Large Vocabulary Automatic Speech Recognition for Children

Hank Liao[1], Golan Pundak[2], Olivier Siohan, Melissa K. Carroll, Noah Coccaro, Qi-Ming Jiang, Tara N. Sainath, Andrew Senior, Françoise Beaufays, and Michiel Bacchiani

Google Inc.

[1]hankliao@google.com, [2]golan@google.com

Abstract

Recently, Google launched YouTube Kids, a mobile application for children, that uses a speech recognizer built specifically for recognizing children's speech. In this paper we present techniques we explored to build such a system. We describe the use of a neural network classifier to identify matched acoustic training data, filtering data for language modeling to reduce the chance of producing offensive results. We also compare long short-term memory (LSTM) recurrent networks to convolutional, LSTM, deep neural networks (CLDNN). We found that a CLDNN acoustic model outperforms an LSTM across a variety of different conditions, but does not specifically model child speech relatively better than adult. Overall, these findings allow us to build a successful, state-of-the-art large vocabulary speech recognizer for both children and adults.

Index Terms: speech recognition, children's speech, neural networks.

1. Introduction

Speech recognition for adults has improved significantly over the last few years; however less progress has been made in recognizing speech produced by children well [1, 2]. Many factors make recognizing children's speech challenging. As children learn to speak, their ability to accurately realize speech sounds properly changes [3, 4]. Spectrally, children's smaller vocal tracts lead to higher fundamental and formant frequencies. Children's overall speaking rate is slower, and they have more variability in speaking rate, vocal effort, and spontaneity of speech [5]. Linguistically, children are more likely to use "imaginative words, ungrammatical phrases and incorrect pronunciations" [6]. By training directly on children's speech it was shown that this mismatch in performance can be reduced on a digit recognition task, although accuracy is still worse than on adults [7]. All these aspects evolve rapidly as children grow [2].

When Google surveyed the Voice Search habits of Americans, it was found that more than 50% of teens, and 41% of adults surveyed used it every day [8]. Recognition performance on this very large vocabulary task is quite good for adults, but we still find it rather poor for children. This paper looks at some of the techniques we have examined to produce a better recognizer for children. We experimented with pitch features, spectral smoothing and vocal tract length normalization (VTLN), and applied improved acoustic modeling; however, we found that the best results were obtained by training on a large amount of data that we selected, aided by a neural network classifier, to better match children's speech. To improve noise robustness, we synthesized a multi-style training database. We used this approach to train a child-friendly Voice Search system that is used to recognize queries in the YouTube Kids app launched earlier this year [9]. It improves the recognition rate of children's speech while reducing the possibility of offensive phrases and content being recognized. As far as the authors are aware, this is the first publicly launched, large vocabulary continuous speech recognition (LVCSR) system that works well for children.

2. Previous Work

There have been a variety of approaches to improving speech recognition for children. Gray et al. [6] present an LVCSR systems that is used for interacting with electronic devices, however it is difficult to determine how well it works without access to it, absolute error rate figures, or vocabulary size; in comparison, this paper uses improved acoustic modeling, an order of magnitude more data, and shows how these factors affect the recognition of adult speech versus child.

For children's speech recognition, Ghai and Sinha [10] suggest that the standard mel-spaced filterbanks used for producing features for recognition are sub-optimal. They observed that in high pitch speech there are distortions in the lower frequency filters, and they suggest increasing the filter bandwidth to smooth them out. While their experiments showed good improvements, they were performed with MFCC features and a weaker Gaussian mixture model (GMM) acoustic model.

VTLN [11, 12, 13] is an approach that was designed to alleviate the impact of different vocal tract shapes among speakers on recognition performance. It is defined as a speaker normalization procedure which implements a frequency warping of the short-term spectral features onto a canonical space. This warping is usually applied both during training and recognition time. One attractive aspect of VTLN is that the transformation is represented by a single scalar parameter, the warping factor, which can be estimated from a small amount of data. Several studies have reported that VTLN is effective in improving the recognition performance on child speech [6, 14, 15]. However, most of those studies were limited to either DNN-based acoustic models with small training sets, in the order of tens of hours in [14], or to medium-size training sets ($\approx$200 hours) with GMM-based acoustic models.

