SIP (2016), vol. 5, e15, page 1 of 13 © The Authors, 2016.

doi:10.1017/ATSIP.2016.16

ORIGINAL PAPER

Adaptive feature truncation to address acoustic mismatch in automatic recognition of children's speech

SHWETA GHAI AND ROHIT SINHA

An algorithm for adaptive Mel frequency cepstral coefficients (MFCC) feature truncation is proposed to improve automatic speech recognition (ASR) performance under acoustically mismatched conditions. Using the relationship found between MFCC base feature truncation and degree of acoustic mismatch of speech signals with respect to recognition models, the proposed algorithm performs utterance-specific MFCC feature truncation for test signals to address their acoustic mismatch in context of ASR. The proposed technique, without any prior knowledge about the speaker of the test utterance, gives 38% (on a connected-digit recognition task) and 36% (on a continuous speech recognition task) relative improvement over baseline in ASR performance for children's speech on models trained on adult speech, which is also found to be additive to improvements obtained with vocal tract length normalization and/or constrained maximum likelihood linear regression. The generality and effectiveness of the algorithm is also validated for automatic recognition of children's and adults' speech under matched and mismatched conditions.

Keywords: Speech recognition, Children's speech, Acoustic mismatch, MFCC features, Cepstral truncation

Received 27 August 2015; Accepted 14 July 2016

I. INTRODUCTION

The automatic recognition of children's speech under mismatched conditions, i.e. on models trained on adult speech, is a well-known challenging problem due to differences in speech of adult and child speakers. Many acoustic and linguistic characteristics of speech such as pitch, formant frequencies, average phone duration, speaking rate, pronunciation, accent, and grammar have already been noted to differ between adults and children [1, 2]. These differences result in degradation in automatic speech recognition (ASR) performance on children's speech under mismatched conditions [3, 4]. Apart from this, ASR performance on children's speech is significantly inferior to that for adults under matched conditions [1, 5–7]. This is attributed to the higher inter- and intra-speaker acoustic variability in children's speech relative to adults' speech [2]. Various model adaptation and speaker normalization techniques reported in the literature for addressing speaker differences have been investigated for improving ASR performance on children's speech. The foremost include vocal tract length normalization (VTLN) [8, 9] speaker adaptation techniques such as maximum *a posteriori* and maximum likelihood linear regression (MLLR) adaptations [8], constrained MLLR (CMLLR) adaptation [8, 10], speaker adaptive training (SAT) [10], constrained MLLR-based speaker normalization [11], and their combinations [8]. Significant improvements have been reported in ASR performance on children's speech under mismatched conditions using each of these speaker adaptation methods.

It is already known that children's speech has higher fundamental frequency and formant frequencies in comparison with those of adult speech [2]. A few studies have found improvement in ASR performance on children's speech with reduction of pitch of the signals using Mel frequency cepstral coefficients (MFCC) [12]. Improvement in ASR performance for children's speech using models trained on adult speech with pitch normalization is further supported by the results and observations already reported in the literature. In [13], pitch-adaptive MFCC features have been shown to improve adult ASR performance on models trained on adult speech, particularly for female speakers, on large vocabulary ASR tasks. Also, improvement in phone classification performance has been obtained with pitch-dependent normalization of the Mel cepstrum [14]. However, since in the MFCC features the pitch information is not captured but rather smoothed out by the filterbank, thus reducing the speaker dependence, one would expect the performance to be rather insensitive to pitch variations among speakers.

Department of Electronics and Electrical Engineering, Indian Institute of Technology Guwahati, Guwahati-781039, Assam, India. Phone: +1 404 785 9340

Corresponding author:
S. Ghai
Email: shweta.ghai@emory.edu

Table 1. Details of the speech corpora used for automatic speech recognition experiments.

	Speech corpus								
	TIDIGITS					WSJCAMo		PFSTAR	
Recognition task	Connected-digit					Continuous speech		Continuous speech	
Sampling freq.	8 kHz					8 kHz		8 kHz	
Language	American English					British English		British English	
Data set	ADtr	ADts	CHtr	CHts1	CHts2	CAMtr	CAMts	PFtr	PFts
Purpose	Training	Testing	Training	Testing	Testing	Training	Testing	Training	Testing
Speaker type	Adults	Adults	Children	Children	Children	Adults	Adults	Children	Children
No. of speakers	197	81	64	101	49	92	20	122	60
No. of words	35 566	10 813	14 725	25 525	10 800	132 778	5320	24 208	5067
Data (in h)	5.3	1.6	2.5	4.4	1.9	15.5	0.6	4.8	1.1
Language model	Equi-probable Wordnet					5 k word bigram		1.5 k word bigram	

Motivated by the results presented in [12], in our previous work [15] we studied the effect of pitch variations on MFCC features. Based on the analysis in that study, the degradation in ASR performance for children's speech on models trained on adult speech is attributed to the increase in variance in the higher dimensions of the MFCC feature space for children's speech due to higher pitch, while the variances of the lower dimensions were comparable with respect to those for adult speech. Although some studies in the literature have reported that higher-order MFCCs correlate with speech source information in general [16], the degree and nature of selective impact of pitch on the higher dimensions of MFCC feature vectors, as seen in multi-mixture hidden Markov model (HMM)-based speech recognition models, has not been explicitly demonstrated in the literature. Therefore, in this paper, we conduct a detailed experimental analysis illustrating the behavior of different coefficients of MFCC features with respect to the multi-mixture multi-state ASR model distributions, before and after explicit pitch normalization of speech signals.

In addition, an automatic algorithm for utterance-specific MFCC feature truncation is also proposed that does not require any prior knowledge about the test utterance being spoken by an adult or a child speaker for improving ASR for children's speech on acoustically mismatched models. In order to verify the generality of the proposed algorithm, the recognition results are evaluated for both children's and adults' test data on matched as well as mismatched ASR models. The effectiveness of the proposed algorithm is validated in combination with the already existing speaker normalization techniques such as VTLN and CMLLR as well.

II. MATERIAL AND METHODS

The ASR experiments are conducted on a connected-digit recognition task and a continuous speech recognition task in this work.

