END-TO-END WORD-LEVEL DISFLUENCY DETECTION AND CLASSIFICATION IN CHILDREN'S READING ASSESSMENT

Lavanya Venkatasubramaniam[*]*, Vishal Sunder*[*]*, Eric Fosler-Lussier*

The Ohio State University

ABSTRACT

Disfluency detection and classification on children's speech has a great potential for teaching reading skills. Word-level assessment of children's speech can help teachers to effectively gauge their students' progress. Hence, we propose a novel attention-based model to perform word-level disfluency detection and classification in a fully end-to-end (E2E) manner making it fast and easy to use. We develop a word-level disfluency annotation scheme using which we annotate a dataset of children read speech, the reading races dataset (READR). We also annotate disfluencies in the existing CMU Kids corpus. The proposed model significantly outperforms traditional cascaded baselines, which use forced alignments, on both datasets. To deal with the inevitable class-imbalance in the datasets, we propose a novel technique called **HiDeC** (**Hi**erarchical **De**tection and **C**lassification) which yields a detection improvement of 23% and 16% and a classification improvement of 3.8% and 19.3% relative F1-score on the READR and CMU Kids datasets respectively.

Index Terms— disfluency detection, attention-based models, children's speech, mispronunciation detection

1. INTRODUCTION

Many modern day deep learning based computer-aided pronunciation training (CAPT) tools operate at the phonological level. These approaches predict the phoneme sequence of a mispronounced utterance [1, 2, 3, 4]. A correct diagnosis constitutes the case when the predicted phoneme is the same as the human transcribed phoneme which in turn is different from the canonical phoneme. In reading assessment for children, it can be difficult and time consuming for a teacher to analyze phone-level predictions for each individual child.

An easy way to assist teachers in correcting children's disfluencies is to get disfluency categories at the word level. Having a finite set of disfluency categories over a child's vocabulary helps in identifying words where a child needs help. An approach by Black et al. [5] follows this route, but requires that the word boundaries be known in advance for training a classifier. This is done by having the children read one word only instead of fluent readings of paragraphs. In this paper, we explore ways to classify disfluent words in a continuous stream of read passage where word boundaries are not known in advance. For this, we annotate two datasets of children's speech with word level disfluency tags. These are the reading races (READR) corpus [6] and the CMU Kids corpus [7] which are recordings of children reading passages.

It is also important to make these models as streamlined and robust as possible. Approaches by Proencca et al. [8] and Duchateau et al. [9] use automatic speech recognition (ASR) to get transcripts for children's speech and then analyse these to determine the presence of disfluencies. This approach requires constructing grammar structures which might not be very robust and do not allow for diverse categories of error types. In this work, we propose a fully end-to-end (E2E) model which operates on speech directly to classify spoken words. In particular, we adapt the attention-based listen, attend and spell (LAS) model [10] for this purpose. The cross-modal attention mechanism in the LAS model serves as a natural choice for extracting relevant speech segments.

A major challenge in developing a machine learning model that operates on a finite set of disfluency classes is the high degree of class imbalance [11] as only a small fraction of a child's vocabulary actually has disfluencies. We explore two different techniques to address this issue. The first is the recently proposed major feature weakening strategy [12] which scales gradients of the majority class by injecting noise into it's features. The second is a novel proposal which we call **HiDeC** (**Hi**erarchical **De**tection and **C**lassification) which reduces the degree of class imbalance by training one classifier only to detect disfluencies and a second classifier to classify the type of disfluency. The second classifier is only trained on disfluent examples. We show that HiDeC deals with the catastrophic effects of class imbalance to a substantial degree. In the disfluency detection task, HiDeC yields a relative improvement of 23% and 16% F1-score on the READR and CMU Kids corpus. In the disfluency classification task, we get a relative improvement of 3.8% and 19.3% F1-score on the two datasets using HiDeC.

2. DATASETS AND ANNOTATION

We annotate two datasets of children's disfluent speech.

[*] Equal contribution. Order alphabetical.

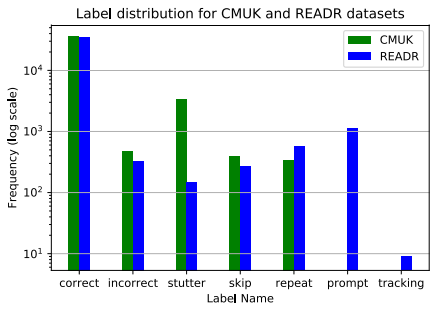

Fig. 1. Label distribution in the CMUK and READR datasets. Label frequency is shown in y-axis in the log scale.