One method commonly used alongside VTLN to adapt a general acoustic model to a speaker is linear transformations of features or model parameters (MLLR/CMLLR) [16]. This technique has shown gains when adapting an existing GMM adult model for a specific child speaker (e.g [6, 2]). We find this adaptation technique unsuitable for our Voice Search task for several reasons. First, our utterances are only a few seconds long, making reliable parameter estimation very challenging. Second, our low latency requirement restricts the use of two-pass decoding strategies. Third, the improvement from using CMLLR adapted features does not appear to transfer from GMM to neural network acoustic models [17].

3. A Speech Recognizer for Children

Many of the techniques in the previous section were shown to be effective on small vocabulary tasks, with a small amount of training data, or applied to GMM acoustic models. However our users expect a virtually unlimited vocabulary system. This section describes the key areas that we focused on to produce a deployable LVCSR system that works well for children.

3.1. Data Selection And Transcription

Typically thousands of hours of speech are required to build a state-of-the-art acoustic model. Our generic training sets were obtained by sampling from our Voice Search traffic. The data is then anonymized and hand-transcribed. An obvious method to improve children's speech recognition is to collect a matched training set in the same manner. To account for children's vocal tract lengths we collect wide-band audio so reliable estimation of high frequency filter banks can be made [18].

To obtain labeled data, humans are asked to only transcribe intelligible child utterances sampled from our Voice Search traffic and reject the rest. Since no explicit adult or child labeling is provided, the rejection is based on audio only. Only about 6% percent of our American English Voice Search traffic comes from children, so for efficiency we try to filter out adult utterances in advance. To perform this filtering, a child speech DNN classifier was built, similar to the child speech tagger described in [6]. The classifier has two output classes for each frame: child and adult. Utterance level classification is obtained by averaging the frame scores. The DNN is composed of two hidden layers with 320 hidden nodes each, and a softmax output layer. As input, a 40-dimensional log mel filterbank feature was used, with 20 context frames stacked to form a super-vector. The classifier was trained on 20k human labeled utterances. By tuning the acceptance threshold on the child class, 89% precision and 40% recall was obtained on a held-out set. Combining our existing adult and the "child" speech data obtained using this approach, we produced the datasets described in Table 1.

Table 1: *Data sets.*

Data Set	# Utts	# Frames	Description
`vs2014`	2.6M	607M	Voice Search **adult**, 16kHz
`hp2014`	1.9M	459M	Voice Search **child**, 16kHz
`vshp2014`	4.5M	1066M	`vs2014` and `hp2014`

While the `vs2014` data set is described as adult speech, as a sample of actual traffic, it contains some high pitch speech which are likely children. For evaluation, Adult and Child test sets were similarly created with about 25k utterances each. The Child test set differs from the Adult in that to get more reliable test transcriptions, majority vote among three human transcribers was taken for each utterance, and utterances without agreement were discarded.

We also sought to explore the robustness of our model to noise. Creating an artificially noisy data set to train a system for a mismatched test environment can be described as multi-style training [19]. Adding noise to training data can produce multi-style acoustic models that are significantly better on artificially generated and real-world noisy data without impairing performance on the original data [20]. Although the original data already contains background noise, we create a noisier version by adding varying degrees of noise and reverberation at the utterance level, such that overall SNR is between 5dB and 30dB. Samples of noise were taken from YouTube and daily life noisy environmental recordings. During training, target labels were generated using the original speech data. The test sets were corrupted in the same way to create matched "noisy" versions.

3.2. Acoustic Modeling

Given the recent success with LSTM models on adult speech [21], we began our experiments with LSTM acoustic models. The features are 40-dimensional log mel-spaced filterbank coefficients, without any temporal context. We experimented with adding pitch features, as in [22], but found no improvement. Note that Ghahremani et al. did not find gains on Switchboard, and here we use more data that is also noisier which can affect the pitch tracker. The acoustic model is comprised of 2 LSTM layers, where each LSTM layer has 800 cells, a 512 unit projection layer for dimensionality reduction, and a softmax layer with 13,522 context-dependent triphone states [23] clustered using decision trees [24]. The same clustering is used for all the models presented in this paper and were estimated on an older Voice Search data set. We use a reduced X-SAMPA phonetic alphabet of 40 phones plus silence.

