Evaluating Automatic Speech Recognition for Child Speech Therapy Applications

Adam Hair[1], **Kirrie J. Ballard**[2], **Beena Ahmed**[3,4], **and Ricardo Gutierrez-Osuna**[1]
[1]Texas A&M University, [2]University of Sydney,
[3]University of New South Wales, [4]Texas A&M University at Qatar
adamhair@tamu.edu

ABSTRACT
Automatic speech recognition (ASR) technology can be a useful tool in mobile apps for child speech therapy, empowering children to complete their practice with limited caregiver supervision. However, little is known about the feasibility of performing ASR on mobile devices, particularly when training data is limited. In this study, we investigated the performance of two low-resource ASR systems on disordered speech from children. We compared the open-source PocketSphinx (PS) recognizer using adapted acoustic models and a custom template-matching (TM) recognizer. TM and the adapted models significantly out-perform the default PS model. On average, maximum likelihood linear regression and maximum a posteriori adaptation increased PS accuracy from 59.4% to 63.8% and 80.0%, respectively, suggesting that the models successfully captured speaker-specific word production variations. TM reached a mean accuracy of 75.8%.

CCS Concepts
•**Computing methodologies** → **Speech recognition;**

Author Keywords
Assistive Technology; Computer-Assisted Pronunciation Training (CAPT)

INTRODUCTION
Child speech therapy practice should be frequent and high-intensity [6], and therefore, sessions at the clinic need to be supplemented with considerable home practice, which can become tedious. Primary caregivers typically administer home practice, but busy schedules decrease practice frequency [7]. To address issues stemming from boring and infrequent home practice, we developed a mobile speech therapy game called Apraxia World [3]. However, in order for children to be able to practice independently and take ownership of their therapy, Apraxia World and similar systems need to include automated speech recognition (ASR) capabilities to provide utterance feedback. In this study, we investigate ASR performance on disordered speech from children using limited training data and mobile-device-friendly techniques. We focus on verifying that an utterance is close to the intended target, which ensures that the child is making an appropriate effort to say the word (i.e. they cannot say something completely different than the target word) while reserving deeper analysis (i.e., phonological) for trained SLPs.

For this task, we examine two low-resource ASR methods: adapting existing acoustic models and template matching. Acoustic model adaptation uses sample recordings from a speaker to better align the model with how that person speaks, creating a speaker-dependent model. In contrast to adaptation, template matching uses the utterances directly to represent how specific words should sound. In our approach, examples of words spoken by the speaker are used as templates to determine if a new recording matches the target word. Our findings demonstrate that although template matching performs well, PocketSphinx with an adapted model achieves significantly higher accuracy. This suggests that limited training data can successfully capture speaker-specific word production variants. The main contributions of this article are (1) an empirical test of PocketSphinx performance on disordered speech from children and (2) a direct comparison of performance of two ASR frameworks using limited training data.

AUTOMATIC SPEECH RECOGNITION
PocketSphinx (PS) is a mobile-ready version of the Sphinx ASR engine developed at CMU [4]. An acoustic model provides information about how specific speech elements "sound" to the recognizer. PS acoustic models can be better aligned with a specific speaker through two types model adaptation, maximum likelihood linear regression (MLLR) [5] and maximum a posteriori (MAP) [2]. Both of these methods take speech from the target speaker to update the model. MLLR estimates linear transformations for the Gaussian means and variances, whereas MAP uses prior information about the parameter distribution combined with the adaptation data to re-estimate all model parameters [10]. Recognition performance can be further improved by using a constrained lexicon that only contains words the child should be practicing in their speech therapy session; this keeps the ASR from searching for irrelevant words when decoding the speech.

ASSETS'19, October 28–30, 2019, Pittsburgh, PA, USA

ACM ISBN 978-1-4503-6676-2/19/10.
DOI: https://doi.org/10.1145/3308561.3354606

In contrast to the statistical models used by PS, template matching recognizes words by directly comparing an utterance against previously-collected utterances. Prior to performing template matching, utterances must be transformed into sequences of acoustic feature vectors. We take an audio recording, trim leading and trailing silence with an energy threshold, pre-emphasize the signal, and extract 13 Mel-frequency cepstral coefficients (MFCCs). We discard the first coefficient (spectral energy) and normalize the remaining coefficients with cepstral mean normalization (CMN) [1].

Following feature extraction, template and test utterances are aligned end-to-end using dynamic time warping (DTW). We compute the distance (root-mean-squared-error) between the template and test utterance using the DTW frame alignment. The framework classifies a test utterance by comparing it against templates for all possible word classes. Let Γ_ω be the set of templates for word ω. The test utterance score is the mean template matching score from comparing a new utterance u against all templates in Γ_ω, where $score(u,\Gamma) = \text{mean}(\{d(j,u)|\forall j \in \Gamma\})$. The test utterance is assigned to the class ω with the lowest score: $label(u) = \arg\min_\omega score(u,\Gamma_\omega)$.

EXPERIMENTS

Although some child speech datasets exist (e.g., the OGI kids' speech corpus [9]), they usually contain speech from typically-developing children, with few mispronunciations. Therefore, we gathered a real-world corpus to better evaluate speech recognition performance on disordered speech. Disordered speech data were collected from seven Australian children (1f, 6m, 7-9 y.o.) with speech sound disorders while they played Apraxia World for eight weeks under caregiver supervision. This provided a large set of scripted single-word recordings. Audio was recorded at 16 kHz using a headset attached to a Samsung Tab A 10.1 tablet. The children started and stopped the recordings on their own, so some recordings were stopped prematurely or "clipped." In total, 21,198 utterances were recorded. Recordings that were clipped or that contained substantial background noise were discarded, leaving 10,415 recordings.

