2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU) | 979-8-3503-0689-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ASRU57964.2023.10389704

DETECTION OF VOWEL ERRORS IN CHILDREN'S SPEECH USING SYNTHETIC PHONETIC TRANSCRIPTS

Ilja Baumann[1], *Dominik Wagner*[1], *Korbinian Riedhammer*[1], *Elmar Nöth*[2], *Tobias Bocklet*[1,3]

[1]Technische Hochschule Nürnberg Georg Simon Ohm, Germany
[2]Friedrich-Alexander-Universität Erlangen-Nürnberg, Pattern Recognition Lab, Germany
[3]Intel Labs, Germany

ABSTRACT

The analysis of phonological processes is crucial in evaluating speech development disorders in children, but encounters challenges due to limited children audio data. This work focuses on automatic vowel error detection using a two-stage pipeline. The first stage uses a fine-tuned cross-lingual phone recognizer (wav2vec 2.0) to extract phone sequences from audio. The second stage employs a language model (BERT) for classification from a phone sequence, entirely trained on synthetic transcripts, to counteract the very broad range of potential mistakes. We evaluate the system on nonword audio recordings recited by preschool children from a speech development test. The results show that the classifier trained on synthetic data performs well, but its efficacy relies on the quality of the phone recognizer. The best classifier achieves an 94.7% F1 score when evaluated against phonetic ground truths, whereas the F1 score is 76.2% when using automatically recognized phone sequences.

Index Terms— children's speech, vowel errors, nonwords

1. INTRODUCTION

The evaluation of children's language is an important topic, especially with regard to the earliest possible detection of speech disorders as a prerequisite for adequate speech and language support. Developmental tests at preschool age are mandatory in Germany before entering elementary school to identify possible speech development disorders. The SETK (Sprachentwicklungstest für Kinder – Language Development Test for Children) [1] is one of the instruments used to assess the level of speech development in children in the age range of 3;0 – 5;11 (years;months). A subtest of the SETK, consisting of a nonword recitation test designed to evaluate the phonological working memory, is used in this work. Nonwords consist of regular syllables in a specific language, forming non-existent and new words. This way, the acquired language level can be tested, and auditory memory performance can be determined. This work focuses on German-speaking children. Several studies [2, 3] have verified the effectiveness of reciting nonwords in identifying children with language development disorders. Automatic detection was performed in [4] as a first step towards automatic speech evaluation. The authors employ a binary classification system to determine whether a nonword was recited correctly. A VGG-like [5] model was used for classification, trained on various feature types including speaker embeddings, wav2vec 2.0 (W2V2) utterance embeddings and senones of an acoustic model. In order to provide support for speech and language, it is necessary to conduct a detailed analysis of pronunciation errors. An analysis of phonological processes, for example, is an important method used by speech therapists.

1.1. Phonological processes

Phonological processes refer to sound patterns that young children use during language acquisition to simplify their speech. A phonological process involves the replacement of a group or sequence of speech sounds that share a common difficulty with an alternative group of lesser difficulty. The replacement occurs due to a lack of coordination of the vocal apparatus. Such processes are normal as children learn to talk and use their vocal apparatus. A phonological disorder occurs when phonological processes persist beyond an age at which typically developing children are unlikely to make these mistakes. This work focuses on the phonological process of vowel substitutions. In this case, a vowel is replaced by another inadequate vowel, e.g., in Banane (banana): [*bananə*] changes to [*bonanə*]. According to So & Dodd's concept of *phonological saliency* [6], shown also for German by Fox-Boyer [7], vowels are highly salient due to their low contrast spectrum to other elements and their unavoidable occurrence. The German language has numerous consonants that differ significantly in their articulation, e.g., in voicing (voiced vs. voiceless consonants) or in the place of articulation (labial, alveolar, etc.). These differences enable a clear distinction and differentiation of consonants from one another. In contrast, there are few vowels and their differences are less salient, these differ only in nuances such as the pitch, length, or the rounding of the lips. This results in a lower contrast between the vowels compared to the consonants. Although this concept should result in relatively few vowel errors, they are most common in our audio recordings.

1.2. Related work

Previous work mostly relies on multimodule classification, partly based on audio recordings. In [8], phonological processes are classified utilizing two modified AlexNet [9] systems using MFCC heat maps, one for correct/incorrect pronunciation and the other for the kind of phonological process found. The authors record a phonological processes adult dataset and pitch shift audios to imitate children's average pitch. The authors of [10] propose a two-module system for recognizing pronunciation and phonological processes. Both modules employ tree-based models. The first consists of a decision tree model for correct/incorrect pronunciation classification. The second consists of a random forest model, which utilizes a matrix provided by an expert to determine the probabilities of various phonological processes in each phoneme of a given word set. Text-based data augmentation in form of additional random sentence generation was used in [11] to detect stuttering events in orthographic transcripts of children's speech.

