PRONUNCIATION MODELING FOR CHILDREN'S SPEECH

Christopher Gebauer[*], *Lars Rumberg*[*], *Jörn Ostermann*
Leibniz Universität Hannover, Institut für Informationsverarbeitung
{gebauer, rumberg}@tnt.uni-hannover.de

Abstract: The accuracy of automatic speech processing systems for children's speech lags heavily behind the accuracy of systems for adult speech. One of the reasons is a high pronunciation variability in children's speech. Modeling this variability can be effective to increase performance. We investigate whether MAUS, a system developed for phonemic segmentation and trained on adult speech, which explicitly models deviations from canonical pronunciations, can be applied to children's speech. We compare it to a recently presented system trained on children's speech. We evaluate whether the systems can capture pronunciation variability as well as the performance on phonemic segmentation.

1 Introduction

As all machine learning applications, automatic speech processing systems depend on similarity of the processed data to the data used during training. When applying speech processing systems designed for adult speech to children's speech, the dissimilarity is often too high, leading to high error rates [1]. One of the properties of children's speech, making it difficult to apply systems designed for adult speech, is the high pronunciation variability. Explicitly considering pronunciation variability might therefore increase performance of systems trained on adult speech. MAUS [2], a system developed for phonemic segmentation of adult speech, explicitly models variations for a given canonical pronunciation. We investigate whether MAUS can be successfully applied to model pronunciation variability of children's speech. We compare it with a recently proposed approach especially trained for children's speech [3].

Both approaches use an available canonical pronunciation of the processed utterance and an acoustic model as well as a pronunciation model to capture deviations of that canonical pronunciation. While MAUS is based on hidden Markov models (HMMs), the system of [3] relies on a deep learning acoustic model trained end-to-end using the connectionist temporal classification (CTC) criterion [4]. The comparison is done on the kidsTALC corpus [5], a recently published corpus of typically developing, monolingual German speaking children. We evaluate the performance in capturing the pronunciation variability as well as on phone-level segmentation.

2 Method

In this section we will describe the systems compared in our work. The main goal is to investigate, whether the default German acoustic and pronunciation model of MAUS [2] is suitable to correct deviations of a canonical pronunciation. We will compare MAUS with a CTC-based system proposed in [3] that is trained on children's speech and models deviation given a canonical pronunciation in a similar way as MAUS.

[*]Contributed equally

2.1 MAUS

MAUS [2] targets phonemic segmentation and labeling using a data-driven Markov process. Given an audio signal x, a label l, and the corresponding sequence of label s, the conditioned probability can be rewritten as follows:

$$P(l|x) \propto P(x|l)P(l). \tag{1}$$

The acoustic model $P(x|l)$ is determined using a HMM, whereas the key contribution of MAUS is the modeling of the language model $P(l)$. As the system focuses on segmentation and not automatic speech recognition (ASR), it is assumed that an approximated label sequence s' for the input signal based on a canonical pronunciation is known. Usually this is obtained by an orthographic transcription coupled with a pronunciation dictionary, which usually leads to a slight deviation between the two sequences s and s'.

MAUS represents the approximated label sequence by a linear weighted finite state transducer (WFST), which is further extended by data-driven patterns, building the full decoding graph $\mathscr{D}$. This extension allows to recover the target sequence s from the decoding graph $\mathscr{D}$, if the extension patterns are chosen correctly. Each of these patterns consists of a left and right context (c_l, c_r), and the original label as well as the replacement $(l_a, l_b) \in \mathscr{T}$. To allow for insertions and deletions the label set $\mathscr{T}$ is extended by an empty label. The patterns are constructed by aligning manual target sequences s from a given dataset with the approximated sequence s' based on a canonical pronunciation from a dictionary and the manual orthographic transcription. To further increase the available extension patterns, the left and right context is nullified, i. e., assuming the replacement pattern given (c_l, c_r, l_a, l_b) is also common when either the left or the right context is omitted.