To evaluate the template-matching framework, we developed a prototype system using the librosa audio processing library for Python [8]. For convenience, we ran tests on a desktop computer, but this framework can also be used on mobile devices and lends itself to parallelization. We randomly selected 15 child-specific templates per target word and used the remaining recordings of the child saying the target word as test data. This process was repeated 5 times, each with new templates selected at random. Since the children only practiced 10 words at a time, each recording can only be labeled as one of the words practiced in that phase (using 10 sets of 15 templates).

For tests with PS, we started with the default American English acoustic model trained on adult speech[1]. To account for dialect and age differences, we also created speaker-dependent acoustic models by adapting the default model with both MLLR and MAP. The acoustic model was adapted using 15 examples for each of the 20 practiced words (300 utterances total per speaker), the same amount of data used in the template-matching approach. The PS decoder was configured with a 10-word lexicon to only recognize words practiced in the respective treatment phase. The remaining 8,315 utterances were used as test data, which were passed to the PS decoder without any additional preprocessing.

[1] https://github.com/cmusphinx/pocketsphinx/tree/master/model/

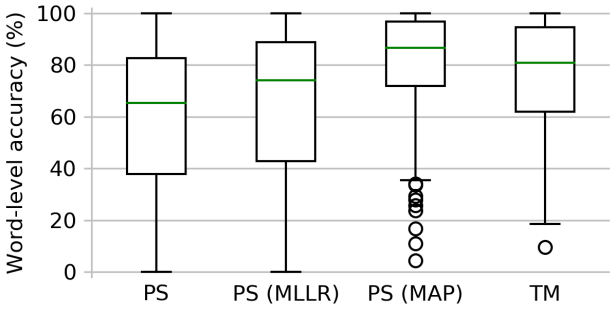

Figure 1: Accuracy measured per word across all speakers (15 utterances per word for adaptation and template matching)

Results and Conclusions

Figure 1 shows the per-word accuracy for all speakers. The MAP-adapted models yield the best recognition performance. Both the MAP-adapted models and template matching correctly recognize all words at least some of the time; this is contrasted with PS with the non-adapted and MLLR-adapted models, both of which never correctly recognized some of the words. Regardless, MLLR- and MAP-adapted models both performed significantly better than the non-adapted model (paired $t(139) = 6.0, p << 0.01$ and paired $t(139) = 13.5, p << 0.01$, respectively). Template matching performed significantly better than both the non-adapted model (paired $t(139) = 6.7, p << 0.01$) and the MLLR-adapted model (paired $t(139) = 5.1, p << 0.01$). The MAP-adapted model performed significantly better than template matching (paired $t(139) = 2.2, p = 0.03$). We used an alpha level of 0.05 for tests of significance.

Based on these results, adapting acoustic models is a viable method for automatic speech recognition with limited disordered speech data. Accent and age-related speaking differences are reduced by the MAP adaptation, which significantly improves performance over the default American English acoustic model. Additional improvements may be gained by training a speaker-independent child model and adapting that to each child, but we leave that analysis for future study. Although this article focuses only on word recognition, further work should investigate the relationship between child pronunciation quality and speech recognition accuracy with a pathologist-annotated corpus.

ACKNOWLEDGEMENTS

This work was made possible by NPRP Grant # [8-293-2-124] from the Qatar National Research Fund (a member of Qatar Foundation). The statements made herein are solely the responsibility of the authors.

REFERENCES

[1] S. Furui. 1986. Speaker-independent isolated word recognition using dynamic features of speech spectrum. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 34, 1 (1986), 52–59.

[2] J.-L. Gauvain and C.-H. Lee. 1994. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. *IEEE transactions on speech and audio processing* 2, 2 (1994), 291–298.

[3] A. Hair, P. Monroe, B. Ahmed, K. J. Ballard, and R. Gutierrez-Osuna. 2018. Apraxia World: A Speech Therapy Game for Children with Speech Sound Disorders. In *Proceedings of the 2018 Conference on Interaction Design and Children*.

[4] D. Huggins-Daines, M. Kumar, A. Chan, A. W. Black, M. Ravishankar, and A. I. Rudnicky. 2006. Pocketsphinx: A free, real-time continuous speech recognition system for hand-held devices. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, Vol. 1. I–I.

[5] C. J. Leggetter and P. C. Woodland. 1995. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Computer speech & language* 9, 2 (1995), 171–185.

[6] E. Maas, C. E. Gildersleeve-Neumann, K. J. Jakielski, and R. Stoeckel. 2014. Motor-based intervention protocols in treatment of childhood apraxia of speech (CAS). *Current developmental disorders reports* 1, 3 (2014), 197–206.

[7] L. McAllister, J. McCormack, S. McLeod, and L. J. Harrison. 2011. Expectations and experiences of accessing and participating in services for childhood speech impairment. *International Journal of Speech-Language Pathology* 13, 3 (2011), 251–267.

[8] B. McFee, C. Raffel, D. Liang, D. Ellis, M. McVicar, E. Battenberg, and O. Nieto. 2015. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*. 18–25.

[9] K. Shobaki, J.-P. Hosom, and R. A. Cole. 2000. The OGI kids' speech corpus and recognizers. In *Sixth International Conference on Spoken Language Processing*.

[10] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland. 2002. *The HTK book*. Vol. 3. Cambridge University Press. 175 pages.