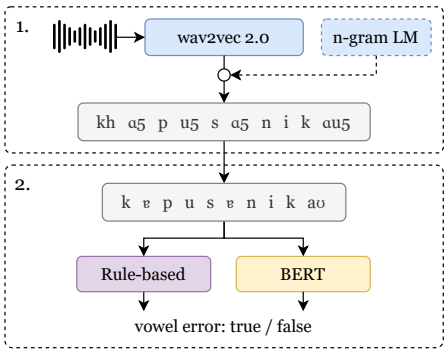

Fig. 1. Schematic diagram of the two-stage vowel error detection system comprised of W2V2 phone recognition and rule-based or phone-level BERT classification.

1.3. Motivation and contribution

Due to the large number of potential phonological processes, it is difficult to collect audio recordings from several children and all the processes. Previous works rely on a large amount of data and additional expert feedback [10] or perform augmentation through pitch shifting at the audio level [8]. Our approach also addresses augmentation to counteract the lack of children data at the textual level. We rely on synthetically generated phone sequences to replicate the very broad range of potential mistakes and test our system with vowel errors. To classify vowel errors in nonwords, we build a two-stage system, illustrated in Figure 1. We first predict a phone sequence using audio recordings of children as input to a W2V2 fine-tuned phone recognizer with an optional n-gram language model. The second stage employs different text-based approaches, including rule-based Phonex matching implemented in Phon [12] and pre-trained phone-level BERT (Bidirectional Encoder Representations from Transformers) [13] system fine-tuned for classification. Since audio recordings from children are only available to a very limited extent, we solely use synthetically generated and out-of-domain phone sequences based on prior knowledge to train the classification models. We use articulatory features to generate the synthetic corpora. Therefore, all audio recordings in our corpus remain available for testing. Finally, we use Integrated Gradients [14] to verify the vowel attribution of the trained classifier.

Our main contributions are (1) fine-tuning a pre-trained cross-lingual phone recognizer (W2V2) for German-speaking children, (2) synthetic vowel error training transcripts generation, (3) language model (BERT) pre-trained and fine-tuned on synthetic phone-level sequences for classification.

2. DATA

We use the SETK corpus as a test set. For fine-tuning the phone recognizer of the first stage, we use three corpora containing audio recordings of children: Erlangen-CLP, Fox-Boyer, and kidsTALC. In the second stage, we use the Wikipedia text corpus for the pre-training of the language models. The respective corpora are briefly described in the following sections.

2.1. Test corpus

We use SETK data collected from the authors of [15], a corpus of 140 children $6;2 \pm 2;11$ years old, containing recordings of seven nonwords of the SETK subtest for phonological working memory and build upon the phonetically transcribed audio in [4] as test data for vowel error detection. The SETK test consists of four subtests for three-year-olds and five subtests for four- to five-year-olds. Three-year-olds are asked to verbally encode concretely presented events, while four- to five-year-olds have to incorporate abstract rule information into the task. The subtests can be mainly grouped into language development, language comprehension, language production and language memory. The SETK subtest used in this work consists of two two-syllable nonwords, one three-syllable nonword, and two four- and five-syllable nonwords. The children recited the nonwords in a screening scenario conducted by a speech therapist, starting with two-syllable nonwords and increasing the complexity to five-syllable nonwords. The seven nonwords are *Maluk, Bilop, Glösterkeit, Ronterklabe, Seregropist, Pristobierichkeit* and *Kabusaniker*. The same recording, pronounced by a speech therapist, was played to the children for each nonword to ensure matching conditions. The recordings were collected in various German Kindergartens on-site. Each nonword is orthographically and phonetically transcribed and labeled whether it contains a vowel error. The complete dataset is only used for testing in order to have as much test data available as possible. Table 1 lists further details of the corpus.

2.2. Fine-tuning audio corpora

In addition to the SETK dataset, we use three children corpora for fine-tuning the W2V2-based phone recognizer: Erlangen-CLP control, Fox-Boyer and kidsTALC. These are briefly described in the following sections, details for each corpus are listed in Table 1.

2.2.1. Erlangen-CLP

The Erlangen-CLP (ER-CLP) corpus [16] contains children with Cleft Lip and Palate (CLP) and age-matched control speakers recorded using the PLAKSS (Psycholinguistische Analyse kindlicher Sprechstörungen - Psycholinguistic analysis of children's speech disorders) picture-naming task. PLAKSS consists of 99 distinct word stimuli encompassing all consonants, vowels, and consonant clusters used in German phonotactics. This task also incorporates words of various lengths and stress patterns.