There has been recent success in using convolutional LSTM deep neural network (CLDNN) models on adult speech [20]. Convolutional layers provide robustness to vocal tract length variation similar to VTLN, however do so by normalizing small shifts in frequency rather than feature warping [25]. CLDNNs seem a natural fit for our task of building a model that can recognize child speech while still performing well on adult speech.

We briefly describe the CLDNN architecture here, however more details can be found in [20]. The filterbank features are fed into a single convolutional layer, which has 256 feature maps. We use an 8×1 frequency$\times$time filter for the convolutional layer. Our pooling strategy is to use non-overlapping max pooling and only in frequency [26]. A pooling size of 3 was used. The CNN output is followed by 2 LSTM layers, where each LSTM layer has 832 cells and a 512 unit projection layer for dimensionality reduction. Finally, we pass the output of the LSTM to 2 fully connected DNN layers with rectified linear unit activations. Each fully connected layer has 1,024 hidden units. To reduce the network size, the softmax layer is factored into two with an intermediate 512-node linear low-rank layer [27].

During training, recurrent networks are unrolled for 20 time steps for training with truncated back-propagation through time (BPTT) [28]. In addition, the output state label is delayed by 5 frames, as we have observed with DNNs that information about future frames helps to better predict the current frame.

3.3. Language Modeling

For a child-friendly experience, we had two goals: minimize the likelihood of offensive mis-recognitions, and better model the types of queries that children were likely to utter. Our training data, consisting of logged typed queries, sentences from web pages, and high confidence machine transcriptions of spoken queries, contains two types of offensive language: those matching a human constructed set of bad words that are offensive in any context, and those consisting of individually innocuous words that put together are classified as an offensive seeking query. Simply eliminating all bad words from our training data could lead to the situation where a child said one of these words and we recognized it as a misspelled or acoustically similar version. Instead, this type of training data was greatly reduced, but not quite eliminated. Offensive queries were eliminated with an offensive query classifier. This classifier was trained using query character n-grams as well as the fraction of offensive web pages in the query's search results reducing the likelihood of offensive phrases composed of individually innocuous words.

There were three things we did to select content for our training data that was likely to match the queries we expected to get from children. First, highly confident machine tran-

Table 2: *Perplexity for textual Child dev sets.*

Data Set	Baseline	Post-filtering
Child-friendly Queries	412.7	209.6
Offensive Queries	353.2	2896.6

scriptions of utterances acoustically classified as likely from children (described in Section 3.1) were used as an unsupervised corpus (80M voice queries). Second, we suppressed web pages that were rated as relatively difficult to read using a reading level classifier [29] inspired by work by Collins-Thompson and Callan [30] and boosted data at the 2nd to 6th grade reading level. Third, search queries from search sessions that led to clicks on child-friendly web sites were added to our training data (3B typed queries). These web sites were identified through an algorithm using user click behavior. Roughly, a small set of sites of no interest to children was chosen as negative scored seeds, and a small set of sites of high interest to children was chosen as positive scored seeds. Web sites that were clicked in the same session where a seeded site was clicked received some of the seed's score. This was iterated through all sessions—in total about 0.12% of web pages were deemed child friendly (3B sentences from web pages). Separate 5-gram language models were trained on each of the above sources and then combined using Bayesian interpolation [31] to yield a 100M n-gram model. The vocabulary is a list of 4.6M tokens extracted from written domain text. As shown in Table 2, we were able to model our target child audience with lower perplexity, but greatly reduce the chance of offensive queries being predicted compared to our baseline Voice Search language model.

4. Experimental Results

The experiments are conducted on proprietary, anonymized mobile search data sets described in Section 3.1. Language models are trained in a distributed manner [32]. All acoustic models are trained using asynchronous gradient descent [33], first with a cross-entropy criterion and then further sequence-trained with a sMBR criterion [34] until convergence unless otherwise noted. Cross-entropy models typically converge after 25-50 epochs, and less than 10 for sequence training. An exponentially decaying learning rate was used.