A) Database

For experiments on a connected-digit recognition task, speech data for both adults and children are taken from the TIDIGITS corpus [17]. For experiments on a continuous speech recognition task, speech data for adults are taken from the WSJCAMo Cambridge Read News corpus [18], while speech data for children are taken from the PFSTAR British English corpus [19]. The average pitch of all speech signals is estimated using the ESPS tool available in the Wavesurfer software package [20]. All speech data are resampled to 8 kHz in order to study the impact of our proposed algorithm under constrained data conditions like telephone quality speech. The details of the speech corpora are given in Table 1. To build a schildren's connected-digit recognition system, the "CHts1" data set, which contains all child speech data of TIDIGITS corpus, is split into two data sets "CHtr" (training set) and "CHts2" (test set). In order to have significant sample size while maintaining coverage across all age groups and gender in both training and test data sets, CHtr and CHts2 data sets are constructed such that they are disjoint in terms of speech data but not in terms of speakers. Out of total 101 speakers in CHts1 set, speech data from 52 speakers between 6 and 13 years of age are solely assigned to CHtr data set, while data from 37 speakers between the age group of 8 and 13 years are uniquely fed into CHts2 data set. However, speech data from remaining 12 speakers (three speakers each from 6–7 years, 8–9 years, 12–13 years and 14–15 years age groups) are distributed equally between CHtr and CHts2 data sets such that both data sets have some speech data from each of those 12 speakers. The different age groups of the child test sets are given in Tables 2 and 3.

B) Speech analysis method

Spectral analysis is carried out using a Hamming window of length 25 ms, frame rate of 100 Hz and pre-emphasis factor

Table 2. Division of different age groups of the child test set "CHts1'" used in the connected-digit recognition task.

	Age group (years)				
	6–7	8–9	10–11	12–13	14–15
No. of speakers (Males/females)	8 (5/3)	31 (12/19)	42 (27/15)	17 (5/12)	3 (1/2)
No. of utterances	615	2386	3231	1309	231

https://doi.org/10.1017/ATSIP.2016.16 Published online by Cambridge University Press

Table 3. Division of different age groups of the child test set "PFts" used in the continuous speech recognition task.

	Age group (years)				
	4–5	6–7	8–9	10–11	12–13
No. of speakers (Males/females)	1 (1/0)	12 (5/7)	16 (5/11)	28 (18/10)	3 (3/0)
No. of utterances	2	20	45	58	4

of 0.97. The 13-dimensional (13-D) MFCC base feature vector (C_0 to C_{12}) is computed using a 21-channel filter-bank using the HTK Toolkit [21]. In addition to the base features, their first- and second-order derivatives, computed over a span of five frames, are also appended to create a 39-D feature vector that is referred to as the "default" MFCC features. Cepstral mean subtraction is also applied to all features.

C) Recognition systems

The ASR systems used for the experimental evaluations for both the connected-digit recognition and the continuous speech recognition tasks are developed using the HTK Toolkit [21]. The word error rate (WER) is used to evaluate the speech recognition performance.

1) Connected-digit recognition task

For experiments on the connected-digit recognition task, the recognizer is developed following the setup described in [22]. The 11 digits (0–9 and OH) are modeled as whole-word HMMs using 16 states per word. Each state is a mixture of five diagonal-covariance Gaussian distributions with simple left-to-right moves without any skips over the states. A three-state model with six diagonal-covariance components is used for modeling silence. A single-state model with six diagonal-covariance components (allowing skip) is used for the short-pause model tied to the center state of the silence model. The adults' connected-digit recognition system is trained using "ADtr" and tested against "CHts1" and "ADts". The children's connected-digit recognition system is trained using "CHtr" and tested against "CHts2" and "ADts". For sake of comparison of ASR performances obtained for same child test set on models trained on "ADtr" and "CHtr" data sets, the adults' connected-digit recognition system is tested against "CHts2" as well. The baseline recognition performances (in WER) for adult and child test sets on the adults' connected-digit recognizer and the children's connected-digit recognizer are given in Table 4.

2) Continuous speech recognition task

The continuous speech recognition system is developed using cross-word tri-phone acoustic models along with decision-tree-based state tying as given by the HTK Toolkit [21]. The tri-phone models have three states with eight diagonal-covariance components for each state. A three-state model with 16 diagonal-covariance components is used for modeling silence, and a short-pause model (allowing skip) is constructed with all states tied to the silence model. Each component is modeled by a Gaussian density function. The adults' continuous speech recognition system is trained using "CAMtr" resulting in 2499 tied states after state tying, while the children's continuous speech recognition system is trained using "PFtr". To evaluate the ASR performance of the continuous speech recognizers on adult speech and child speech, "CAMts" and "PFts" data sets are used, respectively.

Table 4. Baseline ASR performances (in WER) for adult and child test sets on models trained on adult and child speech for both connected-digit and continuous speech recognition tasks.

		Test set Baseline performance (% WER)				
Recognition task	Training data set	ADts	CHts1	CHts2	CAMts	PFts
Connected-digit	ADtr	0.43	11.37	9.19	–	–
	CHtr	13.28	–	1.01	–	–
Continuous speech	CAMtr	–	–	–	9.92	56.34
	PFtr	–	–	–	68.36	12.41

The standard WSJo 5000-word closed non-verbalized vocabulary set and the standard MIT Lincoln Labs 5k Wall Street Journal bigram language model are used for recognition of "CAMts" with no out-of-vocabulary (OOV) words. For "PFts", a 1500-word non-verbalized vocabulary set and a 1.5k bigram language model trained using the transcripts of "PFtr" such that "PFts" has perplexity of 1.02% OOV are used. The pronunciations for all words are obtained from the British English Example Pronunciation (BEEP) dictionary [18]. The baseline recognition performances (in WER) for adult and child test sets on the adults' continuous speech recognition system, and the children's continuous speech recognizer are also given in Table 4.