We only use the control set in this work, which contains 598 children in the age range 3;3 to 18;10. Children's utterances are orthographically transcribed. Therefore, we use eSpeak[1] to generate phonetic International Phonetic Alphabet (IPA) transcriptions of the 99 stimuli. eSpeak is a text-to-speech synthesizer which also offers the conversion from text to phonemes.

2.2.2. Fox-Boyer

The Fox-Boyer (FB) corpus from PhonBank [17] contains 32 typically developing children's audio recordings in the age range 2;3 to 9;2 years collected from the PLAKSS picture-naming task. Audio recordings were collected in various kindergartens located in the northeast region of Germany and private practices in the country's western and northern regions. We use the provided phonetic transcription of the 99 stimuli recordings.

2.2.3. kidsTALC

The kidsTALC corpus [18] contains spontaneous speech of 46 German children aged 3;6 to 10;11 years. The recordings were collected in an examiner-child interaction conducted by speech therapists or

[1] https://github.com/espeak-ng/espeak-ng

speech-language therapy students. The basis for the recordings are seven different wordless picture books suitable for children of different age groups. The average children's utterance length is 15.6 min. The utterances are accompanied by both orthographic and phonetic transcripts. We use the provided phonetic transcription in this work.

2.3. Text corpus

We incorporate a text corpus that has been transformed from graphemes to phonemes in both stages of our pipeline. In the initial stage of audio-based phone recognition, we utilize the converted text corpus to train n-gram language models, enhancing the accuracy of phone recognition. In the subsequent stage, which involves classifying vowel errors, we employ the converted text corpus to pre-train the language model used for classification. The German Wikipedia text corpus serves as our data source, which consists of article texts in German and is utilized along with synthetically generated text data in our work. We use the 2022 version of the corpus[2], randomly select 1 million lines and convert the graphemes to phones of the IPA using eSpeak.

Table 1. Audio corpora details used in our work for fine-tuning the phone recognizer and pre-training the language model based vowel error classifier.

Corpus	Age	Speaker	Hours	Content
ER-CLP	$8; 8 \pm 13; 4$	598	39.5	PLAKSS
Fox-Boyer	$5; 1 \pm 4; 1$	32	0.7	PLAKSS
kidsTALC	$6; 6 \pm 4; 4$	46	11.4	Spontaneous speech
SETK	$6; 2 \pm 2; 11$	140	0.5	Nonwords

3. METHOD

The proposed classification system operates in two stages, illustrated in Figure 1. In the first stage, we employ a cross-lingual phone recognizer that is fine-tuned using German children audio corpora (refer to Section 2.2). In the second stage, a classification module utilizes a language model to identify vowel errors. To ensure compatibility between the outputs of the phone recognizer in the first stage, which can produce any character from the IPA, and our desired lexicon containing only German IPA characters, we utilize articulatory features to map the phone recognizer outputs to the target lexicon. In the second stage, we leverage articulatory features to generate synthetic text-based phone sequences, which serve as training inputs for the classification module. Specifically, we train a phone-level language model to identify vowel errors within phone sequences obtained from transcriptions of actual audio recordings. We also compare this approach to a rule-based detection method.

The subsequent chapters provide detailed descriptions of each of these methods.

3.1. Self-supervised phone recognition

In the initial stage of our approach, we employ the cross-lingual W2V2 phone recognizer introduced in [19]. This model is based on XLSR-53 [20], which is pre-trained on speech audio sampled at 16kHz from 53 languages. It consists of 24 transformer blocks and 16 attention heads. The model is fine-tuned on CommonVoice [21] to predict cross-lingual phonetic labels using a classifier that represents the phonetic vocabulary. The vocabulary size is 392. For the fine-tuning process, we utilize the Connectionist Temporal Classification (CTC) loss [22]. The model is fine-tuned using the Adam optimizer [23] with an epsilon value of 10^{-7}. The learning rate is set to 2×10^{-4}, and a batch size of 64 is used for five epochs. We employ different corpora, which are described in Section 4.1.

We further build {2-4}-gram language models (LM) based on different data sources. On the one hand, we use the synthetically generated phone sequences to build language models. On the other hand, we use the grapheme-to-phone converted German Wikipedia corpus described in Section 2.3. Finally, we combine the synthetically generated phone sequences with the Wikipedia phone corpus. In decoding, we use KenLM [24], with a language model weight $\alpha = 0.5$ and the compensation term $\beta = 1.5$, which have experimentally proven to be the best values.

