In Proceedings of Interspeech 2020, Shanghai, China

Automatic prediction of confidence level from children's oral reading recordings

Kamini Sabu, Preeti Rao

Department of Electrical Engineering,
Indian Institute of Technology Bombay, Mumbai, India
`kaminisabu,prao@ee.iitb.ac.in`

Abstract

Perceived speaker confidence or certainty has been found to correlate with lexical and acoustic-prosodic features in the spontaneous speech of children interacting with an adult. We investigate the prediction of confidence in the context of oral reading of stories by children with good word recognition skills where we must rely purely on prosodic features. We report a dataset of oral reading recordings that has been manually rated for confidence at the level of text paragraphs of 50-70 words. Several acoustic features computed at different time scales are evaluated via a trained classifier for the prediction of the subjective ratings. Features based on pausing, pitch and speech rate are found to be important predictors of perceived confidence. Also it is seen that the ratings are influenced by signal properties computed across the utterance. When trained on recordings with strong rater agreement, the system predicts low confidence readers with an F-score of 0.70.

Index Terms: confidence prediction, oral reading, children's speech, acoustic-prosodic features

1. Introduction

The acquisition of reading skills is a crucial component of early education. Organizations use oral reading assessment to test literacy as well as spoken language skills. A teacher listens to the student reading and provides scores based on word decoding accuracy, reading speed and fluency [1]. Further, it has been noted that confidence in reading is highly correlated with future reading achievement [2]. Measuring a child's confidence (as an attitude to reading) can thus be useful in deciding the type of reading instruction to provide [3]. In this work, we aim to develop a method for the automatic prediction of the speaker's confidence from an audio recording of oral reading. We present a dataset of recordings by children of grades 5-8, reading age-appropriate texts, and rated by two language teachers for perceived confidence. The dataset comprises oral reading recordings that are largely free of lexical disfluencies and word recognition errors and thus suitable for the investigation of the role of prosodic features in signaling speaker confidence. While automatic confidence prediction has been attempted in the context of children's speech in question-answer scenarios, this is the first attempt on oral reading of fixed text to the best of our knowledge (although there have been several empirical studies of the phenomenon in education research [2, 3]).

Zhang et al. detected whether a child is confident, puzzled or hesitant while speaking to a tutor during a Lego task [4]. They used syntactic, lexical, prosodic and spectral cues to obtain 91.3% accuracy with respect to subjectively rated speaker turns. Part-of-speech tags and bigram probabilities formed the syntactic cues. Lexical cues referred to the classification of the uttered word/phrase as one of: affirmation, digit, filled pauses, indicating knowledge of the topic, indicating no knowledge, indicating uncertainty, reasoning, auxiliary verbs/phrases. Prosodic features comprised percent silence, syllable rate, word and utterance duration, and the F0 and log energy variation features computed across final regions of the utterances. Spectral cues referred to the confidence measures with respect to acoustic phone models trained for each labeled target attitude class. Liscombe et al. [5] considered only acoustic-prosodic features to determine certainness in student's turns as they answered a tutor's question. Features were computed locally in each breath group within a student's turn as well as across the turn that comprised functionals of F0, energy and voicing duration to achieve an accuracy of 76.4% in two-way classification.

Confidence or certainness refers to the belief we place in our own and is signaled through specific acoustic-prosodic combinations. Ponsot et al. [6] observed patterns of high or rising pitch to be associated with lack of confidence, and low or falling pitch with dominance or self-confidence. The same was also observed in answering questions by Pon-Barry [7]. They further found temporal features like speech duration and silence duration to be strongly correlated with the confidence ratings. The features used to realize or perceive confidence may also vary with the regional background of speakers and listeners. Jiang et al. [8] observed that both native and non-native English speakers relied on the systematic variation of pitch to indicate confidence and doubt, while intensity and duration cues differed in their relative importance across the two groups.