The recognition performance for child test set (9.19% WER on connected-digit recognition task and 56.34% WER on continuous speech recognition task) is far worse than that obtained for adult test set (0.43% WER on connected-digit recognition task and 9.92% WER on continuous speech recognition task) on recognizers trained on adult speech. This is attributed to the large acoustic mismatch between the adult training and the child test sets and to the loss of spectral information for children's narrowband-filtered speech. The poor ASR performance for children's speech on matched acoustic models trained on child speech as well (1.01% WER on connected-digit recognition task and 12.41% WER on continuous speech recognition task) in comparison with that for adults' speech on matched models trained on adult speech (0.43% WER on connected-digit recognition task and 9.92% WER on continuous speech recognition task) can be attributed to greater inter-speaker variability among children than adults.

The trend observed in the ASR performances obtained from the above data sets and experimental setups is consistent with that already reported in the literatures [9, 23].

https://doi.org/10.1017/ATSIP.2016.16 Published online by Cambridge University Press

III. ANALYSIS

A) Effect of pitch on lower- and higher-order MFCC distributions

The effect of pitch differences between training and test data on the MFCC feature distribution is explored in a connected-digit recognition task to assess the impact of acoustic mismatch on recognition performance. The digits whose recognition models and features are most affected by the pitch mismatch are identified by determining the frequency of substitution, deletion, and insertion errors across different digit models from the ASR performances of "CHts1" on models trained with "ADtr".

Explicit pitch normalization of each child test signal is performed with respect to the pitch distribution of the adult training data set using a maximum likelihood (ML) grid search approach. To carry out the ML grid search, the pitch of each child test signal is first transformed to seven different pitch values ranging from 70 to 250 Hz in steps of 30 Hz using the pitch synchronous time scaling method [24]. This range of transformed pitch values was chosen based on the fact that the adult training data has a pitch distribution from 70 to 250 Hz. Given the various pitch-transformed versions within the specified range, the optimal value $\hat{p}$, to which the pitch of each signal is to be transformed to, is estimated as:

$$\hat{p} = \arg\max_{p} \Pr\left(X_i^p | \lambda, W_i\right), \quad (1)$$

where X_i^p is the feature for the ith utterance with pitch transformed to p, λ is the speech recognition model, and W_i is the transcription of the ith utterance. W_i is determined by initial recognition pass using original feature (i.e. without pitch transformation). The ML search is then performed over the original speech signal and its corresponding seven different pitch-transformed versions.

The recognition performances (in WER) obtained for CHts1 test set on models trained on ADtr using default 13-D base MFCC features with and without explicit pitch normalization of child speech and that using truncated 4-D base MFCC features are 11.37% WER, 9.64% WER, and 5.21% WER. For recognition, the base MFCC features are also appended with their corresponding first- and second-order derivatives. The frequency of substitution, deletion, and insertion errors in these ASR outputs are given in Table 5. It is shown that explicit pitch normalization results in substantial reduction in the number of substitution errors, deletion errors, and insertion errors. The deletion and insertion errors are mainly governed by the differences in speech rate in training and test data sets. Substitution errors are therefore used as the main criterion for evaluating performance. About 35% reduction is obtained in the substitution errors, which constitute 90% of total substitution errors in the original ASR system output on children's speech, while about 40% reduction is obtained in the substitution errors, which constitute 75% of total substitution errors in the original ASR system output on children's speech, after pitch normalization. The underlined numbers highlight examples from the top 10 highest occurring substitution errors (together constituting 67% of total substitution errors) in the original ASR system output obtained using default MFCC features, which were improved by more than 70% after pitch normalization. All substitution errors were verified, but due to space limitations analysis for only two of the underlined cases, which have the highest frequency in the original ASR system output, i.e. original digit "FIVE" recognized as "NINE" and original digit "OH" recognized as "TWO" is shown.

The higher-order coefficients C_{11} and C_{12} are extracted from the 13-D base MFCC feature vectors for 200 digit "FIVE" utterances and 200 digit "NINE" utterances from the original "ADtr" corpus. Similarly, the C_{11} and C_{12} coefficients are also extracted for 200 digit "FIVE" utterances from the original "CHts1" and 200 digit "FIVE" utterances from the explicitly pitch-normalized "CHts1" corpus. The distributions of the C_{11} and C_{12} coefficients extracted from MFCC features of digit "FIVE" utterances from original "ADtr" are shown in yellow, from original "CHts1" in blue, from explicitly pitch normalized "CHts1" in magenta and of digit "NINE" utterances from original "ADtr" are shown in black in Fig. 1. For ease of comparison, distributions of the C_{11} and C_{12} coefficients extracted for all these four different data sets are shown in both subplots of Fig. 1. The multivariate Gaussian distributions estimated using the means and variances of the corresponding coefficients and weighted by the corresponding mixture weights for each mixture in each state of the digit "FIVE" and digit "NINE" recognition models trained on "ADtr" are then computed and are also shown in gray scale in Fig. 1. These distributions are similarly obtained for the lower-order coefficients C_1 and C_2 for the same digit utterances and recognition models, which are shown in Fig. 2 using same color coding as given in Fig. 1.

Scatterplots given in black and yellow show the examples of perfect matching of distributions of MFCCs of adult utterances for each of the two digits with the Gaussian distributions for the corresponding MFCCs of the respective digit models shown in gray scale. As shown in Fig. 1, the spread of the C_{11}–C_{12} distribution for digit "FIVE" utterances from the "CHts1" test set (given in blue) is significantly reduced after explicit pitch normalization of the test set (given in magenta), which explains the reduction in variance of those coefficients observed in [15]. The C_{11}–C_{12} distribution for digit "FIVE" utterances from the "CHts1" test set is better mapped for digit "FIVE" after explicit pitch normalization of the test set. On the other hand, no significant change is observed in the distribution for the lower-order C_1 and C_2 coefficients after pitch normalization, as evident in Fig. 2.

Similar observations hold for the C_{11}–C_{12} and C_1–C_2 distributions of digit "OH" and digit "TWO" utterances and recognition models shown in Figs 3 and 4, respectively. These observations indicate that variations in the pitch of the signals predominantly affect the higher-order MFCCs, which typically carry less linguistically relevant information. Differences in pitch between training and test speech lead to pattern mismatch between the higher-order MFCCs,

https://doi.org/10.1017/ATSIP.2016.16 Published online by Cambridge University Press

Table 5. Frequency of substitution, deletion, and insertion errors as resulting by evaluating the outputs of different ASR systems for "CHts1" on models trained on "ADtr".