Reading Races (READR) [6]: This is a 15-hour corpus of speech data consisting of one-minute audio clips of children of ages 5 to 8. This dataset has audio from children who have reading difficulties owing to various factors, making it a very challenging dataset. The children are asked to read an english passage in the supervision of a teacher who assists them.
CMU Kids (CMUK) [7]: This is a 9 hour corpus of speech data comprising of sentences read by children aging from 6 to 11 years old. It consists of 24 male and 52 female speakers. We are releasing the annotation for this dataset[1].

Each word in the read passage is assigned one of the following 7 disfluency categories given the corresponding speech:
Repeat - A word is uttered more than once. Eg: "...every every every day...", the word "every" is labeled as a "Repeat".
Incorrect - An uttered word does not match the actual word. Eg: "...get her the mason...", the word "medicine", incorrectly pronounced as "mason", will be labeled "Incorrect".
Stutter - Some syllables in a word are repeated. Eg: "...car will not st st start....", "start" will be labeled as "Stutter".
Skip - The speaker omits a word while reading the sentence. Eg: Uttered: "...she a lot of fun...". Actual: "...she had a lot of fun...". The word "had" will be labeled as "Skip".
Prompt - When a child is struggles to pronounce a word, the teacher intervenes and pronounces the word in the background which is captured in the audio. Such words are labelled as "Prompt". This label is only present in the READR dataset as it is collected in the presence of a teacher.
Tracking - The child loses track of the sentence they are reading. Eg: Uttered: "...I came home and Saturday...". Actual: "...I came home and saw...Today is a Saturday...". The words from "and" till "a" will be labeled as "Tracking".
Correct - The rest of the uttered words that do not fall under the previous 6 classes will be labeled as "Correct".

The READR dataset was annotated manually with these classes. The CMUK dataset, on the other hand, already contained generic remarks on the mispronunciations and disfluencies in each of the speech data instances. We transformed these remarks into the above-mentioned classes.

[1] https://github.com/OSU-slatelab/Annotating_CMUK

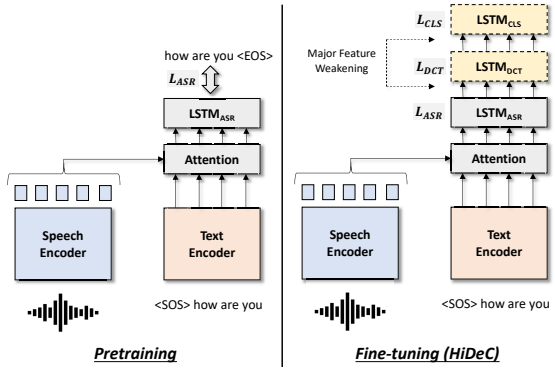

Fig. 2. Model Overview. *Left:* Pretraining follows the LAS ASR setup [10]. *Right:* Proposed HiDeC framework. The pretrained model is adapted to perform hierarchical detection and classification by adding two LSTMs on top of the speller.

The distribution of classes in the two datasets is shown in figure 1. Note that most words in the datasets are marked as correctly pronounced which means that only a small fraction of the words contain disfluencies. This makes the dataset highly class-imbalanced due to the number of correct labels being orders of magnitude higher than other labels.

Given the speech signal and the canonical transcript, we try to solve two related tasks.
Disfluency Detection: This is a binary classification problem where the task is to detect whether a disfluency is present or not for each word in the canonical transcript given the speech.
Disfluency Classification: This is a 7-class classification problem which unlike detection is more fine-grained in that we need to predict the actual class of the disfluency present in each word of the canonical transcript given the speech.

3. MODEL OVERVIEW

Unlike the traditional ASR task, where canonical transcripts are only used during training, ours is a reading assessment task where canonical transcripts are available at test time also. At test time, we use the available text to extract speech sections corresponding to words and classify these into disfluency classes. A natural choice for building an E2E reading assessment system is to use attention-based models. These models have the advantage that the explicit attention mechanism produces attention weights over relevant parts of the speech signal corresponding to the given word units. Learning how to attend over speech given the text can be effectively done by pretraining an attention based ASR model.

3.1. Pretraining

We pretrain a Listen-Attend-Spell (LAS) [10] based ASR model on 960 hours of Librispeech data [13] prior to using it

for disfluency detection and classification. A speech encoder extracts a sequence of speech features which are attended to by the previously predicted context to predict the next token in an auto-regressive manner. Given the speech, $\mathbf{X}$ and the context $\{y_{i-1}, y_{i-2}, ..., y_0\}$, we compute the the ASR loss as,

$$L_{ASR} = -\log(\mathrm{P}(y_i|y_{i-1}, y_{i-2}, ..., y_0, \mathbf{X}))$$

We model this following Tüske et al. [14] but using a unidirectional LSTM so that our model can be used in real-time.