While the state probabilities in the Markov process are directly given from the acoustic model, the transition probabilities need to be weighted based on the resulting decoding graph $\mathscr{D}$. This is necessary to ensure that each decoding path is assigned the correct probability depending on the contained extension patterns. MAUS has a weighted and an unweighted method to compute the resulting transition probabilities. For simplicity we will describe the unweighted method, i. e., assuming that all paths have an equal probability, and refer for the weighted approach to Kipp [6]. The assumption of equal weights for all paths in the decoding graph simplifies the computation of the transition probabilities to a path counting problem. MAUS computes in a forward pass the sum N_i of all paths leading to Node i. Setting the probability of the final Node of being part in the most probable decoding path to one, the probability of all other Nodes is computed in a backward pass

$$P_i = \sum_j P_j \frac{N_i}{N_j}, \tag{2}$$

where $\sum_j$ represents the sum over all subsequent nodes. Combining the results of the forward and backward pass leads to the transition probability of two adjacent nodes, given by

$$P_{i \to j} = \frac{P_j N_i}{P_i N_j} \quad \text{with} \quad i < j. \tag{3}$$

2.2 CTC Decoding

We compare the results of MAUS to those of a recent approach constraining the decoding of an end-to-end speech recognition system [3]. Its general idea is comparable to MAUS. Using canonical pronunciations from a pronunciation dictionary and data-driven pronunciation rules,

a graph is build, which describes different possible realizations of an orthographic transcript. However, instead of using a traditional Hidden Markov model based system, it uses an end-to-end deep neural network acoustic model trained with the CTC criterion [4] in an end-to-end fashion.

CTC is a method for training of sequence labeling models without need for a known alignment between input and output sequences. For each (time-)frame in the input sequence, CTC models compute a probability distribution over the label set $\mathscr{T}$ extended by a special blank token. The blank token allows for repetitions in the finale output sequence, i. e., the output sequence is given by selecting one token at each frame, then collapsing consecutive identical tokens and finally removing the blank token. Each uncollapsed path through the output defines one alignment between the input and output sequences. The probability of a path is computed by multiplying the probabilities of the selected tokens at each frame. Summing up the probability of all paths which lead to the same output sequence after collapsing, results in the probability of that output sequence. CTC thus defines the probability of an output sequence by the sum of the probabilities of all possible alignments between input and output sequences. Since the derivative of the probability of a given target sequence with respect to the model output can be efficiently computed using a forward-backward algorithm, CTC models can be trained in an end-to-end fashion.

Greedy CTC decoding can be described using WFST as finding the shortest path in the decoding graph $\mathscr{G}$

$$\mathscr{G} = \mathscr{C} \circ \mathscr{H}. \tag{4}$$

$\mathscr{H}$ is a dense emission graph of the acoustic model. It has one node for each time-frame in the input audio and one transition for each token in the label set $\mathscr{T}$ between each node. The weight of each transition is given by the models output probability for the corresponding token at that time-frame. $\mathscr{C}$ is a graph modeling the collapsing and blank-removal of CTC by transducing repetitions and emissions of the blank label to no output. $\mathscr{G}$ accepts all possible sequences given the label set, limited by the length of the input.

In [3] CTC decoding is constrained by introducing a graph $\mathscr{S}$ and computing its composition with $\mathscr{C}$ before computing the composition with $\mathscr{H}$:

$$\mathscr{D} = (\mathscr{C} \circ \mathscr{S}) \circ \mathscr{H}. \tag{5}$$

The decoding graph $\mathscr{D}$ now only accepts those sequences which are accepted by $\mathscr{S}$. $\mathscr{S}$ is constructed such that it accepts all expected phonetic realizations given an orthographic transcript of the utterance.

First, it accepts multiple pronunciations for each word. In the present work the pronunciations are given by a pronunciation dictionary containing all pronunciations seen in the manual phonetic transcription of the training set of kidsTALC [5]. In $\mathscr{S}$, the different pronunciations are weighted according to their relative frequency in the training set. For words not seen in the training set, an external pronunciation dictionary is used. Only using the pronunciations from the training data and the external dictionary limits the system to seen pronunciations. A data-driven approach to model sub-word pronunciation pattern is added. This approach is similar to the extension patterns MAUS uses. Substitutions, deletions and insertions between the manual phonetic transcription and the canonical pronunciation given the orthographic transcription are counted. Transitions are added to $\mathscr{S}$, weighted by the frequency of these deviations. [3] also investigates the effect of allowing deviations based on phonological error patterns. Since this does not lead to improved results, we do not use this approach in the present work.

2.3 Phone Alignment

MAUS can intuitively obtain the alignment by back-tracing the Viterbi path of the decoding graph. The decoding graph can include the pronunciation variants as described in Sec. 2.1 or can be a linear graph representing the manual phonetic transcription for forced alignment. While MAUS is trained on a frame-wise alignment yielding to a good approximation using the Viterbi path, CTC is trained end-to-end. As mentioned in Sec. 2.2, CTC learns only an implicit alignment. The outputs of CTC are usually prone to spike [7], i. e., each token often only has one time-frame with a high probability with the remaining time-frames having a high probability for the blank token. When back-tracing the Viterbi path of the CTC decoding graph, therefore only this one time-frame can be clearly assigned to the token. Boundaries between tokens can not be identified intuitively. For frame-wise classification the frames emitting a blank token need to be assigned to the neighboring non-blank tokens. In this work we test a central split, i. e., splitting the frames assigned to a blank token equally between the neighboring valid token.