4.1. Feature Compensation

We explored spectral smoothing, as described in Section 2, for children's speech recognition with an LSTM acoustic model. By only smoothing at test time, the word error rate more than doubled from 15.0% to 36.6%. This is likely because the filterbank introduced a further mismatch between test and training data, so we applied the same filter widths in the training data. Utterances are aligned with the same alignment model and default filterbank, but three different feature sets were produced with different minimum filter widths of 150, 200 and 250 Hz. We found that the greater minimum filter width also resulted in progressively worse training and test frame accuracy during cross entropy training; the baseline word error rate worsens from 12.9% to 15.7%. We suspect the smoothing destroys useful low-frequency information.

We also investigated the impact of VTLN on training sets consisting of a few thousand hours of child speech for LSTM acoustic models. For experimentation, we ignore real-time and latency constraints and use a 2-pass decoding strategy where the hypothesis from a first-pass decoding is used as supervision to obtain a maximum likelihood estimate of the warping factor per utterance from 0.85 to 1.15. This is challenging because each utterance is short: on average 4.4 seconds long. First an oracle experiment was conducted where the ground truth reference transcript was used to estimate the optimal warp factor at test time. Under this scenario, we found that the WER would decrease from 14.0% to 13.5% on the Child test set. However, when the first pass recognition hypothesis is used to estimate the warp factor, the WER increased to 14.2%. We found that a third of the utterances in the test set had a much higher error rate: around 27%; when the warp factor was estimated using these poor hypotheses, the results were worse than not using VTLN at all. On these difficult utterances, the quality of the first pass supervision, combined with short utterance duration, did not lead to a robust estimate of the warping factors. We conclude that VTLN was not effective for our application.

4.2. Combining Adult and Child Training Data

Although YouTube Kids is targeted for children, the ASR should still work well for parents that are also likely to use it. An obvious solution for this is to use an additional model dedicated to adults which will be selected in decoding time by some classifier, e.g. the same one we used for data selection. The main drawback from this solution is the added recognition complexity. The approach we chose was to train one model that will serve both adults and children. We found that adding adult utterances to the training set works well, and in fact improves the children's speech recognition, as shown in Table 3.

Table 3: *Word error rate (%) for LSTM models trained on different fractions of* `vs2014` *added to* `hp2014`.

% of `vs2014`	# Utts	Adult	Child
0	1.9M	37.2	11.2
20	2.4M	14.2	10.3
40	2.9M	13.8	10.2
60	3.5M	13.6	10.1
80	4.0M	13.4	10.0
100	4.5M	13.4	10.2

We also wanted to evaluate accuracy improvements from a smaller amount of child data. To check this we added portions of `vshp2014` to `vs2014` and trained an LSTM model. We found that adding the entire child data set helps significantly more than adding a portion of it, as can be seen in Table 4. The second line in Table 4 shows the result for a model trained with 200k child and 2.6M adult utterances, which achieved a 0.2% improvement over a model trained only on adult speech.

Our results in combining `vs2014` and `hp2014` data for CLDNN models were similar, however we found that using all the data, rather than some fraction gave the best results for both Adult and Child data sets. Table 5 summarizes the effect of adding the entire `hp2014` data set to `vs2014` on both Adult and Child test sets. The CLDNN is clearly better than the LSTM model by 6-10% relative depending on test and training condition. The CLDNN model trained solely on `vs2014` contains one additional LSTM layer, which was removed when training with `vshp2014` because it increased training time without affecting performance. The LSTM models and `vs2014` trained CLDNN have comparable number of parameters: 12.8M versus 13.1M. In [21], it was that shown larger LSTM models did not improve accuracy.

4.3. Results By Pitch

To further understand how LSTM and CLDNN results compare with respect to adult and child speakers, we apply the YIN fundamental frequency algorithm [35] to estimate the average pitch

Table 4: *Word error rate (%) for LSTM models trained on different fractions of* `hp2014` *added to* `vs2014`.