Recognized \ Original		Substitution errors										
		/Oh/	/One/	/Two/	/Three/	/Four/	/Five/	/Six/	/Seven/	/Eight/	/Nine/	Insertion errors
/Oh/	Original	–	6	14	34	**78**	40	12	**79**	281	9	75
	F_o_normalized	–	6	47	58	**48**	8	6	**55**	299	11	79
	4-D base MFCC	–	2	27	51	**6**	1	2	**1**	205	0	36
/One/	Original	4	–	0	3	112	3	1	76	2	4	12
	F_o_normalized	6	–	1	1	69	0	0	17	1	3	7
	4-D base MFCC	4	–	0	0	67	0	0	5	0	1	13
/Two/	Original	**143**	4	–	**102**	4	3	15	**84**	35	4	85
	F_o_normalized	**24**	0	–	**40**	1	0	5	**17**	1	1	17
	4-D base MFCC	**9**	0	–	**26**	0	0	0	**0**	3	0	16
/Three/	Original	1	2	4	–	0	0	0	0	2	2	6
	F_o_normalized	0	4	3	–	1	0	0	0	1	2	2
	4-D base MFCC	0	4	3	–	0	0	0	0	0	1	0
/Four/	Original	1	11	1	0	–	1	0	6	0	0	11
	F_o_normalized	4	9	3	0	–	0	0	3	0	0	10
	4-D base MFCC	9	10	0	0	–	0	0	1	0	0	3
/Five/	Original	14	5	1	1	14	–	1	66	7	11	61
	F_o_normalized	28	6	6	2	39	–	5	91	13	12	96
	4-D base MFCC	31	3	1	0	85	–	1	6	4	2	7
/Six/	Original	0	0	9	0	0	1	–	1	0	0	4
	F_o_normalized	2	0	8	0	0	0	–	0	0	0	3
	4-D base MFCC	3	0	7	3	0	0	–	0	2	0	18
/Seven/	Original	7	0	2	7	0	1	0	–	0	0	1
	F_o_normalized	11	0	4	9	0	0	1	–	0	0	3
	4-D base MFCC	10	0	20	7	0	1	2	–	0	0	3
/Eight/	Original	38	10	35	144	2	1	58	52	–	1	19
	F_o_normalized	31	4	72	143	0	0	49	29	–	0	27
	4-D base MFCC	17	3	42	96	0	0	24	8	–	1	29
/Nine/	Original	180	27	1	4	5	**278**	0	17	11	–	49
	F_o_normalized	123	25	11	4	3	**64**	0	21	8	–	26
	4-D base MFCC	127	43	15	14	0	**74**	0	7	15	–	21
Deletion errors	Original	69	25	19	44	23	23	32	82	57	10	–
	F_o_normalized	40	14	29	20	6	4	3	15	27	8	
	4-D base MFCC	25	3	15	12	4	0	0	0	10	2	

The ASR performances are computed on outputs of ASR systems making use of different set of features including: 13-D base MFCC features of children's original and explicitly pitch normalized test signals and 4-D base MFCC features of children's original test signals. For recognition, the base MFCC features are also appended with their corresponding first- and second-order derivatives. The bold numbers highlight examples from the top 10 highest occurring substitution errors (together constituting 67% of total substitution errors) in the original ASR system output obtained using default MFCC features, which were improved by more than 70% after using 4-D base MFCC features, while the underlined numbers highlight examples, which were improved by more than 70% after pitch normalization.

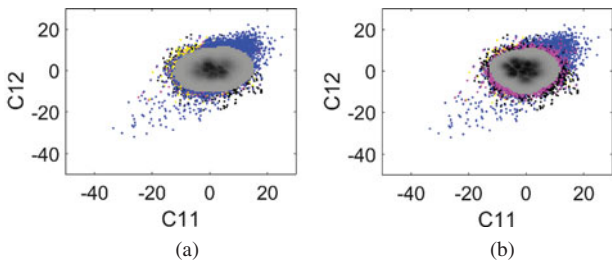

Fig. 1. Scatter plots showing distributions of C_{11} and C_{12} coefficients of digit "FIVE" utterances from (a) original "CHts1" (in blue) along with Gaussian distributions (in gray scale) of those coefficients in digit "NINE" models trained with "ADtr", and (b) explicitly pitch normalized "CHts1" (in magenta) along with Gaussian distributions (in gray scale) of those coefficients in digit "FIVE" models trained with "ADtr". The spread of the C_{11}–C_{12} distribution for digit "FIVE" utterances from "CHts1" test set is significantly reduced and is better mapped for digit "FIVE" models after explicit pitch normalization of the test set.

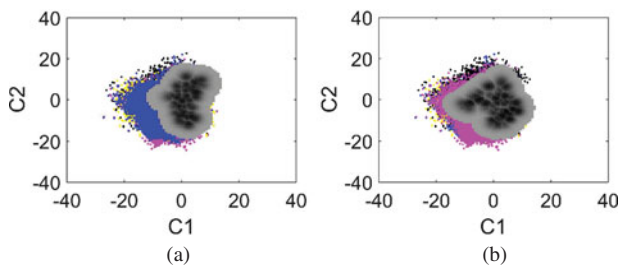

Fig. 2. Scatter plots showing distributions of C_1 and C_2 coefficients of digit "FIVE" utterances from (a) original "CHts1" (in blue) along with Gaussian distributions (in gray scale) of those coefficients in digit "NINE" models trained with "ADtr", and (b) explicitly pitch normalized "CHts1" (in magenta) along with Gaussian distributions (in gray scale) of those coefficients in digit "FIVE" models trained with "ADtr". No significant change is observed in the distribution of lower-order C_1 and C_2 coefficients for digit "FIVE" utterances from "CHts1" test set after pitch normalization.

https://doi.org/10.1017/ATSIP.2016.16 Published online by Cambridge University Press

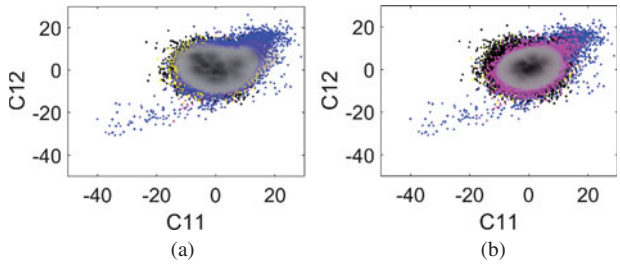

Fig. 3. Scatter plots showing distributions of C_{11} and C_{12} coefficients of digit "OH" utterances from (a) original "CHts1" (in blue) along with Gaussian distributions (in gray scale) of those coefficients in digit "TWO" models trained with "ADtr", and (b) explicitly pitch normalized "CHts1" (in magenta) along with Gaussian distributions (in gray scale) of those coefficients in digit "OH" models trained with "ADtr". The spread of the C_{11}–C_{12} distribution for digit "OH" utterances from "CHts1" test set is significantly reduced and is better mapped for digit "OH" models after explicit pitch normalization of the test set.

