP)



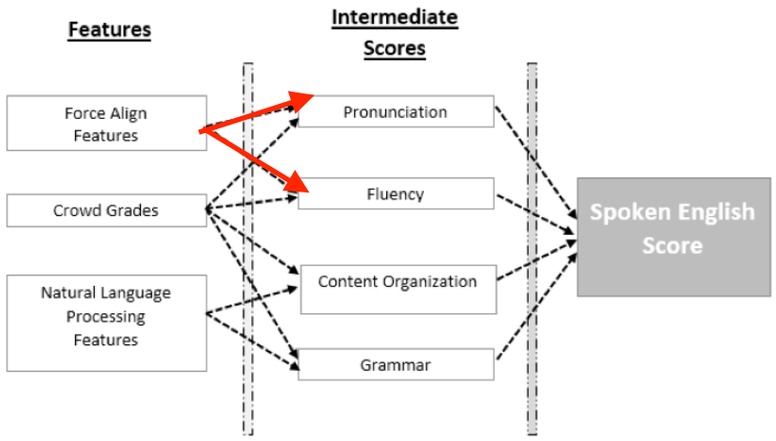
Approach

- Crowd Grades(CG)

Composite score :

pronunciation+fluency+content organization+grammar.

These grades are combined to form a composite score per worker per candidate. These are further averaged across workers to give a final score.



Approach

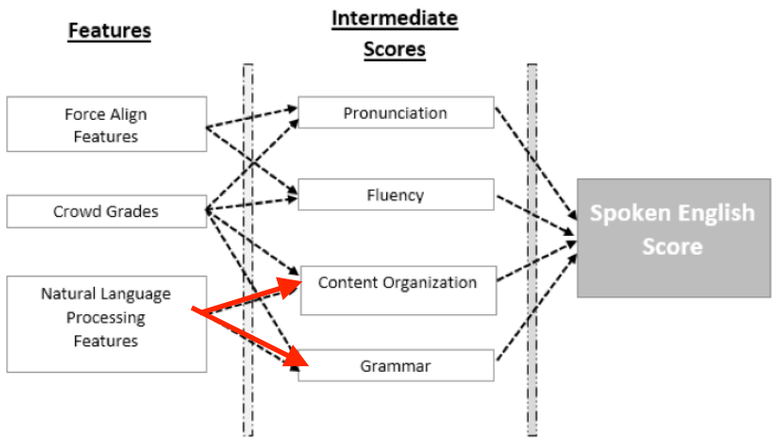
- Force Alignment features(FA)

HTK speech recognizer

TIMIT : corpus of phonemically and lexically transcribed speech of American English speakers of different sexes and dialects.

pronunciation and fluency's score :

rate of speech, position and length of pauses, log likelihood of recognition, posterior probability, hesitations and repetitions, etc.



Approach

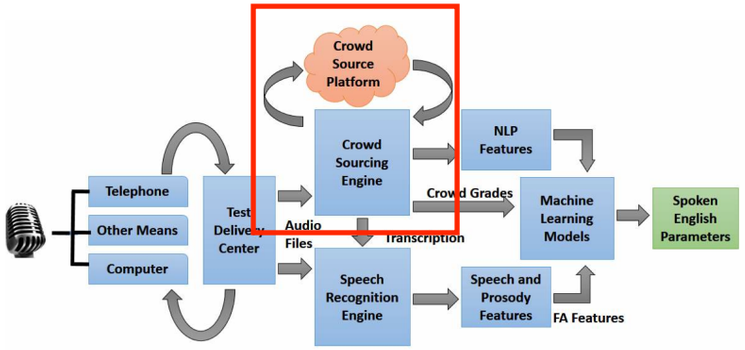
- Natural Language Processing(NLP)

<u>standard NLP packages</u>

grammar and content organization’s score :

a.surface level features —> the number of words complexity or difficulty of words and the number of common words used.

b.semantic features —> the coherency in text, context of the words spoken, sentiment of the text and grammar correctness