The reviewed works relate to question-answer style of speech where the choice of words in the response provides important cues to speaker certainty. Further, the subjective ratings are obtained for short sentences or turns ranging between 1.5 to 4 seconds in duration. In the reading context, however, the words are fixed by the provided text and judgments are made over a few tens of seconds. Disfluencies, hesitations and word recognition errors by the child would indicate poor reading skills and may be expected to be associated with low speaker confidence. However, lack of confidence can be a trait of strong readers as well due to low self-belief about reading skill, with other reasons being shyness or fear of speaking in public. Low self-belief influences reading behaviors and would be useful to detect for the teacher to employ specific instructional strategies [9]. In our own dataset of oral reading in English, described later, we observe that nearly 20% of the recordings with high word recognition scores are rated as displaying low reading confidence. This is the target group of interest in the present work where we seek to develop a method for the automatic detection of confidence level from a recording of oral reading of known text using acoustic-prosodic analysis. In the next section, we describe our dataset and annotation. This is followed by a discussion of the acoustic-prosodic features and their experimental evaluation.

2. Dataset

Oral reading recordings in English by L2 learners in grades 5-8 (age group 10-14 years) were carried out across eight schools in rural and urban areas near Mumbai. The reading material was short English stories in 2-3 paragraphs of about 50-70 words each. A text paragraph is displayed on an Android app screen with recording performed in a relatively quiet school room using a headset microphone. With 80 stories available in the application, the students are each assigned age-appropriate texts for the oral reading exercise with an average of 4 stories per child. Each recording of a paragraph is 20-40 s long based on the reading speed of the child. All recordings comprise of audio at 16 kHz sampling frequency and 16-bit PCM format.

The larger goal of the project is to develop a tool for the automatic assessment of reading skills that can be used in large-scale feedback and monitoring programs by education administrators [10]. Among the dimensions chosen for the reading assessment are word recognition accuracy and speed. Expressiveness in terms of the proper use of prosody is another important attribute that is also indicative of a reader's comprehension [11]. In the present work, we consider the detection of reading confidence by the analysis of the recording of a single paragraph. While lexical miscues and disfluencies can contribute to a listener's perception of low confidence, we restrict ourselves in this work to examining the more challenging scenario of detecting low confidence in otherwise skilled readers given our dataset labeled by human raters.

In order to obtain recordings free of lexical miscues, we used screening by an ASR system [12]. The ASR has hybrid DNN-HMM acoustic models trained for 47 phones corresponding to Indian English and a trigram language model trained on paragraph text with zero gram garbage model in parallel. 'Lexical miscues' by the speaker are defined as the word deletions, insertions and substitutions. Only recordings with number of lexical miscues below 20% of the total number of words were considered for the confidence rating task.

The paragraph utterances were each rated for confidence by each of two English teachers (identified as jrx and cbx) using 3 categories (low, medium, high) as presented on a web-based interface where they also rated other attributes related to fluency and expressiveness. The recordings were randomized (across story text and child) to prepare the ordered sequence for rating. The same order was used by both the raters. The raters were also given an option to mark a recording as not ratable (NA) for any reason. We had a total 2295 recordings from 195 students rated by each expert. Before starting the task, the raters were provided a set of 10 recordings capturing roughly the diversity of skills in the dataset in order to help anchor their judgments. A summary of the labeled dataset appears in Table 1.

Table 1: *Ratings distribution for confidence ratings*

jrx/cbx	low	medium	high	NA
low	196	700	196	16
medium	40	445	509	0
high	0	47	162	0
NA	8	16	3	28

Figure 1 shows the distribution of the ratings across the 3 categories for each of the raters. We see a clear bias toward the lower levels of low and medium confidence in the case of jrx while the opposite trend holds for cbx. A Pearson linear correlation coefficient of 0.43 and Spearman's rank-order correlation of 0.44 is obtained which is considered 'moderate' inter-agreement [13]. There are 196 recordings where the raters differ greatly (low vs high) which needs looking into.