Another approach was recently introduced by Kürzinger *et al.* [8]. The authors present a method based on a model trained with CTC to align large German corpora. However, it is important to notice that this method was introduced to align utterances within large German corpora and does not directly target phone-level segmentation. In a forward pass, the decoding is carried out by suppressing any repetitions of non-blank token, i. e., only allowing each label in the target sequence s to spike once. In a backward pass the most probable path is back-traced similarly to the Viterbi path. During the backward-pass every frame, after the last valid token was consumed, is assigned to this token. As soon as the next token is consumed, the frame is assigned to the new token, which leads to a non-central assignment.

3 Experimental Setting

In this section we will explain the used dataset and the trained CTC-models. Furthermore, we give details to the different pronunciation dictionaries.

3.1 kidsTALC

We utilize kidsTALC [5], which is a corpus of typically developing, monolingual standard German speakers in the age range of 3½ to 11 years. It provides a manual orthographic and phonetic transcription, as well as utterance level segmentation. The transcription relies on a reduced token set of IPA symbols, which can be directly mapped to the X-SAMPA Symbols used by MAUS. The token set of kidsTALC does not differentiate between diphthongs and separate vowel pairs. Furthermore, the glottal stop is not part of the reduced token set.

All evaluations are based on the validation split of the kidsTALC corpus. Within this set 8 children, one female and one male child from each of the four age groups, are present and not part of the training data. We neglect hard- and unintelligible utterances, which contain words that are semantically meaningless or not transcribable at all by a speech language therapist. In total this results in 50 min of children's speech for evaluation and about 6 h of training data.

3.2 CTC model

The CTC-model is trained on the kidsTALC corpus, enlarged using Mozilla Common Voice [9]. For training we use SpeechBrain [10], whereas the training script is closely related to the TIMIT recipes provided by the SpeechBrain community. For the alignment on TIMIT [11], we used a pretrained CTC-model provided by the SpeechBrain community, as well. The constrained decoding is efficiently implemented using k2 [12].

3.3 Grapheme-to-Phoneme

In the our evaluation we rely on two different canonical pronunciations. First, we use the Grapheme-to-Phoneme (G2P) of the BAS web services [13], which we refer to as a baseline. We do not have access to the system and cannot outline further details, but assume it focuses on adult speech. Secondly, we use a domain specific dictionary, which is based on the most common pronunciations existing in the kidsTALC train set. For words not existing in the train set we use pronunciations from the BAS repository[14] extended by a trained seqitur-G2P model [15] for out-of-vocabulary (OOV) words.

4 Results

In this section we will present the results from our experiments. First, we compare the performance in pronunciation modeling. Furthermore, we show the alignment capabilities of MAUS and a CTC-based system. This is done only qualitatively on the kidsTALC corpus, since it unfortunately does not provide time alignments on phone-level. To also provide some quantitative comparisons, we additionally evaluate on the TIMIT-dataset [11].

4.1 Pronunciation Modeling

In this section we evaluate the pronunciation modeling capabilities. We compare the canonical pronunciations provided by the different G2P approaches, as well as the adjustments made by the different pronunciation modeling systems. We also evaluate the end-to-end acoustic model performance of the CTC-based system. For all, we compute the phone error rate (PER) to the manual phonetic transcription given in the kidsTALC dataset. All numbers in Tab. 1 are based on the validation set, since the manual transcriptions for the test set of the kidsTALC corpus are not publicly available. For a fair comparison we allowed MAUS to output diphthongs if desired and split the diphthongs before aligning to the ground truth. The same applies to glottal stops, which are removed before comparison. Furthermore, we noticed inconsistency for the pauses, which we neglected in the PER as well.

Table 1 – Phone error rate (PER) with respect to the manual phonetic transcript on the validation set of the kidsTALC corpus[5].