Approach

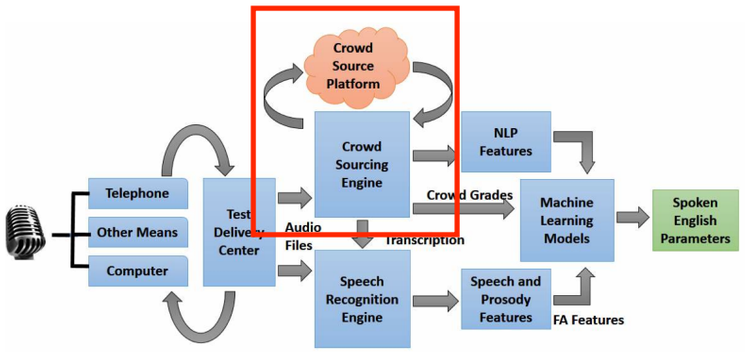
- Crowdsourcing :
  - Amazon Mechanical Turk (AMT)
  - Transcription

-Advanced audio player

-ROVER algorithm :

a.First stage —> the outputs are aligned and a single word transcription network (WTN) is built

b.Second stage —> consists of selecting the best scoring word (with the highest number of votes) at each node.



Approach

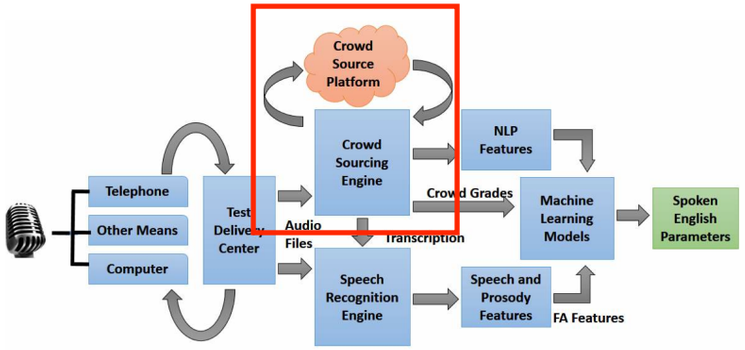
- Crowdsourcing :
  - Amazon Mechanical Turk (AMT)
  - Gold standard tasks :