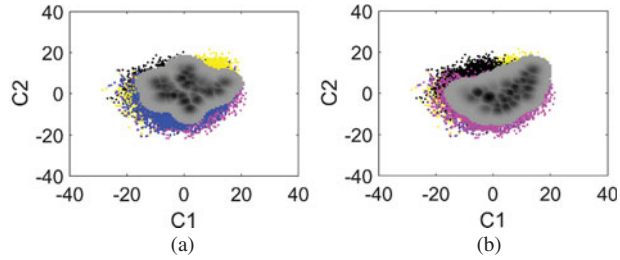

Fig. 4. Scatter plots showing distributions of C_1 and C_2 coefficients of digit "OH" utterances from (a) original "CHts1" (in blue) along with Gaussian distributions (in gray scale) of those coefficients in digit "TWO" models trained with "ADtr", and (b) explicitly pitch normalized "CHts1" (in magenta) along with Gaussian distributions (in gray scale) of those coefficients in digit "OH" models trained with "ADtr". No significant change is observed in the distribution of lower-order C_1 and C_2 coefficients for digit "OH" utterances from "CHts1" test set after pitch normalization.

resulting in poor ASR performance under mismatched conditions.

B) Effect of higher-order MFCCs on Mel spectrum

To better comprehend the impact of higher-order MFCCs on the MFCC feature representation, the spectra corresponding to MFCC features of different feature lengths are illustrated. The plots of the smooth spectra corresponding to various truncated base feature lengths including C_0 for the stable portion of vowel /*iy*/ with average pitch value 300 Hz along with its linear discrete Fourier transform (DFT) spectrum are shown in Fig. 5. The frequency of the first two pitch harmonics in the linear DFT spectrum in Fig. 5(b) exactly matches that of the two harmonics in the lower-frequency region of the smoothed. Mel spectrum in Fig. 5(a) in case of default MFCC base feature length.

Truncation in quefrency domain is equivalent to convolving the spectral envelope of the MFCC filterbank with the spectrum of the truncation window in frequency domain. As the degree of truncation increases, the bandwidth of the truncation window spectrum increases resulting in greater smoothening of the pitch harmonics and the spectral peaks (formants) in the spectrum. Therefore, with increasing cepstral feature truncation, the two harmonics in the lower-frequency region of the smoothed Mel spectrum in Fig. 5(a) are getting smoothed out.

Thus, excluding higher-order MFCCs for speech recognition under mismatched conditions is hypothesized to reduce the gross acoustic mismatch between training and test data by minimizing the effect of pitch differences, VTL differences (one of the foremost source of acoustic mismatch between adult and child speech [25]) and any other sources of speaker-specific acoustic mismatch which induce rapidly varying changes in spectra.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

A) Effect of MFCC feature truncation on ASR performance

In order to study the extent of acoustic mismatch captured by higher-order MFCCs, the effect of exclusion of the higher-order coefficients from the MFCC feature vector is tested by truncating the MFCC base features for the children's test data from 13 (C_0-C_{12}) down to 3 (C_0-C_2) in steps of 1. For recognition, the various truncated MFCC base features are augmented with their corresponding first- and second-order derivatives. The resulting feature set is then decoded using ASR models of the same dimensionality extracted from the baseline 39-D models. The recognition performance for "CHts1" for different dimensions of the truncated test features on corresponding models trained on "ADtr" with matching feature dimensions are given in Table 6. The table also shows the pitch group-wise breakup of performances corresponding to each of the truncations. The recognition performance for "CHts1" improves consistently with feature truncation. The best relative improvement of about 54% is obtained over the baseline for "CHts1" with 4-D base MFCC features. The age group-wise breakup of this best recognition performance obtained for "CHts1" using 4-D base MFCC features is given in Table 7.

This improvement in ASR performance with feature truncation beyond the default feature length is an outcome of consistent reduction in all three categories of recognition errors as given in Table 5. The bold numbers in Table 5 highlight examples from the top 10 highest occurring substitution errors (together constituting 67% of total substitution errors) in the original ASR system output on children's speech obtained using default MFCC features, which are improved by more than 70% after using 4-D base MFCC features.

To validate the improvement in performance with increasing MFCC feature truncation on a continuous speech recognition task, the recognition performance for "PFts" is evaluated on the corresponding models trained on "CAMtr" with matching feature dimensions, as given in Table 8. When the higher-order MFCCs are excluded from the feature vector, improvement is obtained in recognition performance on a continuous speech recognition task as

https://doi.org/10.1017/ATSIP.2016.16 Published online by Cambridge University Press