% of `hp2014`	# Utts	Adult	Child
0	2.6M	13.4	11.9
10	2.8M	13.6	11.7
20	3.0M	13.7	11.5
40	3.4M	13.7	11.1
80	4.1M	13.7	10.7
100	4.5M	13.4	10.2

Table 5: *Word error rate (%) by training data and AM.*

Training Data	Acoustic Model	Test Set Adult	Child
`vs2014`	LSTM	13.4	11.9
`vshp2014`	LSTM	13.4	10.2
`vs2014`	CLDNN	12.6	10.7
`vshp2014`	CLDNN	12.5	9.4

of each utterance in the Adult test set which recall contains a small amount of child speech. The fraction of all utterances that fall in each pitch bin is shown in Figure 1 as well the average WER on the test set for the combined `vshp2014` trained CLDNN model. There is a large peak at around 120Hz representing the majority of male speakers; a smaller peak should appear around 200Hz for females, but in this region we find pitch estimation accuracy is poorer. While the overall error rate is relatively low and flat from about 100Hz-300Hz, it increases in the extreme low and high pitch ranges that occur much less frequently in the training data.

An alternative analysis was conducted by having annotators determine solely from the audio whether they thought the utterance was from a male, female or child. Since its not actually possible to do this with complete certainty, we use three transcribers for every utterance and we find that all three agree on a label 93% of the time for the 28k utterances we examined. In Figure 2 we plot out the average WER for 3 different types of acoustic models broken down into five bins: adult male (62% of test set), adult female (25%), high-pitch/child (5.7%), M-F for disagreement on adult male-female label (2.5%), and F-H for disagreement on adult female-high pitch label (3.7%). The distribution of error rates over these bins are similar for both LSTM and CLDNN although the CLDNN is better for all labels. Conventional wisdom has it that recognizers perform better for males than females, but here we have the opposite. Recognition is still significantly worse for high pitch children and surprisingly bad for those utterances where there is confusion whether the speaker is a male or female. This may be due to under-representation in the training data, however recall

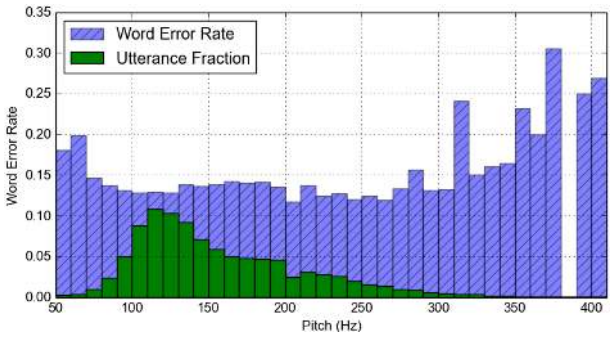

Figure 1: *Word error rate (%) versus pitch. CLDNN acoustic model trained on* `vshp2014`*, tested on Adult speech.*

this model was trained on data where 50% is high pitched. The gap that remains between the adult speaker error rates and child speech may indicate further research is needed.

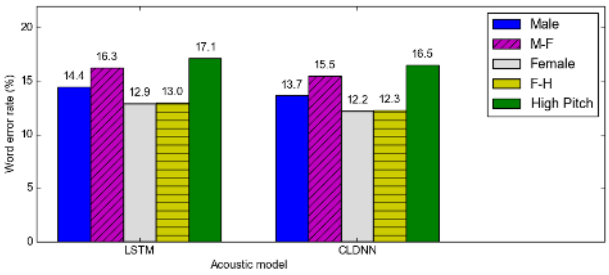

Figure 2: *Word error rate (%) by hand-labeled populations for models trained on* `vshp2014` *and tested on Adult speech.*

4.4. Multi-Style Training

As discussed in Section 3.1, we can artificially create more diverse training data to improve overall noise robustness. Table 6 compares an acoustic model trained on `vshp2014`, to an artificially noise-corrupted `vshp2014`, as described in Section 3.1, and tested on the Adult and Child test sets with and without noise added. Since the multi-style training set contains a version of `vshp2014`without noise added, there was no degradation in performance on either clean Child or Adult speech. However, as expected, the performance on noisy versions of these test sets is greatly improved. While these results are on artificially created test sets that match the multi-style training, we find in practice that this anecdotally improves our in-field recognition in noisy conditions without affecting performance in quieter ones.