Training on various languages can capture more information, which is especially important for the developing child's language, where unpredictable sounds may be produced. For example, we have found that interdentality is revealed by the replacement of the phones [s] and [z] in German by the phones [θ] and [ð], which are not part of the German phone inventory. By using a cross-lingual phone recognizer, which is trained on multiple languages, we can capture the diverse range of vowel sounds present in different languages. This provides quantitative measures for future analysis through, e.g., the assignment of phonetic similarity scores or acoustic distances between the target and produced vowels.

3.2. Articulatory features

We adopt phones as the modeling units, utilizing the symbols available in the widely used IPA standard. As the phone recognizer has the capability to predict any IPA character, we follow the approach described in [19] and map these predictions to the specific German phone set used in our test inventory. The target lexicon consists of 79 characters representing standard German language pronunciations. For this purpose, we use articulatory features that map each phone or sound by global attributes. To do this, we use PanPhon [25], which relates IPA segments to their definitions in 21 articulatory features (positive or negative). These can be grouped into six classes widely used by phonologists: major class (syllable, sonorant, consonantal, continuant), laryngeal (voice, spread glottis, constricted, glottis), major place (anterior, coronal, labial, distributed), minor place (high, low, back), manner (nasal, lateral, delayed release, strident), and minor manner (round, tense). Figure 2 illustrates the mapping process. The first step is the phone relation to articulatory vectors, followed by conversion to numeric vectors. We use the lowest Hamming distance between the source (phone recognizer) and target vocabulary numeric features to obtain the final mapping.

3.3. Synthetic pronunciations generation

Recordings of children are only available to a very limited extent. The number of available samples shrinks further by focusing on a specific phonological process since a single process, is contained only in a fraction of the utterances. Therefore, we generate synthetic training transcriptions (phone sequences) using prior knowledge and articulatory features in this work.

We first convert the orthographic ground truths of the nonwords to phones and operate further on the phone sequences. We identify in which position a vowel error may occur in the seven nonwords. At the identified locations, we calculate the Hamming distances be-

[2]https://github.com/GermanT5/wikipedia2corpus

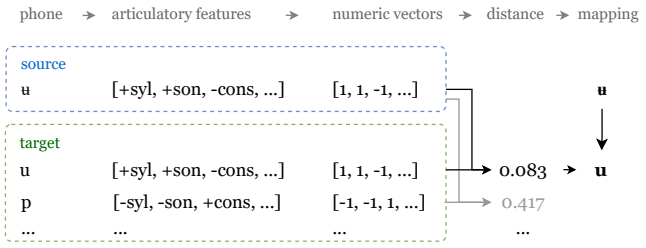

Fig. 2. Phone mapping from source vocabulary of the phone recognizer to the target vocabulary based on articulatory features.

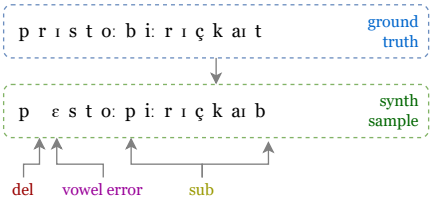

Fig. 3. Vowel error synthetic (synth) sample with possible augmentations for the nonword *Pristobierichkeit*.

tween the respective vowels in the nonword and the remaining vowels in the test inventory across numeric feature vectors from PanPhon. Based on the distances, we determine the number of examples n to generate using the Equation 1, where s is the number of syllables of the respective nonword and d is the distance between the articulatory features. We generate more examples on the one hand for phones that are articulatory close, and on the other hand by the length of the nonwords. The more syllables a nonword consists of, the more examples we generated.

$$n = 10^s ln(\frac{1}{d} + 1) \tag{1}$$

Based on the number of synthetic vowel error sequences, we upsample the training dataset with nonword sequences that don't contain a vowel error by 97%. Thereby, we map previous findings, according to which around 3% vowel error rate of children is observed in younger child groups in German [7, p. 68].

To introduce further variation and improve the generalization ability of our classifiers, we randomly augment the generated sequences by substitutions with nearby phones, random phones, and deletions on the training data. The intention of substituting close phones is to model other possible phonological processes. One augmentation is applied to two-syllable words, while two to three augmentations are applied to the remaining words. We randomly deleted 10% of the phones in the generated transcriptions. According to Fox-Boyer [7, p. 69], this value was chosen since children still perform 10-20% of consonant and consonant cluster extinction up to age 3;11. Nearby substitutions are calculated in the same way as for synthetic vowel errors. Random phone substitutions are applied at the end with a randomly chosen probability of 20%. All augmentations are applied to all phones except the vowels to not generate wrong labels. Figure 3 shows a synthetic sample generated for the nonword *Pristobierichkeit*, augmented with a deletion and two substitutions.