3.2. Finetuning

To the pretrained model, we cascade $\mathrm{LSTM}^{DCT/CLS}$ on top of the speller LSTM to extract word level features for detection/classification. If the output of the speller is a sequence $\mathbf{x}^{ASR}$ of length T, then the prediction is computed as,

$$\mathbf{x} = \mathrm{LSTM}^{DCT/CLS}(\mathbf{x}^{ASR})$$
$$\mathrm{P}(.|x_t) = \mathrm{softmax}(\mathbf{W}x_t + b)$$

Here, $\mathbf{x} = (x_1, ..., x_T)$ is the output sequence which is passed to a single classification layer with weights $\mathbf{W}$ and b. Then, the detection/classification loss is computed as,

$$L_{DCT/CLS} = -\frac{1}{T}\sum_{t=1}^{T}\log(\mathrm{P}(y_t|x_t))$$

Here, y_t is the ground-truth label for the representation x_t. To help with the adaptation to the new speech domain, we add the in-domain ASR loss to the above loss,

$$L_{MTL} = L_{DCT/CLS} + L_{ASR}$$

As the datasets are highly class-imbalanced, we incorporate two different techniques to the above framework.

Major feature weakening (MFW): This technique was introduced by Ye et al. [12] and proposes to prevent overfitting by adding noise to the majority training feature by interpolating between two data points in a batch during training. This scales the gradients for majority features such that the training is balanced for all classes. For a given data point $\mathbf{x}$ with label y, we sample a random data point $\tilde{\mathbf{x}}$ from the batch. Then,

$$\mathrm{MFW}(\mathbf{x}) = (1-\lambda)\mathbf{x} + \lambda\tilde{\mathbf{x}}$$
$$\lambda \in (0.0, 0.5] \text{ and } \lambda \propto N_y$$

where N_y is the frequency of label y in the dataset. λ can be sampled using any policy provided it follows the above constraints. We followed the same policy as Ye et al. [12].

Hierarchical detection and classification (HiDeC): While MFW deals with the issue of class imbalance to some extent, we further improve performance by co-training the model to perform detection and classification together in a hierarchical manner (see figure 2). First, the detection module, LSTM_{DCT} predicts whether a disfluency was present in the word or not, then, if a disfluency was present, LSTM_{CLS} classifies the type of disfluency. Formally,

$$\mathbf{x}^{DCT} = \mathrm{MFW}(\mathrm{LSTM}^{DCT}(\mathbf{x}^{ASR}))$$
$$\mathbf{x}^{CLS} = \mathrm{MFW}(\mathrm{LSTM}^{CLS}(\mathbf{x}^{DCT}))$$
$$\mathrm{P}(.|x_t^{DCT}) = \mathrm{softmax}(\mathbf{W}^{DCT}x_t^{DCT} + b^{DCT})$$
$$\mathrm{P}(.|x_t^{CLS}) = \mathrm{softmax}(\mathbf{W}^{CLS}x_t^{CLS} + b^{CLS})$$

Here, $\mathrm{P}(.|x_t^{DCT})$ is the probability whether a disfluency is present in a word or not and $\mathrm{P}(.|x_t^{CLS})$ is the distribution over the possible set of disfluency classes. The detection and classification losses are then computed as,

$$L_{DCT} = -\frac{1}{T}\sum_{t=1}^{T}\log(\mathrm{P}(y_t^{DCT}|x_t^{DCT}))$$
$$L_{CLS} = -\frac{1}{T}\sum_{t=1}^{T}\log(\mathrm{P}(y_t^{CLS}|x_t^{CLS}))$$

Here, y_t^{DCT} is 1 if a disfluency is present in the word otherwise it is 0. y_t^{CLS} is the actual disfluency class. The final loss, $L_{HiDeC} = L_{DCT} + L_{CLS} + L_{ASR}$.

During training, L_{CLS} is set to 0 for cases where disfluency is not present. Thus, LSTM^{CLS} is trained only on disfluent examples. At test time, an instance is first passed through LSTM^{DCT} and only if it is predicted to have disfluencies, it is passed on to LSTM^{CLS} to predict the actual class.

This type of training has the advantage that the class-imbalance is explicitly reduced for the classification task. As the lower level LSTM is already tasked with predicting whether the word was fluent, the upper level LSTM never needs to see the fluent cases (the major contibutor to class imbalance) during training resulting in balanced training.

