ROLE OF PROSODIC FEATURES ON CHILDREN'S SPEECH RECOGNITION

Hemant K. Kathania[1], S. Shahnawazuddin[2], Nagaraj Adiga[3] and Waquar Ahmad[1]
[1]Department of Electronics and Communication Engineering, NIT Sikkim, India
[2]Department of Electronics and Communication Engineering, NIT Patna, India
[3]Department of Computer Science, University of Crete, Greece
hemant.ece@nitsikkim.ac.in, s.syed@nitp.ac.in, nagaraj@csd.uoc.gr, waquar@nitsikkim.ac.in

Introduction

- Role of combining prosodic variables with the existing acoustic features in the context of children's speech recognition under mismatched conditions has been explored.
- The prosodic variables considered here are loudness, voice-intensity and voice-probability.
- The explored acoustic features are Mel-frequency cepstral coefficients (MFCC) and perceptual linear prediction cepstral coefficients (PLPCC).
- An analysis presented here shows that, for the same textual content, the considered prosodic variables exhibit very similar contours for adults' and children's speech.
- At the same time, the contours differ a lot when the context is different.
- Therefore, inclusion of prosodic information reduces the inter-speaker differences and increases the class discrimination.
- This improves the recognition performance in the case of children's mismatched ASR.
- Further improvements are obtained by projecting the feature vectors to a lower-dimensional subspace.
- This is experimentally verified using deep neural network (DNN) based automatic speech recognition system.
- On combining MFCC (PLPCC) and prosodic features, a relative improvement of 16% (14%) is noted on decoding children's speech using adult data trained DNN models.

Motivation

- Inclusion of prosodic information is first explored on the connected digit recognition task.
- The details of the speech corpus used for developing the digit recognition system are as follows:
  - The training and the test data for the connected digit recognition was obtained from the TIDIGITS database.
  - The age of the adult speakers contributing to this database varies from 17 to 70 years.
  - The child speakers, on the other hand, belong to an age group of 6 to 15 years.
  - A train set comprising of 5.3 hours of speech data from 197 adult male/female speaker was created from this database.
  - Two different test sets were derived for testing.
  - The first test set was composed of 1.6 hours speech data from 81 adult speakers.
  - The other test set comprised of 1.9 hours speech data from 49 children.
- The specifications of the connected digit recognition system are as follows:
  - Speech data was first analyzed into short-time frames using overlapping Hamming windows of length 20 ms with frame rate of 100 Hz.
  - A 23-channel Mel-filterbank was employed to compute the 13-dimensional base MFCC features.
  - This was followed by time-splicing of the base features considering a context size of 9, i.e., ± 4 frames.
  - The dimensionality of the resulting time-spliced features was then reduced to 40 using linear discriminant analysis (LDA).
  - Further de-correlation of the feature vectors was done through maximum likelihood linear transform (MLLT).
  - Mean and variance normalization (MVN) was also performed.
  - The 11 digits (0-9 and 'OH') were modeled as whole words using continuous density hidden Markov models (HMM) employing 3 states per word including silence.
  - Each HMM state, in turn, was modeled using 6 diagonal-covariance Gaussian mixture models (GMM).
  - An equillikely wordnet was employed during testing.
  - The metric used to measure the recognition performance is word error rate (WER).
- The WERs for the connected digit recognition system with respect two adults' and children's speech test sets are given in following table.

Test set	WER (in %)	
	MFCC	MFCC+Prosody
Adult	1.65	1.91
Child	9.17	6.93

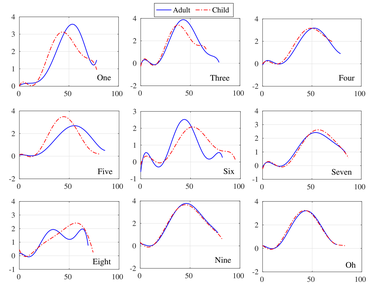
- The WER is noted to reduce for children's speech by the inclusion of prosodic information.
- For children's speech test set, the WER for each of the eleven digits with and without the inclusion of prosodic variables are enlisted separately in the following table.

WER (in %)					
Digit	MFCC	MFCC + Prosody	Digit	MFCC	MFCC + Prosody
One	5.40	7.10	Six	6.00	**3.80**
Two	4.90	**1.60**	Seven	9.60	**5.60**
Three	7.90	**7.20**	Eight	18.80	20.60
Four	6.30	**4.40**	Nine	1.00	**0.50**
Five	16.00	**5.40**	Oh	5.40	**4.00**

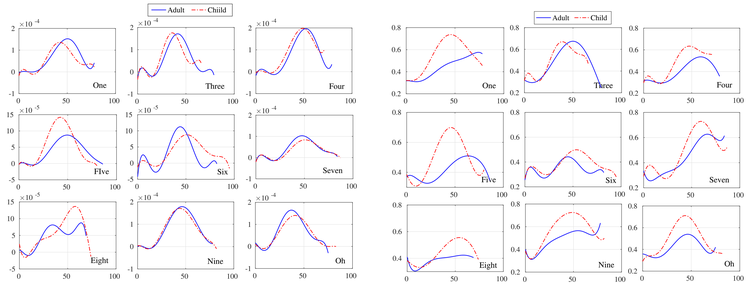
- The WER is noted to reduce for most of the digits by including prosodic variables.
- For majority of the digits, the reduction in WER is very large. The WERs for all those cases have been presented in bold.
- On the other hand, the WER is found to increase for digits *one* and *eight*.
- These observations motivated us to further explore the role of prosodic features in children's mismatched ASR.