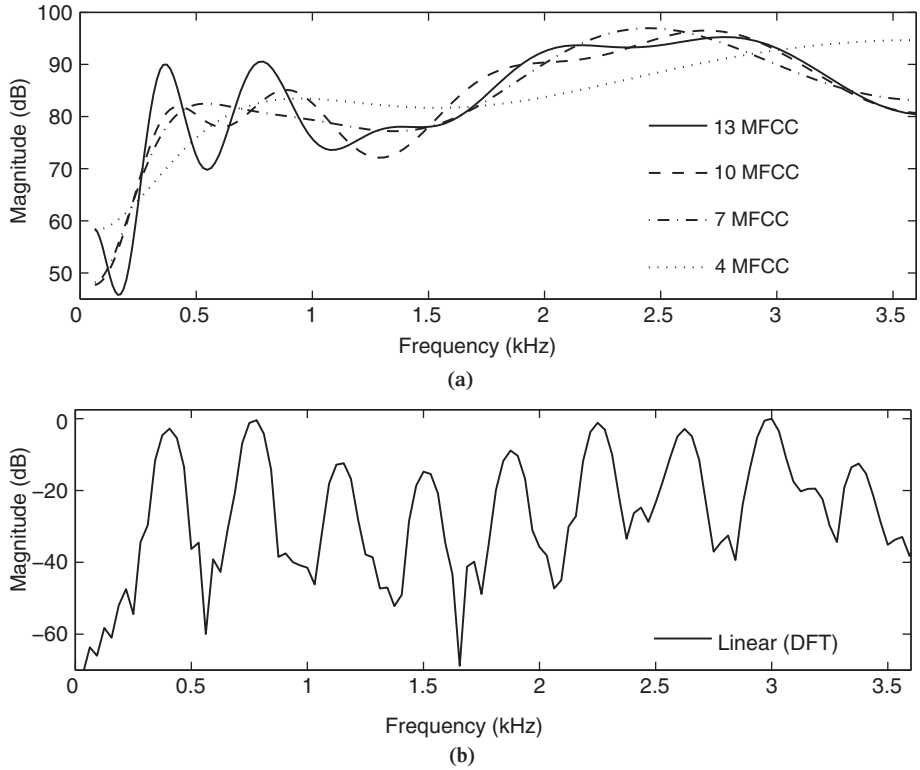

Fig. 5. Plots of (a) smoothed spectra corresponding to the base MFCC features of different dimensions along with corresponding. (b) Linear DFT spectrum for a frame of vowel /*iy*/ having the average pitch value of 300 Hz.

Table 6. Performance of "CHts1" on models trained with "ADtr" for various truncations of the base MFCC features along with its pitch group-wise breakup.

	WER (%)			
MFCC base feature length	All F_0 values (7772)	$F_0 < 250$ Hz (5224)	250 Hz $\leq$ $F_0 < 300$ Hz (2346)	$F_0 \geq 300$ Hz (202)
Default (C_0–C_{12})	11.37	6.54	17.47	39.03
C_0–C_{11}	11.20	6.81	16.80	35.81
C_0–C_{10}	11.38	7.35	16.55	33.58
C_0–C_9	10.71	7.03	15.38	31.60
C_0–C_8	9.03	6.03	13.00	25.29
C_0–C_7	7.80	5.35	10.77	23.30
C_0–C_6	6.77	4.72	9.42	18.09
C_0–C_5	6.21	4.25	8.62	18.09
C_0–C_4	6.03	4.55	7.93	14.25
C_0–C_3	**5.21**	**4.20**	**6.33**	**12.64**
C_0–C_2	5.47	4.51	6.34	14.75

The truncated base MFCC features are also appended with their corresponding first- and second-order derivatives for recognition. The quantity in bracket gives the number of utterances corresponding to that pitch group. The bold numbers highlight the best performances obtained for each pitch group by increasing the truncation of the base MFCC feature beyond the default length of 13 up to 3.

Table 7. Age group-wise breakup of the best recognition performance obtained for "CHts1" on models trained with "ADtr" using 4-D base MFCC features.

	WER (%) Age group (years)					
MFCC base feature length	All (7772)	6–7 (615)	8–9 (2386)	10–11 (3231)	12–13 (1309)	14–15 (231)
Default (C_0–C_{12})	11.37	23.40	19.46	7.00	3.72	0.26
C_0–C_3	5.21	8.28	8.48	3.71	2.21	1.05

The truncated base MFCC features are also appended with their corresponding first- and second-order derivatives for recognition. The quantity in bracket gives the number of utterances corresponding to that age group.

well. The best relative improvement of about 38% is obtained over the baseline for "PFts" with 6-D base MFCC features. The age group-wise breakup of this best recognition performance obtained for "PFts" using 6-D base MFCC features, given in Table 9, shows consistent improvements across all age groups.

These results verify our hypothesis that a large degree of degradation in ASR performance on children's speech under the mismatched conditions is mainly caused by the acoustically mismatched information present in the higher-order coefficients of the MFCC feature vector.

B) Relation between MFCC feature truncation and acoustic mismatch

Although truncating MFCC base features beyond the default length of 13 helps in reducing the gross acoustic mismatch between the child test set and the models trained

https://doi.org/10.1017/ATSIP.2016.16 Published online by Cambridge University Press

Table 8. Performance of "PFts" on models trained with "CAMtr" for various truncations of the base MFCC features along with its pitch group-wise breakup.

	WER (%)			
MFCC base feature length	All F_0 values (7772)	$F_0 < 250$ Hz (5224)	250 Hz $\leq$ $F_0 < 300$ Hz (2346)	$F_0 \geq 300$ Hz (202)
Default (C_0–C_{12})	56.34	33.05	77.25	102.69
C_0–C_{11}	52.10	31.23	69.20	101.71
C_0–C_{10}	48.39	27.89	66.19	92.18
C_0–C_9	44.72	25.61	61.35	85.33
C_0–C_8	42.19	23.37	59.28	78.73
C_0–C_7	39.89	22.30	54.59	80.20
C_0–C_6	39.02	21.69	54.05	76.04
C_0–C_5	**35.13**	**19.45**	**48.03**	72.13
C_0–C_4	38.25	22.34	51.83	73.35
C_0–C_3	41.50	26.63	54.05	75.06
C_0–C_2	40.62	27.55	51.97	**68.46**

The truncated base MFCC features are also appended with their corresponding first- and second-order derivatives for recognition. The quantity in bracket gives the number of utterances corresponding to that pitch group. The bold numbers highlight the best performances obtained for each pitch group by increasing the truncation of the base MFCC feature beyond the default length of 13 up to 3. The numbers indicate that greater truncation is chosen for signals of higher pitch group.

Table 9. Age group-wise breakup of the best recognition performance obtained for "PFts" on models trained with "CAMtr" using 6-D base MFCC features.