similar to real tasks and the workers have no way to distinguish between the two

initial risk level : 0.2

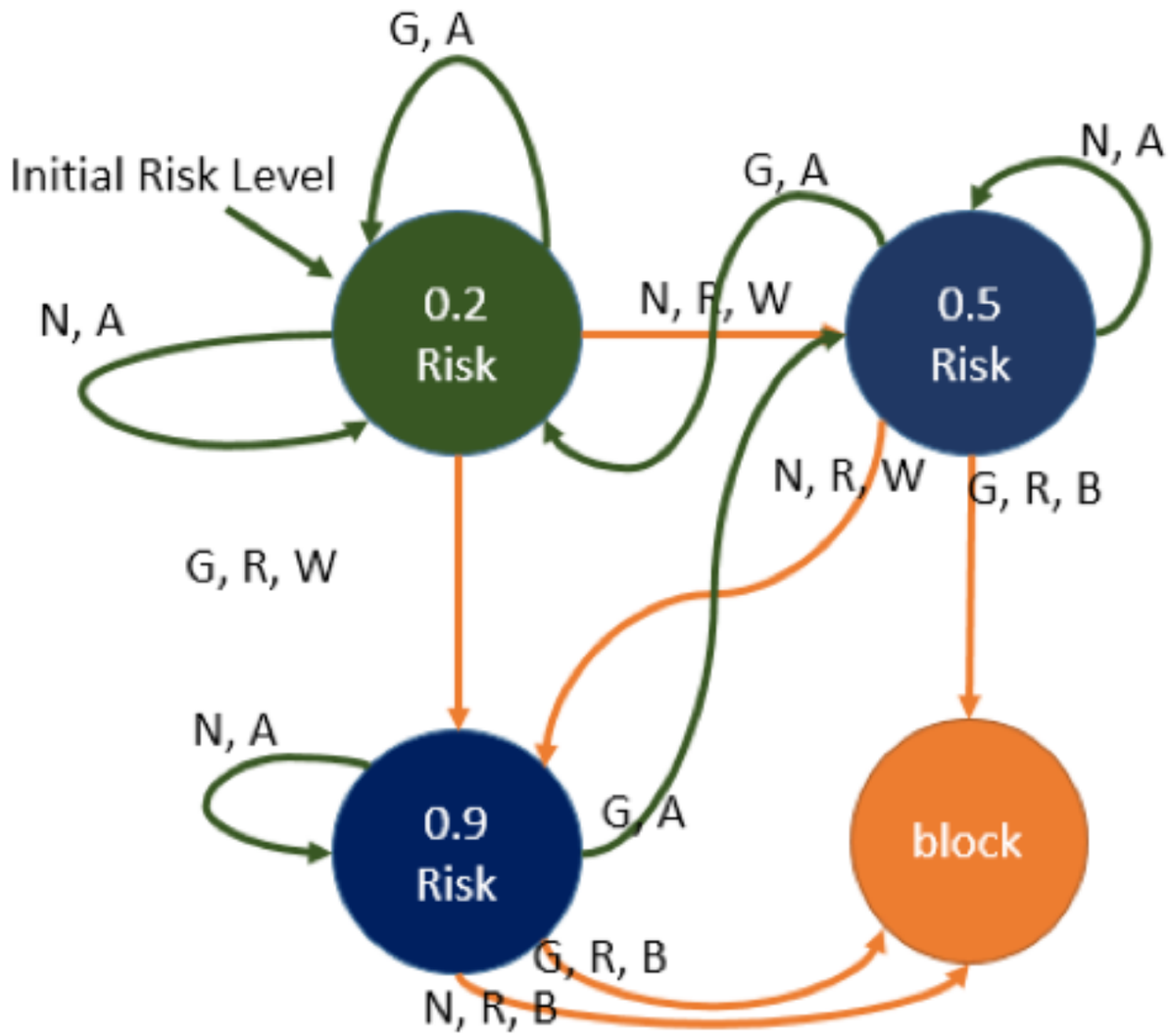
Workers with a higher risk level saw more gold standard tasks.

The risk level of the worker was updated based on his/her performance on the gold standard tasks



Approach

- Crowdsourcing :