4. EXPERIMENTS AND RESULTS

As the size of the READR and CMUK datasets is small, we perform a 5-fold cross-validation on both these datasets.

The results are reported in table 1. Row (1) is a trivial baseline predicting everything as "correct". This acts as the lower bound for the highly skewed datasets. In row (2), we try to match two word segment based models proposed in literature for disfluency detection as closely as possible. Specifically, Lea et al. [15] extract 3 second audio clips and annotate them with different stutter classes. Each audio segment is fed to a neural speech encoder for classification. Models in Black et al. [5] also work at the word level. To build a comparable setup, we run the Montreal Forced Aligner (MFA) [16] to get word boundaries in the children's speech. Speech segments corresponding to these boundaries are extracted and

fed to the ASR-pretrained speech encoder in section 3 for the task of disfluency detection and classification. Row (3) shows results from the recently proposed model by Jouaiti et al. [17]. This model uses phonological class posterior probabilities and the predicted phoneme sequence, extracted from the "Phonet" pretrained model [18], for an audio segment as features for a downstream classifier. Again, we use the MFA to extract the said audio segments.

The baseline results indicate a much lower performance compared to our E2E setup. One of the reasons that we notice is that forced alignment operations can be detrimental for children's speech whose quality is very different from fluent adult speech used to build the acoustic models used in MFA. However, to use a streamlined cascaded setup that extracts word boundaries prior to performing classification, using an off-the-shelf forced aligner like MFA is inevitable. Another reason that cascaded models do not work well is that some of the disfluency categories like "repeat" and "skip" need the spoken context to be identified which only a continuous stream of audio can provide.

The last 3 rows in table 1 show performances of our proposed models. The E2E models outperform the cascaded baselines by a significant margin. Row (4) represents the multi-task learning framework defined in section 3.2. We see a significant improvement in detection performance just by using L_{MTL} compared to the cascaded baselines. This shows that the attention based E2E model is successful in extracting the relevant speech segments corresponding to the words. We also see an improvement in the classification performance but it is not as much as the gains seen in detection. This can be attributed to the small size of the datasets which consequently leads to smaller number of examples in each class.

With MFW, we see gains in the performance on the two datasets and on both tasks as shown in row (5). However, using the proposed HiDeC model, we see the most significant improvements across the board. Classification improves as the information from the lower level detection module helps the classification module on top by filtering only the disfluent cases. Having the classification loss at the top level backpropagate through the lower network also helps detection.

Note that the classification performance of all models on the READR dataset is not very good with the best model reaching only 27.6% F1 score. This shows the difficulty of the task and leaves room for much improvement. Also, the lower results are consistent with the findings of Yang et al. [11] who encounter a similar issue of class imbalance for a similar task at the phone level. To tackle the possible acoustic mismatch between the Librispeech and READR datasets, we tried varying the pitch of female voices in the Librispeech corpus to mimic children voices and reduce the mismatch. Interestingly, it ended up hurting the performance. We note that this is because the readers are children with substantial reading difficulties and hence their speech is hard to process.

Ablation study: Table 2 shows an ablation study on the READR dataset to see the effect of each component on our proposed models. ASR pretraining proves to be an important step for the downstream task when we compare row (1) with the rest. Row (2) shows the results for the case when the ASR task is not performed along with detection/classification and we just use the loss, $L_{DCT/CLS}$. We see a lower performance compared to row (3) which adds the ASR loss. This shows that the ASR component serves as an effective co-training mechanism possibly assisting in acoustic model adaptation.

	READR		CMUK	
Model	Detection	Classification	Detection	Classification
(1) All correct	0.0	14.3	0.0	18.0
Previous work				
(2) Word segment based [5, 15]	13.2	16.3	19.7	22.7
(3) Phonological [17]	0.0	14.3	4.6	18.4
End-to-End (ours)				
(4) L_{MTL}	34.6	24.8	34.4	32.0
(5) + MFW	34.8	26.6	38.3	32.6
(6) L_{HiDeC}	**42.8**	**27.6**	**44.4**	**38.9**

Table 1. Results (in macro F1-score) on READR and CMUK datasets. Row (1): a trivial baseline which predicts "correct" every time. Rows (2), (3): the traditional cascaded baselines using forced alignment. Rows (4)-(6): our proposed methods.

	READR	
Model	Detection	Classification
No ASR pretraining		
(1) $L_{DCT/CLS}$	21.6	18.5
ASR pretraining		
(2) $L_{DCT/CLS}$	28.0	22.7
(3) + L_{ASR}	34.6	24.8

Table 2. Ablation studies on the READR dataset. Row (1): results without ASR pretraining. Rows (2), (3): with ASR pretraining, progressively adding the proposed techniques.