	WER (%) Age group (years)					
MFCC base feature length	All (129)	4–5 (2)	6–7 (20)	8–9 (45)	10–11 (58)	12–13 (4)
Default (C_0–C_{12})	56.34	249.33	107.35	77.51	20.79	58.64
C_0–C_5	35.13	244.00	81.55	39.13	10.50	53.40

The truncated base MFCC features are also appended with their corresponding first- and second-order derivatives for recognition. The quantity in bracket gives the number of utterances corresponding to that age group.

on adult speech, for each test signal the degree of acoustic mismatch with respect to the models would be different. Therefore, we investigate the correspondence between MFCC feature length and degree of acoustic mismatch in speech recognition.

Due to reduction in the dimensionality of feature vectors, the likelihoods of test features increase monotonically with increasing truncation. Thus, the ML criterion cannot be simply used to determine the appropriate truncation for a test feature without using appropriate penalties. Since we have hypothesized that increased MFCC feature truncation reduces the acoustic mismatch mainly due to pitch differences and differences in formant frequencies, we explore the appropriate MFCC base feature truncations for the test signals based on their average pitch values and their optimal VTLN warp factors on default MFCC features.

The speech signals of child test sets are divided into different groups based on their average pitch values. The pitch group-wise performances of "CHts1" and "PFts" for different dimensions of the truncated test features on corresponding models trained on adult speech with matching feature dimensions are also given in Tables 6 and 8. On both tasks, the best ASR performances correspond to the same MFCC base feature truncation for different pitch groups. Therefore, pitch cannot be used for determining the appropriate MFCC feature truncation for a signal. Although the acoustic mismatch due to pitch differences is also addressed by increased MFCC feature truncation, it is not large enough to quantify the gross acoustic mismatch of the signals with respect to the models.

Among the various other sources of acoustic mismatch, the VTL differences are the foremost source of mismatch. VTLN provides an easy quantification of the acoustic mismatch in terms of the frequency warp factor. To investigate appropriate MFCC base feature truncations for children test signals, we now explore optimal VTLN warp factors. The VTLN warp factors for all speech signals of both child test sets are estimated on an utterance-by-utterance basis with respect to the corresponding 39-D baseline models trained on adult speech by carrying out a ML grid search among 13 frequency warp factors (α) ranging from 0.88–1.12 in steps of 0.02 as described in [26]. This range of VTLN warp factors is chosen in order to study the relation between degree of MFCC feature truncation and acoustic mismatch (as quantified by VTLN warp factors) in speech recognition of child speech on models trained on adult speech with the default feature and recognition model settings that are typically used for recognition of adult speech under matched conditions. Piece-wise linear frequency warping of the filterbank, as supported in the HTK Toolkit [21], is used. The optimal warp factor $\hat{\alpha}$, is estimated as:

$$\hat{\alpha} = \arg\max_{\alpha} \Pr\left(X_i^{\alpha} | \lambda, W_i\right), \tag{2}$$

where X_i^{α} is the frequency warped feature for the ith utterance. λ is the speech recognition model, and W_i is the transcription of the ith utterance. W_i is determined by the initial recognition pass. Both test sets are then divided into different groups based on their optimal VTLN warp factors.

The VTLN warp factor-wise recognition performances of "PFts" are given in Table 10. There are not sufficient numbers of signals corresponding to each of the values of the VTLN warp factor in the child test sets. The recognition performances for "CAMts" are therefore also evaluated on models trained on "CAMtr" as given in Table 11.

While the ASR performance improves consistently with increased MFCC feature truncation for "PFts", there is a slight improvement over baseline with increased MFCC feature truncation up to the MFCC base feature length of 12 for "CAMts". The reduced feature truncation and small improvement for "CAMts" is attributed to the smaller degree of gross acoustic mismatch for adult test set with

https://doi.org/10.1017/ATSIP.2016.16 Published online by Cambridge University Press

Table 10. Performance of "PFts" on models trained on "CAMtr" for various truncations of the base MFCC features along with their VTLN warp factor-wise breakup.

	WER (%) VTLN warp factor values								
MFCC base feature length	All (129)	0.88 (91)	0.90 (24)	0.92 (7)	0.94 (1)	0.96 (2)	0.98 (2)	1.06 (1)	1.08 (1)
Default (C_0–C_{12})	56.34	64.07	40.97	26.32	25.00	**8.42**	94.44	42.86	5.13
C_0–C_{11}	52.10	58.70	38.95	23.08	25.00	10.53	93.52	42.86	5.13
C_0–C_{10}	48.39	54.76	34.21	22.67	25.00	**8.42**	93.52	42.86	5.13
C_0–C_9	44.72	50.21	31.48	22.67	25.00	9.47	98.15	28.57	5.13
C_0–C_8	42.19	47.27	28.96	22.27	25.00	**8.42**	99.07	30.95	5.13
C_0–C_7	39.89	44.87	28.15	16.60	25.00	**8.42**	90.74	26.19	5.13
C_0–C_6	39.02	43.99	27.35	16.19	13.64	**8.42**	93.52	21.43	5.13
C_0–C_5	**35.13**	**39.56**	**23.71**	**14.98**	13.64	9.47	88.89	26.19	2.56
C_0–C_4	38.25	42.25	29.26	16.60	22.73	9.47	90.74	23.81	2.56
C_0–C_3	41.50	45.62	35.22	15.38	**9.09**	12.63	86.11	21.43	2.56
C_0–C_2	40.62	44.22	33.70	21.05	**9.09**	25.26	**82.41**	**16.67**	**0.00**

The truncated base MFCC features are also appended with their corresponding first- and second-order derivatives for recognition. The quantity in bracket gives the number of utterances corresponding to that VTLN warp factor. The bold numbers highlight the best performances obtained for each group of signals corresponding to different VTLN warp factors by increasing the base MFCC feature truncation beyond default. The numbers indicate that greater truncation is chosen for signals having greater VTL differences.

Table 11. Performance of "CAMts" on models trained on "CAMtr" for various MFCC base feature truncations along with their VTLN warp factor-wise breakup.