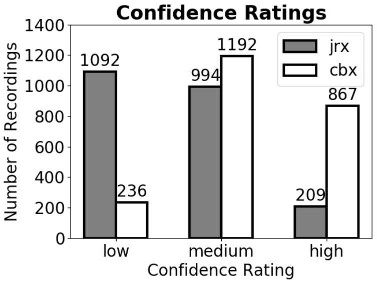

Figure 1: *Distribution of confidence markings by two raters*

Both the raters agree completely on 803 recordings from 155 students. Of these, 196 recordings are rated as low, 445 as medium and 162 as highly confident. Since the two raters fully agree on less than half the recordings, we also attempt to model individual raters separately in the reported experiments.

3. Acoustic-Prosodic Feature Extraction

Based on the literature and observations from our dataset, we see that the students who read rapidly and smoothly are more likely to be rated highly confident. The students who hesitate while reading, read slowly, or take too many pauses are typically perceived as lacking in confidence. Highly confident children also tend to vary their pitch and volume across the utterance. Further, their reading is emphatic and energetic with clear articulation. We compute acoustic signal features motivated by the speech production correlates of these speaker traits.

Pause: Pauses are used as cues to syntactical breaks in normal reading. Outside this context, pauses are indicative of hesitation. ASR based word alignment is used to identify the non-speech regions that occur between words. Non-speech duration longer than 200 ms is termed a 'silence', further classified into pauses (< 500 ms) and long pauses (> 500 ms).

Speech Rate: Perceived reading speed is related to 'speech rate' or the count of sub-word units such as phones and syllables per unit time [14]. The ASR decoder hypothesis provides the segmentation into words from which we further compute phone and syllable boundaries. The recording duration is computed from the onset of the first word to the end of the last uttered word (i.e. neglecting silences at either ends). We also compute an 'articulation rate' as the number of syllables divided by actual speech duration (omitting intervening non-speech regions).

Pitch: Pitch variations can be computed from a contour of the fundamental frequency (F0) across the utterance. Autocorrelation based F0 detection is performed to get the contour sampled at 10 ms frame intervals. We use 20 ms Hamming windows, and restrict the pitch search range to 100 Hz to 520 Hz. The voicing threshold is based on the normalized autocorrelation peak corresponding to the detected F0 and is set at its typical level of 0.45 for speech in quiet [15]. The contour is normalized for speaker dependence by division with the minimum detected F0 value across the recording.

Intensity: Loudness and its dynamics are the perceptual correlates of signal intensity and its variations. While spectral envelope is also affected, this is accounted for by voice quality features. Short-time intensity is computed for every 10 ms frame of audio as the logarithmic energy (in dB) with reference to the silence region energy in order to compensate for microphone gain variations across different recordings.

Voice Quality: Voice quality refers to the perceptual correlates of the glottal excitation in speech production. Spectral envelope tilt reduces with increased vocal effort; this can be captured via the increased high frequency content in the envelope. Relative intensity in four vocal frequency bands [16]: 60 Hz to 400 Hz, 400 Hz to 2 kHz, 2 kHz to 5 kHZ and 5 kHz to 8 kHz are measured over 10 ms frame intervals throughout the speech regions across the utterance. Another important aspect of voice quality is the deviation from modal voice to incorporate some breathiness or hoarseness. This has traditionally been measured in terms of the Harmonic-to-Noise Ratio (HNR) also computed every 10 ms throughout the speech regions.

Articulation Clarity/Enunciation: Clarity in articulation is reflected in the spectral envelope changes associated with large vocal tract movements. Also referred to as 'enunciation', it can be captured by the cepstral distance from the neutral schwa sound cepstrum [17]. The latter can be computed as the average of the voiced cepstrum across the recording. We compute 20 cepstrum coefficients for every 10 ms frame. Cepstral distance is computed for each frame as the Euclidean distance from the estimated average voiced cepstrum.

3.1. Recording Level Features

We compute features across the paragraph audio as follows:

Pause: The features computed in pause category are: count of silences, pauses and long pauses, mean and standard deviation (s.d.) of silence durations, pause durations and long pause durations, percentage of silence duration in the recording.

Speech Rate: Speech rate is computed across recording in terms of number of syllables per second, articulation rate (number of syllables per second excluding silence regions) and number of phones per second.