5. CONCLUSION

In this paper, we propose an end-to-end framework for word level disfluency classification in children's speech. We annotate two datasets for word level disfluency categories. To deal with the high degree of class imbalance in the two datasets, we propose a novel technique called HiDeC for hierarchical detection and classification which implicitly lowers the class imbalance by breaking the classification task into two hierarchical parts. We hope that this paper prompts future work towards collecting more children's speech datasets for building robust reading assessment models.

6. ACKNOWLEDGEMENT

This work was supported by the National Science Foundation under Grant No. 2008043.

7. REFERENCES

[1] Peter Plantinga and Eric Fosler-Lussier, "Towards real-time mispronunciation detection in kids' speech," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 690–696.

[2] Minglin Wu, Kun Li, Wai-Kim Leung, and Helen Meng, "Transformer based end-to-end mispronunciation detection and diagnosis," *Proc. Interspeech 2021*, pp. 3954–3958, 2021.

[3] Binghuai Lin and Liyuan Wang, "Phoneme mispronunciation detection by jointly learning to align," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6822–6826.

[4] Wenxuan Ye, Shaoguang Mao, Frank Soong, Wenshan Wu, Yan Xia, Jonathan Tien, and Zhiyong Wu, "An approach to mispronunciation detection and diagnosis with acoustic, phonetic and linguistic (apl) embeddings," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6827–6831.

[5] Matthew Black, Joseph Tepperman, Sungbok Lee, Patti Price, and Shrikanth S Narayanan, "Automatic detection and classification of disfluent reading miscues in young children's speech for the purpose of assessment," in *Eighth Annual Conference of the International Speech Communication Association*, 2007.

[6] Morris R Council III, Ralph Gardner III, Gwendolyn Cartledge, and Alana O Telesman, "Improving reading within an urban elementary school: computerized intervention and paraprofessional factors," *Preventing School Failure: Alternative Education for Children and Youth*, vol. 63, no. 2, pp. 162–174, 2019.

[7] M Eskenazi, J Mostow, and D Graff, "The cmu kids corpus," in *Linguistic Data Consortium, no. 11*. LDC, 1997.

[8] Jorge Proença, Carla Lopes, Michael Tjalve, Andreas Stolcke, Sara Candeias, and Fernando Perdigao, "Mispronunciation detection in children's reading of sentences," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 7, pp. 1207–1219, 2018.

[9] Jacques Duchateau, Yuk On Kong, Leen Cleuren, Lukas Latacz, Jan Roelens, Abdurrahman Samir, Kris Demuynck, Pol Ghesquière, Werner Verhelst, et al., "Developing a reading tutor: Design and evaluation of dedicated speech recognition and synthesis modules," *Speech Communication*, vol. 51, no. 10, pp. 985–994, 2009.

[10] William Chan, Navdeep Jaitly, Quoc V Le, and Oriol Vinyals, "Listen, attend and spell," *arXiv preprint arXiv:1508.01211*, 2015.

[11] Xuesong Yang, Anastassia Loukina, and Keelan Evanini, "Machine learning approaches to improving pronunciation error detection on an imbalanced corpus," in *2014 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2014, pp. 300–305.

[12] Han-Jia Ye, De-Chuan Zhan, and Wei-Lun Chao, "Procrustean training for imbalanced deep learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 92–102.

[13] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[14] Zoltán Tüske, George Saon, Kartik Audhkhasi, and Brian Kingsbury, "Single headed attention based sequence-to-sequence model for state-of-the-art results on switchboard," *arXiv preprint arXiv:2001.07263*, 2020.

[15] Colin Lea, Vikramjit Mitra, Aparna Joshi, Sachin Kajarekar, and Jeffrey P Bigham, "Sep-28k: A dataset for stuttering event detection from podcasts with people who stutter," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6798–6802.

[16] Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kaldi.," in *Interspeech*, 2017, vol. 2017, pp. 498–502.

[17] Melanie Jouaiti and Kerstin Dautenhahn, "Dysfluency classification in stuttered speech using deep learning for real-time applications," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6482–6486.

[18] Juan Camilo Vásquez-Correa, Philipp Klumpp, Juan Rafael Orozco-Arroyave, and Elmar Nöth, "Phonet: A tool based on gated recurrent neural networks to extract phonological posteriors from speech.," in *INTERSPEECH*, 2019, pp. 549–553.