3.4. Vowel error detection

The second stage of our approach is to detect vowel errors exclusively based on the phonetic transcriptions of the nonword audio recordings. For this purpose, we rely on two methods, first rule-based detection and second classification, using a language model pre-trained on phone sequences and fine-tuned on synthetically generated pronunciation phone sequences.

3.4.1. Rule-based detection

Phon [12] is a software program initially designed to assist in investigating children's language development and facilitates the creation of textual and phonological data collections. The software includes Phonex, a pattern-matching language used for IPA transcriptions, which enables searching transcripts for phone sequences according to segmental and prosodic criteria. We use Phon to indicate vowel errors based on the built-in phonological processes module. We transfer the test dataset and the phone recognizer predictions to the dataset structure used in Phon and analyze these using the provided functions in Phon[3].

3.4.2. Phone-level BERT

Models like BERT [13] have achieved remarkable performance on various natural language processing and downstream tasks by fine-tuning on domain specific data. Given our ability to generate a substantial amount of synthetic data, we pre-train and subsequently fine-tune BERT for sequence classification. Considering that our focus is on vowel error detection, we exclusively utilize phones as the input. In a self-supervised manner, we undertake pre-training of a BERT model at the phone level, following a similar approach as the original BERT. For our specific case, we opt for a linguistically motivated pre-training objective known as masked language modeling (MLM), which in our context is referred to as masked phone token prediction. To accommodate our needs, we employ a modified BERT architecture featuring three attention heads, three transformer layers, and a hidden size of 384. Our vocabulary size is set at 199 to encompass the relevant phone tokens and most common phone combinations. The resulting model has approximately 3.3M parameters. We use the pre-trained BERT model with a classification head for fine-tuning to the vowel error binary classification task. The classification head consists of a single linear layer, a hidden size of 384 and an output size of 2. Cross-Entropy was used as the loss function.

4. EXPERIMENTS

The first stage in our process involves the phone sequence prediction from audio using a W2V2 cross-lingual phone recognizer on the full SETK corpus. We evaluate the performance of the phone recognizer by calculating the phone error rate (PER). In the next stage, we assess the classification performance using ground truth phone sequences to determine how well the classifiers perform when applied to perfectly predicted phone sequences. To accomplish this, we utilize the phonetic transcription from the SETK dataset as the input sequence for classification. Finally, we convert the predicted phone sequence from the phone recognizer into the test language's vocabulary and perform classification using both rule-based and BERT-based approaches. We calculate the macro-averaged F1 scores across the seven nonwords to evaluate the performance of the classifiers.

[3]https://github.com/phon-ca/phon

4.1. Phone recognition

We perform fine-tuning on the pre-trained cross-lingual model, as described in Section 3.1, using additional children only corpora outlined in Section 2. Each fine-tuning process lasts for five epochs. We start with the ER-CLP control data, followed by the inclusion of the Fox-Boyer corpus, and finally, we fine-tune the model using all three corpora jointly. The results of the fine-tuning are presented in Table 2. We evaluate the performance of the phone recognizer on the audio recordings of nonwords from the SETK test. The baseline result, without any fine-tuning, exhibits a high PER of 56.6% for the non-fine-tuned model. However, through fine-tuning with children's utterances we are able to reduce the PER by almost 36% relative to 36.4%. Especially, adding the kidsTALC corpus lowers the PER noticeably. Although the impact of adding the Fox-Boyer corpus is not as pronounced as adding kidsTALC, its inclusion still enhances the performance. This is likely due to the diverse acoustic conditions of the recordings, despite the dataset's small size of 32 speakers and less than an hour of recordings.

Table 2. Cross-lingual W2V2 fine-tuning PER results on the SETK nonwords dataset and additional n-gram phone language models built upon the best fine-tuned model with all three corpora.

System	PER
W2V2 without fine-tuning	56.6
ER-CLP control	47.5
+ Fox-Boyer	43.9
+ kidsTALC	36.4
Wikipedia phonetic corpus	
+ 2-gram LM	35.7
+ 3-gram LM	36.9
+ 4-gram LM	42.5
Synthetic phonetic corpora	
+ 2-gram LM	33.2
+ 3-gram LM	**30.7**
+ 4-gram LM	35.9
Synthetic + Wikipedia phonetic corpus	
+ 2-gram LM	34.8
+ 3-gram LM	35.2
+ 4-gram LM	38.3

To address the persistently high PER observed even after fine-tuning with children's data, we employ probabilistic n-gram language models during the decoding process to enhance the recognition. Especially for the nonword *Ronterklabe*, a wrong prediction of the vowel [ɔ] as [ʊ] often occurred, which is critical for the vowel error classification. The results in Table 2 show that decoding with a 3-gram language model built on synthetic data performs best with an 30.7% PER and achieves a 15.7% relative improvement compared to the absence of an n-gram language model.