Analyzing the cause of improved recognition performance

- In order to develop ASR systems, given the training data, relevant front-end features are extracted first.
- If the front-end features are such that the discrimination among the classes is more, then better performance will be obtained.
- In other words, the front-end features should be such that the *within-class differences* are minimal and the *between-class differences* are large.
- Using this reasoning, we studied the nature of explored prosodic variables for each of the classes.
- For any given digit (class), the deviation in the value of the prosodic variables for adult and child speakers should be very small.
- At the same time, the nature should be dissimilar for those classes where adding prosodic information did not help.
- In order to study the nature of any particular prosodic variable, smooth contours were derived for each of the classes as follows:
  - Forty isolated utterances of a given class, say digit one, were selected at random from the database.
  - Twenty of those were collected from adults while remaining twenty were from child speakers.
  - Next, the prosodic variables were computed for each of the utterances from the adult speakers.
  - The mean value for each of the frames was then computed using all the twenty examples.
  - In a similar manner, the mean was computed using the examples from children.
  - Finally, a seventh-order polynomial function was fitted over the mean data to derive a smooth contour.
  - These steps were repeated for each of the classes and each of the prosodic variable kinds.



- The smooth contours for mean loudness are as shown above.
- The smooth contours for mean intensity and voice probability are shown in the following figures.



- For each of the classes, the blue (solid) curves are for the case when data from adult speakers is used. The red (dash-dot) curves are obtained using data from children.
- The smooth contours for the mean of the explored prosodic variables derived using adults' and children's speech are very similar for those cases where the WERs have reduced.
- The WER had increased significantly for digits *one* and *eight* and the contours for adult and child speakers are also starkly different in those cases.
- For two different digits, the prosodic contours do not look similar thereby enhancing the inter-class differences.

Continuous speech recognition task

- The role of prosodic features in the context of continuous speech recognition task was explored next.
- For experimental evaluations, ASR systems were developed on the 15.5 hours adults' speech data from WSJCAM0 British English speech corpus.
- There are a total of 7861 utterances from 92 adult (male/female) speakers with approximately 90 sentences per speaker in this train set.
- For mismatched testing, the children's speech test set of the PF-STAR British English speech database was used.
- This test set contains 1.1 hours of speech data from 60 child speakers with a total of 5067 words.
- The experimental evaluations were performed on wideband speech.
- For computing the MFCC/PLPCC feature vectors, the earlier described steps were followed with a difference that a 40-channel Mel-filterbank was used.
- To boost the robustness towards speaker variations, feature-space maximum likelihood linear regression (fMLLR) was employed for normalization.
- The observation probabilities for the HMM states were generated using the GMM and DNN.
- Cross-word triphone models consisting of a 3-states HMM with 8 diagonal covariance Gaussian components per state were used in the case of GMM-HMM-based ASR system.
- Decision tree-based state tying was performed with the maximum number of tied-states (senones) being fixed at 2000.
- While learning the DNN-HMM-based ASR system, the fMLLR-normalized feature vectors were time-spliced once again considering a context size of 9.
- The number of hidden layers was chosen as 8 with each layer consisting of 1024 hidden nodes.
- The nonlinearity in the hidden layers was modeled using the *tanh* function.
- The initial learning rate for training the DNN-HMM parameters was set at 0.015 which was reduced to 0.002 after 20 epochs and extra 10 epochs of training were employed.
- The minibatch size for neural net training was selected as 512.
- For decoding the children's speech test set, a domain-specific 1.5k bigram language model (LM) was employed.
- The out-of-vocabulary (OOV) rate and perplexity of the employed bigram LM with respect to the children's test set are 1.20% and 95.8, respectively.
- A lexicon of 1, 969 words including the pronunciation variations was employed.
- The WERs for the children's speech test, with and without the inclusion of prosodic information are given in the following table.
- Inclusion of prosodic variables results in significant reduction in WERs for both GMM and DNN-based ASR systems.
- Further improvements were obtained by projecting the data to lower-dimensional subspace using heteroscedastic linear discriminant analysis (HLDA).

Explored Acoustic Model	Acoustic Feature Kind	WER (in %)		
		Baseline	+ Prosody	+ Prosody + HLDA
GMM	MFCC	32.69	25.81	**20.63**
	PLPCC	33.21	26.96	**21.35**
DNN	MFCC	19.68	16.66	**12.73**
	PLPCC	20.16	17.51	**13.28**

- Another DNN-HMM-based ASR system was developed by pooling together speech data from both adult as well as children train sets.
- The children's speech train set derived from PF-STAR consisted of 8.3 hours of speech data from 122 children.
- Only MFCC features were used in this case.
- The WERs for children's speech test set with respect to this new ASR system are given in the following table.
- Inclusion of prosodic features is observed to improve the recognition performance in this case as well.
- Low-rank feature projection through HLDA results in added reductions in WERs.

Acoustic Model	WER (in %)		
	Baseline	+ Prosody	+ Prosody + HLDA
DNN	11.47	9.98	8.82

Conclusion

- The effectiveness of combining prosodic variables with two of the dominant acoustic features in the context of children's speech recognition using adult data trained system is studied in this work.
- On combining the prosodic variables with MFCC/PLPCC features, significant reductions in WER are noted.
- In order to further improve the system performance, low-rank feature projection is also explored. Additive reductions in WERs are obtained by low-rank feature projection