Method	**PER**
End-to-end	32.43 %
G2P (BAS Webservices)	21.67 %
G2P (BAS Webservices) + MAUS	21.03 %
Domain Specific Dictionary	10.07 %
Domain Specific Dictionary + MAUS	12.36 %
Domain Specific Dictionary + End-to-end	9.50 %

When the CTC-model is greedily decoded, i. e., without any constraining based on the manual orthographic transcription, a performance in terms of PER of 32.43 % is achieved. To naively utilize the manual orthographic transcription, we apply two different pronunciation dictionaries: A G2P from the BAS Webservice and a domain specific dictionary based on the train set of the kidsTALC dataset. We refer to this automatically translated transcript as pseudo phonetic. The pseudo phonetic transcript using the G2P of the BAS web services achieves a PER of 21.67 %. Using the domain specific dictionary generated from the kidsTALC train set reduces the PER to 10.07 %, showing that using a domain specific dictionary is important for children's

speech. While one common reason is the presence of uncommon words, we also noticed that the most prominent pronunciations of the children often contain pronunciation mistakes and deviate from the canonical pronunciation. A good example is the German word *und*, where the most prominent pronunciation does not contain the /t/ at the end.

Applying the MAUS pipeline on both transcripts with canonical pronunciations was only successful when using the G2P of the BAS web services, improving the PER to 21.03 %. MAUS is not able to further improve the transcript generated using the domain specific dictionary, but increases the PER to 12.36 %. This can most likely be explained by the domain gap between the training data for the acoustic models and the pronunciation rules, and the children's speech it is applied to. However, the constrained CTC-based decoding is capable of improving over the pseudo phonetic transcript from the domain specific dictionary. It is important to notice that the utilized model is trained end-to-end on children's speech. This enables the system to achieve a relative improvement of 5.66 %, i. e., an overall PER of 9.5 %.

4.2 Phone Alignment

In this section we evaluate the segmentation capabilities of MAUS and CTC-based systems using the frame accuracy (FA). The FA refers to the proportion of correctly assigned frames, given a phone-level manual segmentation.

Table 2 – Frame accuracy (FA) with respect to the manual phone-level segmentation on the test set of the TIMIT corpus [11].

Method	**FA**
Kürzinger [8]	59.54 %
CTC + Viterbi	77.90 % (96.73 %)
MAUS	87.15 %

We start by evaluating the performance on the TIMIT speech corpus, since it provides manual phone-level segmentation, in Tab. 2. The method based on Kürzinger *et al.* [8] described in Sec. 2.3 only achieves a FA of 59.54 %. We argue the reason lays in the design itself, which targets the segmentation on utterance-level for large audio files. In comparison, when aligning based on the Viterbi path and splitting the blank token centered between the neighboring tokens, the performance in terms of FA is improved to 77.90 %. Since CTC only implicitly learns an alignment, and due to the bi-directional LSTM, the CTC-based model is in theory capable of outputting the entire sequence s in the first $N = |s|$ frames. To demonstrate the capabilities of the model to still correctly place the tokens, and only requiring a more elaborated division of the blank-token to accurately identify boundaries between tokens, we double assigned the frames with blank tokens to both neighboring tokens. The double assignment leads to a score of 96.73 %, which is not comparable to the other models but gives a reasonable intuition of the alignment done by CTC-based models. MAUS outperforms all other systems leading to a performance in terms of FA 87.15 %.

On the kidsTALC corpus, we qualitatively compare MAUS with a CTC-based model trained on children's speech. First, we compute the agreement for the constrained decoding and forced alignment. The agreement refers to the percentage of frames, where MAUS and the CTC-based system agree after aligning both resulting transcripts. For forced alignment the alignment of both transcripts is not necessary as the transcripts are identical. We do not split up diphthongs or remove pauses as done in Sec. 4.1, because only the agreement was of interest and no metric with respect to a ground truth was applied. For the constrained decoding we reach an overlap of 51.63 %, while for forced alignment this value drops to 40.92 %. We argue, that this can be

explained by the models giving the target sequence an even lower probability, i. e., if a target token is highly improbable given the model, the alignment will most likely be wrong as well.

We further qualitatively analyze the two alignments of the constrained decoding for one specific child. We notice that MAUS results in more accurate alignments as long as the acoustic model was capable of recognizing the correct token. The CTC-based system has a special problem with the token boundaries, while being accurate on the token placement. The inaccurate boundaries are originated in the re-assignment of the blank token, as described in Sec. 2.3. However, whenever the audio gets less intelligible, the acoustic model of the CTC-based system is more robust in terms of phone recognition, also leading to a more accurate segmentation.