4.2. Rule-based vowel error detection

We employ Phon [12] for a simple rule-based vowel error detection using phone sequences. Phon's phonological process analysis compares the actual phonetic transcription to the target word's phonetic transcription. This includes the automatic syllabification and alignment of the predicted and the grapheme-to-phone converted ground truth word, which is the reference for comparison. The rule-based detection results in an F1 score of 87.7% on the phonetic ground truth and 71.1% on phone recognizer predicted sequences. All classification results are listed in Table 3.

4.3. BERT vowel error detection

To overcome limitations imposed by available training data, we take advantage of our ability to generate data and train a BERT language model. We explore two datasets for pre-training. Firstly, we utilize synthetically generated transcriptions for pre-training the BERT model for 40 epochs and subsequently fine-tune it for the classification task outlined in Section 3.3. In addition, we leverage the Wikipedia sub-corpus detailed in Section 2.3 for pre-training the BERT model. Secondly, we fine-tune the pre-trained BERT model using the synthetically generated data. By incorporating both of these datasets, we aim to enhance the performance and robustness of the BERT language model for our classification task.

The results in Table 3 show that by using synthetic data for pre-training (PT) and fine-tuning (FT), the performance of 72.0% F1 score is lower than that of the rule-based system. With a pre-training on Wikipedia data, an F1 score of 94.7% can be achieved and thus an increase of 8.0% relative compared to the rule-based approach. To use synthetic and Wikipedia data jointly, we upsample the synthetic samples to approximately match the 1M lines of the Wikipedia text corpus. The combination of synthetic and Wikipedia data in pre-training differs only in two samples, therefore the results are not statistically significant.

Table 3. Vowel error classification using the rule-based approach, compared to different pre-trained (PT) and fine-tuned (FT) BERT classifiers. Results are listed as F1 scores with standard deviation (SD).

Classifier	Phonetic ground truth	Predicted sequences
Rule-based	87.7 (7.5)	71.1 (9.5)
BERT		
PT: Synth, FT: Synth	72.0 (9.5)	66.9 (12.6)
PT: Wiki, FT: Synth	94.7 (5.6)	76.2 (7.3)
PT: Synth + Wiki, FT: Synth	94.5 (5.9)	76.0 (15.3)

4.4. End-to-end vowel error detection

For an end-to-end solution, we combine the fine-tuned phone recognizer with the vowel error classification module. Using built-in PanPhon functionality, we compute feature vectors and compare each pair of phones between the phone recognizer's vocabulary and the German IPA target inventory. We map each phone of the source vocabulary to a phone in the target inventory using the lowest Hamming distance between the numeric feature vectors, where multiple phones can map to a single target phone. An illustration of the process is shown in Figure 2. The resulting mapping is applied to the predicted phone recognizer sequences. By using the fine-tuned recognizer and an n-gram LM, mapping is rarely necessary due to the fact that fine-tuning recordings and n-gram LM data is only in German. We still perform this step to ensure that all predicted phones are in the target inventory.

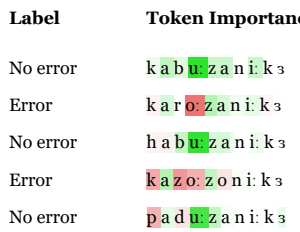

Fig. 4. Token attribution examples of the nonword *Kabusaniker* obtained using IG. A red background contributes to the classification as vowel errors, and a green background contributes to the class not containing vowel errors.

The results on the recognized sequences from audio in Table 3 show that there is a substantial drop in F1 score, compared to the prediction on ground truth sequences. Using the rule-based approach, an F1 score of 71.1% is achieved, which is a relative decline of 18.9% compared to the classification on ground truth. Using the BERT classifier yields an F1 score of 76.2%, a relative drop of 19.5%. This shows that the phone recognizer produces mistakes in substantially important phone locations of vowels.

4.5. Classification verification

Integrated Gradients (IG) [14] models human perception, capturing the significance of features in sensory stimuli. It considers variations in features and assigns importance scores based on their impact on the model's output, mirroring how our brains naturally prioritize salient features in our environment.

In our approach, we utilize IG from Captum [26] to assess the emphasis on vowel errors in phone sequences. This technique assigns importance scores to individual features of our classification models. IG calculates the gradient of the output with respect to the input by traversing a straight-line path from a baseline input (represented by a zero information padding token "PAD") to the actual input. By integrating the gradients along this path, it quantifies the contribution of each feature to the model's output. The analysis shows that the best-performing model focuses mainly on the vowels, i.e., they contribute the most to the classification. Figure 4 shows an excerpt of the influences of each token in the classification of vowel errors in the nonword *Kabusaniker*. Phones highlighted in red contribute to the classification as vowel errors, highlighted in green to the class not containing vowel errors.