Pitch: Min, max, mean and s.d. of values in the normalized pitch contour are estimated across the voiced region of recording. Besides, the slope is computed for every word segment's pitch contour through line fitting and the fraction of words with positive slope (rising pitch) is treated as a feature.

Intensity: Min, max, mean and s.d. of intensity contour are estimated across the non-silence region of recording.

Voice Quality: Mean and s.d. are computed for each of the relative spectral intensity contours and HNR contour. The computation is performed only across the speech regions.

Enunciation: The mean and s.d. of the enunciation contour are computed as features. Only speech regions are considered during this computation.

3.2. Segment Level Features

Besides the recording level features, we also computed features across shorter segments of the recording with the segmentation itself effected in different ways. The temporal sequence of features captures the local variation of the corresponding attributes that could be useful in confidence prediction given that human listeners probably aggregate information across the recording duration in different ways. The following three segmentations were considered:

1. Fixed 10 word long windows with 5 word overlap
2. Speech chunks separated by silence regions
3. Five equal length segments in terms of the number of words contained.

All the features computed across the recording (except pause related) were also computed across the local window. Additionally, the pitch contour slope is computed across the window and across the final 100 ms voiced interval of the window. Next, all the features are aggregated in two ways: Mean and s.d. were taken across the set of segments for all the locally computed features. Instead of s.d., max value was computed for fraction of words with rising pitch. The second method is to concatenate the chunk-level features directly. In the next section, we present the evaluation of the different feature sets both in type and time-scale of observation.

4. Classification Results and Discussion

First we considered the dataset of 803 recordings where both the raters agree. Since the three confidence classes were not balanced in this dataset, we removed some of the medium rated recordings to get 600 recordings from 155 speakers in total. With this, we had 196 recordings with low confidence, 242 with medium confidence and 162 with high confidence. Then we trained a random forest (RF) regressor [18] in 5-fold cross-validation mode with each of the five feature groups separately. RF is implemented using scikit learn library [19] with 500 regression tress, each trained on maximum of 200 training samples with only square root number of all input features. Gini based feature importance values were obtained from the model in each case. The top features with importance values summing up to 0.95 are listed in Table 2 when RF regressor is trained on recording level features. These top features from each feature group were used to train regressor for 'Overall' condition. Further, the top features in this case are indicated in Table 2 under the entry 'Overall'.

Table 2: *Group-wise top features at recording-level from regression based confidence prediction system trained on 600 recordings where both raters agree*

Feature group	Important features
Pause	#sil, #pauses, #long pauses, s.d. of sil and pause dur, mean dur of sil, pause and long pause
Speech rate	articulation rate, phone rate
Pitch	fraction of words with rising slope, pitch mean to s.d. ratio, pitch s.d., pitch mean
Intensity	intensity mean, s.d., min
Voice quality	mean and s.d. of band1, band2, band3 relative intensity and HNR, mean of band3 relative intensity
Enunciation	s.d. of enunciation
Overall	mean sil, pause, long pause dur, # pauses, # long pauses, s.d. of sil, articulation rate, phone rate, pitch mean to s.d. ratio, pitch mean, fraction of words with rising slope, pitch s.d., min intensity, band1 s.d., band2 mean, band2 s.d., band3 mean, band3 s.d., HNR mean, HNR s.d., band4 mean, enunciation s.d.

The performance for the 5-fold cross-validation is reported in terms of Pearson's correlation coefficient in Table 3 for each feature group as well as the overall set. The first column shows results when recording level features were used.

As can be seen from Table 3, pause characteristics, speech rate, and pitch features are the major indicators of confidence

Table 3: *Performance of regression on agreed 600 recordings (in 5-fold cross-validation) in terms of Pearson's correlation coefficient with different feature groups and windowing conditions. Number of features input to the regressor indicated in parentheses*

Features type Features group	Recording level features	Aggregates across 10 word window	Aggregates across silence separated window	Aggregates across equal length window	Concatenation across equal length window
Pause	0.54	-	-	-	-
Speech rate	0.63	0.57	0.59	0.55	0.59
Pitch	0.62	0.70	0.66	0.67	0.63
Intensity	0.37	0.48	0.50	0.46	0.44
Voice quality	0.43	0.39	0.50	0.41	0.42
Enunciation	0.09	0.22	0.00	0.23	0.20
Overall	0.76 (27)	0.72 (42)	0.72 (38)	0.72 (40)	0.69 (118)

level. The intensity and voice quality features contribute similarly. The enunciation features contribute the least. All the features together help in estimating the confidence level with high correlation. However, as can be seen from Table 2, if we input all the selected features from each group together to the regressor, enunciation appears as an important feature meaning that it can help classification, but not on its own.