5 Conclusion

In this work we evaluate MAUS [2] on German children's speech and compare it to a similar system using an end-to-end speech recognition acoustic model [3]. Both apply constraining towards a canonical pronunciation based on an orthographic transcript in a similar fashion. However, the MAUS system is incapable of correctly modeling deviations from the canonical pronunciations given by a domain specific pronunciation dictionary. In contrast, the CTC-based model trained on children's speech improves upon the transcript given by the domain specific pronunciation dictionary by relatively 5.66 %. For phone-level segmentation we first compare both systems on the TIMIT corpus [11], where MAUS outperforms all CTC-based systems. However, for the kidsTALC corpus [5] only a qualitative analysis was possible. While MAUS is more accurate in general for phone-level segmentation, the CTC-based system is more robust, when the acoustic model of MAUS fails due to domain mismatch. An interesting further step is fine-tuning MAUS on children's speech and evaluating the segmentation capabilities of both systems using a children's speech corpus manually segmented on phone-level.

References

[1] POTAMIANOS, A. and S. NARAYANAN: *Robust recognition of children's speech. IEEE Transactions on Speech and Audio Processing*, 11(6), pp. 603–616, 2003.

[2] SCHIEL, F., C. DRAXLER, and J. HARRINTON: *Phonemic Segmentation and Labelling using the MAUS Technique*. In *Workshop 'New Tools and Methods for Very-Large-Scale Phonetics Research'*. 2011.

[3] RUMBERG, L., C. GEBAUER, H. EHLERT, U. LÜDTKE, and J. OSTERMANN: *Improving Phonetic Transcriptions of Children's Speech by Pronunciation Modelling with Constrained CTC-Decoding*. In *Proceedings INTERSPEECH 2022 – 23rd Annual Conference of the International Speech Communication Association*, pp. 1357–1361. ISCA, 2022.

[4] GRAVES, A., S. FERNANDEZ, F. GOMEZ, and J. SCHMIDHUBER: *Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks. International Conference on Machine Learning (ICML)*, p. 8, 2006.

[5] RUMBERG, L., C. GEBAUER, H. EHLERT, M. WALLBAUM, L. BORNHOLT, J. OSTERMANN, and U. LÜDTKE: *kidsTALC: A Corpus of 3- to 11-year-old German Children's Connected Natural Speech*. In *Interspeech 2022*, pp. 5160–5164. ISCA, 2022.

[6] KIPP, A.: *Automatische Segmentierung Und Etikettierung von Spontansprache*. Ph.D. thesis, Technical University Munich, Germany, 1998.

[7] ZEYER, A., E. BECK, R. SCHLÜTER, and H. NEY: *CTC in the Context of Generalized Full-Sum HMM Training*. In *Interspeech 2017*, pp. 944–948. ISCA, 2017.

[8] KÜRZINGER, L., D. WINKELBAUER, L. LI, T. WATZEL, and G. RIGOLL: *CTC-Segmentation of Large Corpora for German End-to-End Speech Recognition*. In A. KARPOV and R. POTAPOVA (eds.), *Speech and Computer*, vol. 12335, pp. 267–278. Springer International Publishing, 2020.

[9] *Mozilla Common Voice 8.0, German*. https://commonvoice.mozilla.org/en/datasets, 2022.

[10] RAVANELLI, M., T. PARCOLLET, P. PLANTINGA, A. ROUHE, S. CORNELL, L. LUGOSCH, C. SUBAKAN, N. DAWALATABAD, A. HEBA, J. ZHONG, J.-C. CHOU, S.-L. YEH, S.-W. FU, C.-F. LIAO, E. RASTORGUEVA, F. GRONDIN, W. ARIS, H. NA, Y. GAO, R. D. MORI, and Y. BENGIO: *SpeechBrain: A General-Purpose Speech Toolkit*. 2021.

[11] GAROFOLO, J. S., L. F. LAMEL, W. M. FISHER, J. G. FISCUS, and D. S. PALLETT: *DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1. NASA STI/Recon Technical Report N*, 93, p. 27403, 1993.

[12] *K2*. https://github.com/k2-fsa/k2, 2022.

[13] *BAS Webservice*. https://clarin.phonetik.uni-muenchen.de/BASWebServices/, last accessed 18.01.2023.

[14] *Bavarian Archive for Speech Signals (BAS)*. http://hdl.handle.net/11858/00-1779-0000-000C-DAAF-B, 2013.

[15] BISANI, M. and H. NEY: *Joint-sequence models for grapheme-to-phoneme conversion*. *Speech Communication*, 50(5), pp. 434–451, 2008.