5. DISCUSSION

Our work aimed to address the challenges associated with analyzing phonological processes in children with speech development disorders, under limited availability of children audio data. To overcome these limitations, a two-stage pipeline was developed for automatic vowel error detection. The first stage employs a fine-tuned cross-lingual phone recognizer, specifically the W2V2 model, to extract phone sequences from the audio recordings. In the second stage, a language model based on BERT was employed for classification using the phone sequences. An important aspect is that the language model was trained exclusively on synthetic transcripts. This allows the modelling of a wide range of potential mistakes encountered in children's speech.

The evaluation of the proposed system was conducted on nonword audio recordings recited by preschool children as part of a speech development test. The results demonstrated that the classifier trained on synthetic transcripts performed well in identifying vowel errors. However, the effectiveness of the classifier heavily relies on the quality of the phone recognizer employed in the first stage of the system. This finding emphasizes the significance of accurate phone sequence extraction for vowel error detection. Inaccurate phone recognition cascades errors throughout the system, resulting in incorrect phonological error classification.

Previous works have achieved an F1 score of 98.4% in [8] on the classification task of phonological processes from audio of adults, pitch shifted to three average pitches of children in the age range from 3 to 5. In [10] an accuracy of 94.6% was achieved. However, only the classification whether a word was pronounced correctly or incorrectly is performed on the audio. The classification itself is based on expert knowledge and preliminary information about which phonological processes can occur in a word.

Our results in the classification of vowel errors are in a similar range of 94.7% F1 score, considering a correctly predicted phone sequence from audio. However, the results are difficult to compare, since we perform the classification on a textual level and data-driven, which, for example, with the analysis using integrated gradients, gives us more precise information about where the errors occur in the respective nonword. After all, the phone recognizer produces errors at the critical locations of the vowels, which are important for the classification, and therefore the error is propagated further to the classifier. This results in an F1 score of 76.2%. The Wikipedia pretraining of the classification module adds more value, most likely due to many correct samples from real texts.

We have shown that prior knowledge from phonological acquisition can be used to generate synthetic data for training, which is well suited for the detection of phonological errors. The utilization of synthetic data for training the language model proved to be beneficial, enabling the model to handle a broad range of potential mistakes encountered in children's speech development. This concept can be easily applied to other phonological processes, extending the amount of textual child data available. The generated transcriptions can further be used for augmentations, such as the use of text-to-speech systems to directly generate audio samples of nonwords.

6. CONCLUSIONS

We addressed the challenges associated with analyzing phonological processes in children with speech development disorders by introducing a two-stage system for automatic vowel error detection. The pipeline utilized a fine-tuned cross-lingual phone recognizer (wav2vec 2.0) to extract phone sequences from audio recordings, followed by a language model (BERT) trained on synthetic transcripts for classification. We examined two approaches for classifying vowel errors based on phonetic transcriptions: a rule-based approach and utilizing a BERT language model. Additionally, we explored the effectiveness of using synthetically generated phone sequences to compensate for the lack of child-specific textual data. When fine-tuned with synthetic phone sequences, a BERT-based classifier achieved an F1 score of 94.7% under the assumption of a perfect phone recognizer. However, when utilizing the output of a fine-tuned wav2vec 2.0 phone recognizer on children's recordings, the F1 score for vowel error classification was 76.2%. Future research will concentrate on leveraging the insights gained from this work to analyze and detect other phonological processes and explore further augmentation techniques for children's speech assessment.

7. REFERENCES

[1] Hannelore Grimm, M Aktas, and S Frevert, "Setk 3–5: Sprachentwicklungstest für 3-bis 5-jährige kinder," *Hogrefe: Göttingen, Germany*, 2001.

[2] Susan E Gathercole and Alan D Baddeley, "Phonological memory deficits in language disordered children: Is there a causal connection?," *Journal of memory and language*, vol. 29, no. 3, pp. 336–360, 1990.

[3] James W Montgomery, "Sentence comprehension in children with specific language impairment: The role of phonological working memory," *Journal of Speech, Language, and Hearing Research*, vol. 38, no. 1, pp. 187–199, 1995.

[4] Ilja Baumann, Dominik Wagner, Sebastian Bayerl, and Tobias Bocklet, "Nonwords Pronunciation Classification in Language Development Tests for Preschool Children," in *Proc. Interspeech 2022*, 2022, pp. 3643–3647.