Table 3 also shows the performance when aggregates of local features across different windows and the concatenation of all chunk level features are considered. The performance improves for intensity and enunciation, but reduces for voice quality features. Though the performance trend is still the same with the highest performance for pitch and lowest for enunciation, we can see overall reduction in performance as we moved from recording level computation to local feature computation.

As the performance improved for intensity and voice quality after taking window level features, we tried concatenating these with the recording level features; but the performance remained the same as the case when only recording level features were used. Further, we also added locally computed pitch features to get Pearson's correlation coefficient of 0.76, still similar to the performance obtained with only recording level features.

Next, we considered the task of detecting 'low confidence' recordings by two-way classification where medium and high confidence recordings are grouped into one class. This emulated the realistic task of flagging low confidence readers. The corresponding results are shown in Table 4. We see the best performance with the lower dimensional recording level features. The concatenation of recording level features with all the chunk level features was also tested.

Table 4: *Performance on agreed 600 recordings (5-fold cross validation) in terms of % accuracy and low confidence detection F-score for two-way classification.*

Features type (No. of features)	Acc. (%)	F-score
Recording level (36)	82.3	0.70
Fixed window (52)	82.0	0.70
Silence separated window (52)	82.0	0.69
Equal length window (52)	82.1	0.69
All chunk features concatenated (130)	80.5	0.65
All chunk features concatenated with recording level features (164)	82.5	0.69

We also performed the predictions for the individual raters to better understand the relative weighing of cues by individual listeners. We used only the recording level features here. Performance is reported in terms of percentage accuracy for every feature subgroup in Table 5. This seems to be almost similar for each feature group, except for speech rate. This indicates that both the raters perceive confidence level using the same features, but their threshold to discriminate the three levels are different. This is in line with the observations in 1, where we find a bias in the ratings of jrx towards the lower level labels. The chance experiment accuracy is also reported for ready reference.

Table 5: *Three-way classification accuracy (%) on 2295 recordings rated by individual rater for (5-fold cross-validation) using recording level features*

Feature group	jrx	cbx
Pause	58.6	56.6
Speech rate	57.7	63.9
Pitch	60.5	60.1
Intensity	52.9	51.7
Voice quality	53.3	53.1
Enunciation	48.0	48.8
Overall	64.0	64.1
Chance Agreement	51.9	47.6

5. Conclusion

We addressed the automatic prediction of confidence from speech of children reading aloud. Perceived confidence was rated at three levels by two language teachers based on story text recordings of duration 20 - 40 s. A number of acoustic-prosodic features, were evaluated for confidence prediction. It was found that the fast speaking students were considered more confident by the raters followed by consistent pausing, and greater pitch and intensity variations. Articulation clarity did not matter much, perhaps because we considered students who had good word decoding skills. Further, features computed across the paragraph recording performed better than the aggregates of locally computed features. We obtained an accuracy of 65% in predicting confidence at three levels, while 82% accuracy was achieved in the binary classification with F-score of 0.70 in flagging students with low confidence.

Future work will involve assigning the ratings task to a larger number of experts using different sequences of presentation of recordings to obtain more reliable collective ratings. The scope of the work will also be extended to weak readers in which case lexical features will be included with the acoustic-prosodic features in the classification task.

6. Acknowledgment

Authors gratefully acknowledge the support by Ministry of Elect. and IT, Govt. of India under Visvesvaraya PhD scheme.