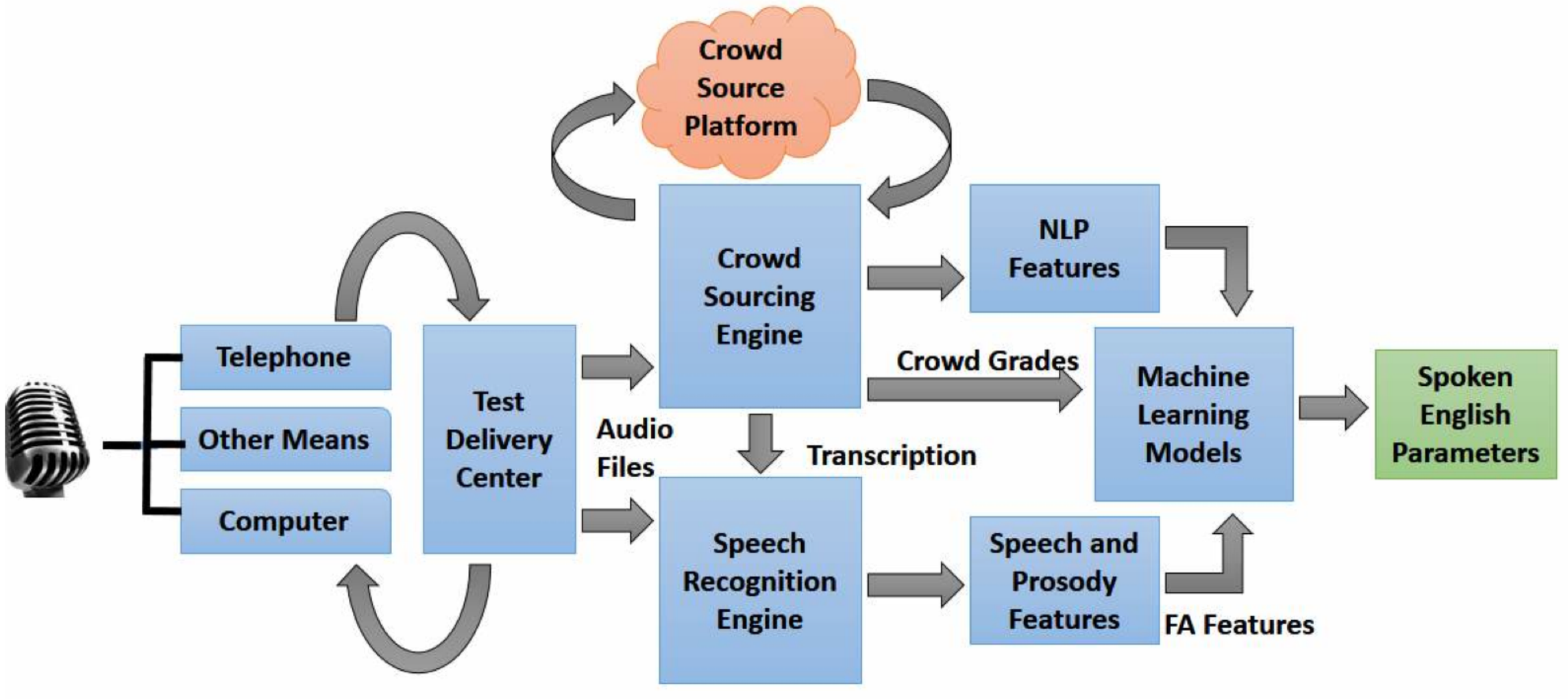
Task Type	
G	Gold Standard Task
N	Normal Task

Task Status	
A	Accept
R	Reject

Proposed Action	
W	Warning
B	Block

Approach

- Framework :



Outline

- Introduction
- Approach
- **Experiment**
- Conclusion

Experiment

- Data Set :

contains a total of 566 spontaneous speech responses comprising of 319 samples from India and 247 samples from Philippines

Experiment

- Feature Sets :