[5] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2015.

[6] Lydia KH So and Barbara J Dodd, "The acquisition of phonology by cantonese-speaking children," *Journal of child language*, vol. 22, no. 3, pp. 473–495, 1995.

[7] Annette V Fox-Boyer, Inula Groos, and Kerstin Schauß-Golecki, "Kindliche aussprachestörungen," *Phonologischer Erwerb–Differenzialdiagnostik–Therapie*, vol. 7, 2016.

[8] Braulio Baldeon, Renzo Ravelli, and Willy Ugarte, "Wawasimi: Classification techniques for phonological processes identification in children from 3 to 5 years old.," in *CSEDU (2)*, 2022, pp. 147–154.

[9] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, Eds. 2012, vol. 25, Curran Associates, Inc.

[10] Maria Helena Franciscatto, Marcos Didonet Del Fabro, João Carlos Damasceno Lima, Celio Trois, Augusto Moro, Vinícius Maran, and Marcia Keske-Soares, "Towards a speech therapy support system based on phonological processes early detection," *Computer Speech & Language*, vol. 65, pp. 101130, 2021.

[11] Sadeen Alharbi, Madina Hasan, Anthony J. H. Simons, Shelagh Brumfitt, and Phil Green, "Detecting stuttering events in transcripts of children's speech," in *Statistical Language and Speech Processing*, Nathalie Camelin, Yannick Estève, and Carlos Martín-Vide, Eds., Cham, 2017, pp. 217–228, Springer International Publishing.

[12] Yvan Rose, Brian MacWhinney, Rodrigue Byrne, Gregory Hedlund, Keith Maddocks, Philip O'Brien, and Todd Wareham, "Introducing phon: A software solution for the study of phonological acquisition," in *Proceedings of the... Annual Boston University Conference on Language Development. Boston University Conference on Language Development*. NIH Public Access, 2006, vol. 2006, p. 489.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[14] Mukund Sundararajan, Ankur Taly, and Qiqi Yan, "Axiomatic attribution for deep networks," in *International conference on machine learning*. PMLR, 2017, pp. 3319–3328.

[15] Tobias Bocklet, Cordula Winterholler, Andreas" Magnet" Maier, Maria Schuster, and Elmar Nöth, "An automatic screening test for preschool children: theory and data collection," in *Proceedings of the 2nd Workshop on Child, Computer and Interaction*, 2009, pp. 1–4.

[16] T. Bocklet, A. Maier, K. Riedhammer, U. Eysholdt, and E. Nöth, "Erlangen-CLP: A large annotated corpus of speech from children with cleft lip and palate," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, May 2014, pp. 2671–2674, European Language Resources Association (ELRA).

[17] Yvan Rose and Brian MacWhinney, "The PhonBank Project: Data and Software-Assisted Methods for the Study of Phonology and Phonological Development," in *The Oxford Handbook of Corpus Phonology*. Oxford University Press, 05 2014.

[18] Lars Rumberg, Christopher Gebauer, Hanna Ehlert, Maren Wallbaum, Lena Bornholt, Jörn Ostermann, and Ulrike Lüdtke, "kidstalc: A corpus of 3- to 11-year-old german children's connected natural speech," in *Proceedings INTERSPEECH 2022 – 23rd Annual Conference of the International Speech Communication Association*, Sept. 2022.

[19] Qiantong Xu, Alexei Baevski, and Michael Auli, "Simple and Effective Zero-shot Cross-lingual Phoneme Recognition," in *Proc. Interspeech 2022*, 2022, pp. 2113–2117.

[20] Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli, "Unsupervised Cross-Lingual Representation Learning for Speech Recognition," in *Proc. Interspeech 2021*, 2021, pp. 2426–2430.

[21] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," in *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, 2020, pp. 4211–4215.

[22] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd International Conference on Machine Learning*, New York, NY, USA, 2006, ICML '06, p. 369–376, Association for Computing Machinery.

[23] Diederik Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *International Conference on Learning Representations*, 12 2014.

[24] Kenneth Heafield, "KenLM: Faster and smaller language model queries," in *Proceedings of the Sixth Workshop on Statistical Machine Translation*, Edinburgh, Scotland, July 2011, pp. 187–197, Association for Computational Linguistics.

[25] David R. Mortensen, Patrick Littell, Akash Bharadwaj, Kartik Goyal, Chris Dyer, and Lori S. Levin, "Panphon: A resource for mapping IPA segments to articulatory feature vectors," in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. 2016, pp. 3475–3484, ACL.

[26] Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and Orion

Reblitz-Richardson, “Captum: A unified and generic model interpretability library for pytorch,” *ArXiv*, vol. abs/2009.07896, 2020.