7. References

[1] M. C. Danne, J. R. Campbell, W. S. Grigg, M. J. Goodman, A. Oranje, and A. Goldstein, "The Nation's Report Card: Fourth-Grade Students Reading Aloud: NAEP (The National Assessment of Educational Progress) 2002 Special Study of Oral Reading," National Center for Education Statistics, U.S. Department of Education, Tech. Rep., 2005.

[2] S. P. McGeown, R. S. Johnston, J. Walker, K. Howatson, A. Stockburn, and P. Dufton, "The relationship between young children's enjoyment of learning to read, reading attitudes, confidence and attainment," *Educational Research*, vol. 57, no. 4, pp. 389–402, 2015.

[3] Q. Gao, H. Wang, F. Chang, Q. An, H. Yi, K. Kenny, and Y. Shi, "Feeling bad and doing bad: student confidence in reading in rural China," *Compare: A Journal of Comparative and International Education*, vol. 16, pp. 1–20, 2020.

[4] M. Hasegawa-Johnson, S. Levinson, and T. Zhang, "Children's emotion recognition in an intelligent tutoring scenario," in *Proceedings of INTERSPEECH*, Jeju Island, Korea, 2004, pp. 1441–1444.

[5] J. Liscombe, J. Hirschberg, and J. J. Venditti, "Detecting certainness in spoken tutorial dialogues," in *Proceedings of INTERSPEECH*, Lisbon, Portugal, 2005, pp. 1837–1840.

[6] E. Ponsot, J. J. Burred, P. Belin, and J.-J. Aucouturier, "Cracking the social code of speech prosody using reverse correlation," *Psychological and Cognitive Sciences*, vol. 115, no. 15, pp. 3972–3977, 2018.

[7] H. Pon-Barry, "Prosodic manifestations of confidence and uncertainty in spoken language," in *Proceedings of INTERSPEECH*, Brisbane, Australia, 2008, pp. 74–77.

[8] X. Jiang and M. D. Pell, "Predicting confidence and doubt in accented speakers: Human perception and machine learning experiments," in *Proceedings of Speech Prosody*, 2018, pp. 269–273.

[9] H. Pon-Barry, K. Schultz, E. O. Bratt, B. Clark, and S. Peters, "Responding to student uncertainty in spoken tutorial dialogue systems," *International Journal of Artificial Intelligence in Education*, vol. 16, no. 2, pp. 171–194, 2006.

[10] P. Rao, P. Swarup, A. Pasad, H. Tulsiani, and G. Das, "Automatic assessment of reading with speech recognition technology," in *Proceedings of International Conference on Computers in Education*, Mumbai, India, 2016.

[11] K. Sabu, K. Kumar, and P. Rao, "Automatic detection of expressiveness in oral reading," in *Proceedings of INTERSPEECH (Show and Tell)*, Hyderabad, India, 2018, pp. 1489–1490.

[12] K. Sabu, P. Swarup, H. Tulsiani, and P. Rao, "Automatic assessment of children's L2 reading for accuracy and fluency," in *Proceedings of SLaTE*, Stockholm, Sweden, 2017, pp. 121–126.

[13] H. Akoglu, "User's guide to correlation coefficients," *Turkish Journal of Emergency Medicine*, vol. 18, no. 3, pp. 91–93, 2018.

[14] C. Cucchiarini, H. Strik, and L. Boves, "Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology," *Journal of the Acoustical Society of America*, vol. 107, no. 2, pp. 989–999, 2000.

[15] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," in *Proceedings of the Institute of Phonetic Sciences*, Amsterdam, 1993, pp. 97–110.

[16] V. V. Patil and P. Rao, "Detection of phonemic aspiration for spoken Hindi pronunciation evaluation," *Journal of Phonetics*, vol. 54, pp. 202–221, 2016.

[17] N. G. Ward, J. C. Carlson, and O. Fuentes, "Inferring stance in news broadcasts from prosodic-feature configurations," *Computer Speech and Language*, vol. 50, pp. 85–104, 2018.

[18] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[19] scikit-learn, "Ensemble methods – scikit-learn," https://scikit-learn.org/stable/modules/ensemble.html#id7, scikit-learn documentation.