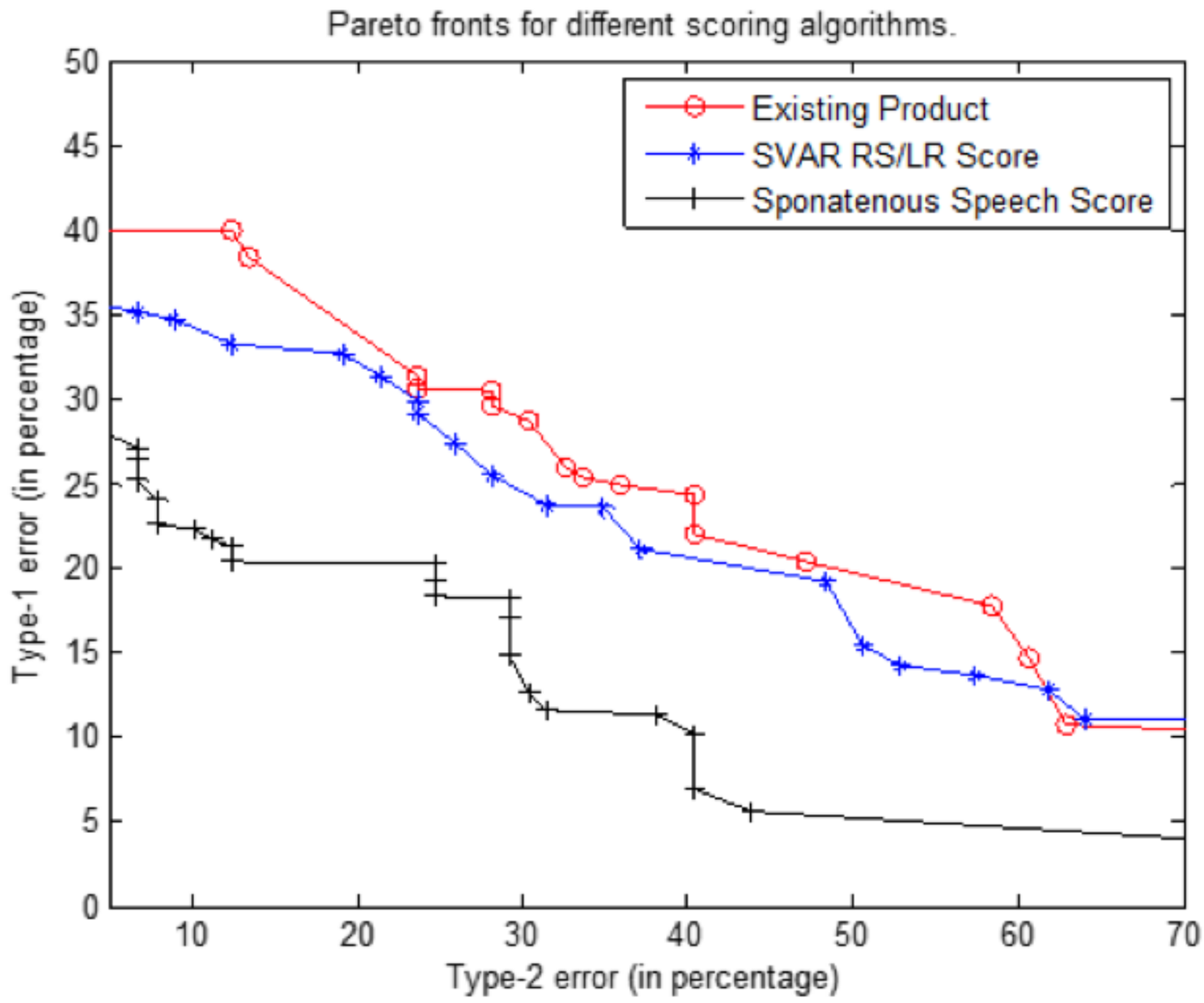
  - Pure ML
  - RS/LR
  - Crowd Grades
  - ML-CS
  - All : NLP, FA features from crowdsourced transcription and Crowd Grades

Experiment

Technique	Model Code	Feature Type	IN Set		PH Set	
			Train r	Validation r	Train r	Validation r
Ridge Regression	RR-1	RS/LR	0.42	0.51	0.47	0.44
	RR-2	Pure ML	0.46	0.48	0.60	0.54
	RR-3	Crowd Grades	0.61	0.67	0.61	0.71
	RR-4	ML-CS	0.64	0.70	0.77	0.60
	RR-5	All	0.74	0.74	0.76	0.79
SVM	SVM-1	RS/LR	0.42	0.51	0.49	0.43
	SVM-2	Pure ML	0.44	0.46	0.60	0.54
	SVM-3	Crowd Grades	0.62	0.57	0.60	0.70
	SVM-4	ML-CS	0.60	0.61	0.76	0.61
	SVM-5	All	0.75	0.74	0.75	0.78
Neural Networks	NN-1	RS/LR	0.47	0.51	0.47	0.47
	NN-2	Pure ML	0.60	0.44	0.55	0.49
	NN-3	Crowd Grades	0.61	0.57	0.58	0.63
	NN-4	ML-CS	0.68	0.58	0.62	0.61
	NN-5	All	0.76	0.75	0.76	0.78

Experiment

- Case Study :



Outline

- Introduction
- Approach
- Experiment
- **Conclusion**

Conclusion

- addressed the problem of evaluating spontaneous speech using a combination of machine learning and crowdsourcing

- able to predict expert grades with much higher accuracy than a machine learning only approach

Conclusion

- Trade-offs :

  - there is a cost for every assessment done and the scalability depends on the number of non-expert workers available
  - our approach doesn’t provide instant grades

- Future work :

  - Real time crowdsourcing