Table 6: *Word error rate (%) with or without noise added to training and test sets using a CLDNN model.*

Training Condition	Adult Child	Adult Noisy	Child Clean	Child Noisy
Clean	12.5	21.0	9.4	20.0
Noisy	12.5	14.3	9.4	11.8

5. Conclusions

This paper has described the methods used to build what we believe to be the first launched, large vocabulary automatic speech recognizer suitable for children, which is used in the YouTube Kids mobile application. To reduce the chance of presenting offensive content, the language model is trained on data that is filtered of such content. Many acoustic modeling techniques for improving children's speech recognition presented in the literature such as spectral smoothing, vocal tract length normalization, and using pitch features were found not to be effective given our task. This may be due to the neural network modeling and the much larger amounts of data our system is trained on. To gather matched acoustic training data, we trained a classifier to identify child speech. We collected a supervised, child speech corpus similar in size to our adult speech. For LSTM and CLDNN models, we found that combining the high pitch and standard training corpus yields models that perform well on both child and adult speech. Overall though, the CLDNN models perform uniformly better for males, females and child speech than LSTMs. Despite the convolutional layers, CLDNNs still appear to perform slightly worse for the lowest and highest pitched speech. The work in this paper also shows that a single acoustic model can be trained to handle both children and adults as well as a variety of noise conditions.

6. References

[1] M. Gerosa, D. Giuliani, S. Narayanan, and A. Potamianos, "A review of ASR technologies for children's speech," in *Proc. Workshop on Child, Computer and Interaction (WOCCI)*, 2009.

[2] P. G. Shivakumar, A. Potamianos, S. Lee, and S. Narayanan, "Improving speech recognition for children using acoustic adaptation and pronunciation modeling," in *Proc. Workshop on Child, Computer and Interaction (WOCCI)*, 2014.

[3] M. Russell and S. D'Arcy, "Challenges for computer recognition of children's speech," in *Proc. Speech and Language Technologies in Education (SLaTE)*, 2007.

[4] A. Hämäläinen, S. Candeias, H. Cho, H. Meinedo, A. Abad, T. Pellegrini, M. Tjalve, I. Trancoso, and M. S. Dias, "Correlating ASR errors with developmental changes in speech production: A study of 3-10-year-old European Portuguese children's speech," in *Proc. Workshop on Child, Computer and Interaction (WOCCI)*, 2014.

[5] A. Potamianos, S. Narayanan, and S. Lee, "Automatic speech recognition for children," in *Proc. Eurospeech*, 1997.

[6] S. S. Gray, D. Willett, J. Lu, J. Pinto, P. Maergner, and N. Bodenstab, "Child automatic speech recognition for US English: Child interaction with living-room-electronic-devices," in *Proc. Workshop on Child, Computer and Interaction (WOCCI)*, 2014.

[7] D. Elenius and M. Blomberg, "Adaptation and normalization experiments in speech recognition for 4 to 8 year old children," in *Proc. Interspeech*, 2005.

[8] "OMG! Mobile voice survey reveals teens love to talk," 2015, http://googleblog.blogspot.com/2014/10/omg-mobile-voice-survey-reveals-teens.html.

[9] "Introducing the newest member of our family, the YouTube Kids app—available on Google Play and the App Store," 2015, http://youtube-global.blogspot.com/2015/02/youtube-kids.html.

[10] S. Ghai and R. Sinha, "Exploring the role of spectral smoothing in context of children's speech recognition," in *Proc. ICASSP*, 2009.

[11] E. Eide and H. Gish, "A parametric approach to vocal tract length normalization," in *Proc. ICASSP*, 1996.

[12] L. Lee and R. Rose, "A frequency warping approach to speaker normalization," *IEEE Trans. on Speech and Audio Processing*, vol. 6, no. 1, pp. 49–60, 1998.

[13] S. Wegmann, D. McAllaster, J. Orloff, and B. Peskin, "Speaker normalization on conversational telephone speech," in *Proc. ICASSP*, no. 1, Atlanta, 1996, pp. 339–341.

[14] R. Serizel and D. Giuliani, "Vocal tract length normalisation approaches to dnn-based children's and adults' speech recognition," in *Proc. of the IEEE Workshop on Spoken Language Technology (SLT)*, 2014.

[15] S. Umesh, R. Sinha, and D. R. Sanand, "Using vocal-tract length normalization in recognition of children speech," in *Proc. of National Conference on Communications*, Kanpur, Jan. 2007.

[16] M. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, Jan. 1998.

[17] T. N. Sainath, A. R. Mohamed, B. Kingsbury, and B. Ramabhadran, "Deep convolutional neural networks for LVCSR," in *Proc. ICASSP*. IEEE, 2013, pp. 8614–8618.

[18] M. Russell, S. D'Arcy, and L. Qun, "The effects of bandwidth reduction on human and computer recognition of children's speech," *Signal Processing Letters, IEEE*, vol. 14, no. 12, pp. 1044–1046, 2007.

[19] R. P. Lippmann, E. A. Martin, and D. B. Paul, "Multi-style training for robust isolated word speech recognition," in *Proc. ICASSP*, 1987.

[20] T. N. Sainath, O. Vinyals, A. Senior, and H. Sak, "Convolutional, long short-term memory, fully connected deep neural networks," in *Proc. ICASSP*, 2015.

[21] H. Sak, A. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," in *Proc. Interspeech*, 2014.

[22] P. Ghahremani, B. BabaAli, D. Povey, K. Riedhammer, J. Trmal, and S. Khudanpur, "A pitch extraction algorithm tuned for automatic speech recognition," in *Proc. ICASSP*, 2014.

[23] L. Bahl, P. de Souza, P. Gopalakrishnan, D. Nahamoo, and M. Picheny, "Context dependent modelling of phones in continuous speech using decision trees," in *Proc. DARPA Speech and Natural Language Processing Workshop*, 1991, pp. 264–270.

[24] S. Young, J. Odell, and P. Woodland, "Tree-based state tying for high accuracy acoustic modelling," in *Proc. ARPA Workshop on Human Language Technology*, 1994, pp. 307–312.

[25] O. Abdel-Hamid, A. Mohamed, H. Jiang, and G. Penn, "Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition," in *Proc. ICASSP*, 2012.

[26] T. N. Sainath, B. Kingsbury, A. Mohamed, G. Dahl, G. Saon, H. Soltau, T. Beran, A. Aravkin, and B. Ramabhadran, "Improvements to deep convolutional neural networks for LVCSR," in *Proc. ASRU*, 2013.

[27] T. N. Sainath, B. Kingsbury, V. Sindhwani, E. Arisoy, and B. Ramabhadran, "Low-rank matrix factorization for deep neural network training with high-dimensional output targets," in *Proc. ICASSP*, 2013.

[28] A. Robinson and F. Fallside, "The utility driven dynamic error propagation network," University of Cambridge, Tech. Rep. CUED/F-INFENG/TR.1, 1987.

[29] N. Coccaro, "Find results at the right reading level," Dec. 2010, http://googleforstudents.blogspot.com/2010/12/find-results-at-right-reading-level.html.

[30] K. Collins-Thompson and J. P. Callan, "A language modeling approach to predicting reading difficulty," in *HLT-NAACL*, 2004, pp. 193–200. [Online]. Available: "http://acl.ldc.upenn.edu/hlt-naacl2004/main/pdf/111_Paper.pdf"

[31] C. Allauzen and M. Riley, "Bayesian language model interpolation for mobile speech input," in *Proc. Interspeech*, 2011.

[32] T. Brants, A. Popat, P. Xu, F. Och, and J. Dean, "Large language models in machine translation," in *Proc. of EMNLP*, 2007.

[33] J. Dean, G. S. Corrado, R. Monga, K. Chen, M. Devin, Q. V. Le, M. Z. Mao, M. Ranzato, A. Senior, P. Tucker, K. Yang, , and A. Y. Ng, "Large scale distributed deep networks," in *Proc. Advances in NIPS*, 2012.

[34] G. Heigold, E. McDermott, V. Vanhoucke, A. Senior, and M. Bacchiani, "Asynchronous stochastic optimization for sequence training of deep neural networks," in *Proc. ICASSP*, 2014.

[35] A. de Cheveigne and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *Journal of the Acoustical Society of America*, vol. 11, no. 4, pp. 1917–1930, Apr. 2002.