	WER (%) VTLN warp factor values										
MFCC base feature length	All (314)	0.92 (8)	0.94 (14)	0.96 (39)	0.98 (79)	1.00 (70)	1.02 (13)	1.04 (46)	1.06 (8)	1.08 (33)	1.12 (4)
Default (C_0–C_{12})	9.92	6.31	8.87	13.76	8.12	9.82	13.71	**13.93**	4.93	**5.01**	9.76
C_0–C_{11}	**9.76**	6.31	**7.39**	13.30	**7.89**	9.42	13.31	14.55	4.93	5.18	9.76
C_0–C_{10}	10.28	**4.50**	8.37	13.15	8.82	10.06	**11.69**	14.68	10.56	5.87	9.76
C_0–C_9	9.92	5.41	7.88	**12.54**	9.13	**9.19**	**11.69**	14.55	**4.23**	6.22	9.76
C_0–C_8	10.60	5.41	9.36	13.46	9.82	9.50	12.90	15.68	5.63	6.56	**4.88**
C_0–C_7	11.02	5.41	10.84	13.91	10.21	10.30	12.10	15.81	6.34	6.56	7.32
C_0–C_6	11.86	5.41	9.85	15.29	10.90	11.66	12.10	17.31	6.34	6.56	7.32
C_0–C_5	12.95	7.21	12.81	16.97	11.60	12.38	13.71	18.70	**4.23**	7.94	9.76
C_0–C_4	15.08	7.21	15.76	18.96	13.77	14.54	14.11	22.08	9.86	8.64	7.32
C_0–C_3	21.52	13.51	24.14	27.68	20.42	20.69	16.94	28.23	13.38	14.51	17.07
C_0–C_2	29.15	24.32	33.99	37.77	27.38	27.16	25.00	37.39	16.90	21.42	14.63

The truncated test features also include their first- and second-order derivatives. The quantity in bracket gives the number of utterances corresponding to that VTLN warp factor. The bold numbers highlight the best performances obtained for each group of signals corresponding to different VTLN warp factors by increasing the base MFCC feature truncation beyond default. The numbers indicate slight improvement over baseline with increased MFCC feature truncation only up to MFCC base feature length of 12 due to smaller degree of gross acoustic mismatch between the test and training data set.

respect to the models trained on adult speech unlike the data set "PFts". However, for both child and adult test data sets, different truncations of MFCC base features are required for signals with different degrees of VTL differences. Greater truncation is chosen for signals having greater VTL differences. Child speech has greater acoustic mismatch with respect to ASR models trained on adult speech. As a result, larger improvements are obtained in speech recognition performance on children's speech, while no significant change is obtained for adults' speech using increased MFCC feature truncation.

Thus, increasing MFCC feature truncation helps to improve ASR performance under mismatched conditions by efficiently reducing significant gross acoustic mismatch between the test signals and the training set.

C) Proposed algorithm for adaptive MFCC feature truncation for ASR

We have already found that the appropriate MFCC base feature truncation increases as the degree of acoustic mismatch

https://doi.org/10.1017/ATSIP.2016.16 Published online by Cambridge University Press

increases between the test speech and the speech recognition models. Under matched and mismatched conditions, however, on the whole, less MFCC feature truncation is required for matched test speech, while greater truncation is required for mismatched test data. In order to automate the procedure of determining the appropriate MFCC feature truncation for a test signal, the test speech must first be categorized as matched or mismatched.

The frequency spectrum of child speech may require compression by a frequency warp factor of as low as 0.88 relative to adult speech. On the other hand, the frequency spectrum of adult speech would often need expansion by a frequency warp factor of as high as 1.12 relative to child speech. Thus, to categorize the test signal as belonging to a child speaker or an adult speaker, we compare the log likelihood for the test signal using the default MFCC feature set and the feature sets corresponding to different VTLN warp factors.

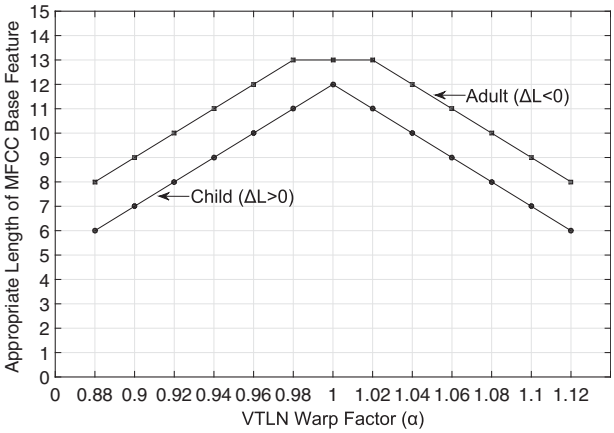

Fig. 6. Graph showing the relation proposed between appropriate length of MFCC base feature and VTLN warp factor for both adult and child test signals.

On models trained on adult speech, if the likelihood is greater for the feature set corresponding to VTLN warp factor of 0.88 than that corresponding to the default MFCC features, the test signal is categorized as belonging to a child speaker. However, on models trained on child speech, the test signal is categorized as adult speech if its likelihood is greater for the feature set corresponding to VTLN warp factor of 1.12 than that corresponding to the default MFCC features. For the sake of generality, based on the ASR performances obtained for adult and child speech corresponding to different VTLN warp factors using different MFCC feature lengths, we propose a piece-wise linear model for the relationship between MFCC base feature truncation for both adult and child test signals and their acoustic mismatch, as assessed by their VTLN warp factor ($\hat{\alpha}$), depicted in Fig. 6.

Once the test signal is categorized as belonging to adult or child, the appropriate MFCC base feature length is determined using the optimal VTLN warp factor ($\hat{\alpha}$). The flow diagram of the algorithm proposed for models trained on adult speech or child speech is shown in Fig. 7.

The recognition performances for all test sets using the proposed algorithm are given in Table 12. The numbers given in brackets give the relative improvements obtained using the proposed algorithm over the corresponding baseline. Relative improvements of 38% (on a connected-digit recognition task) and 36% (on a continuous speech recognition task) are obtained for child test sets over baseline that are close to the best performance improvements of 54% (on a connected-digit recognition task) and 38% (on a continuous speech recognition task) obtained using the same MFCC base feature length for all test signals. However, when tested on the adult test set, no significant improvement in performance is obtained compared with baseline on the models trained on adult speech. In fact, the performance

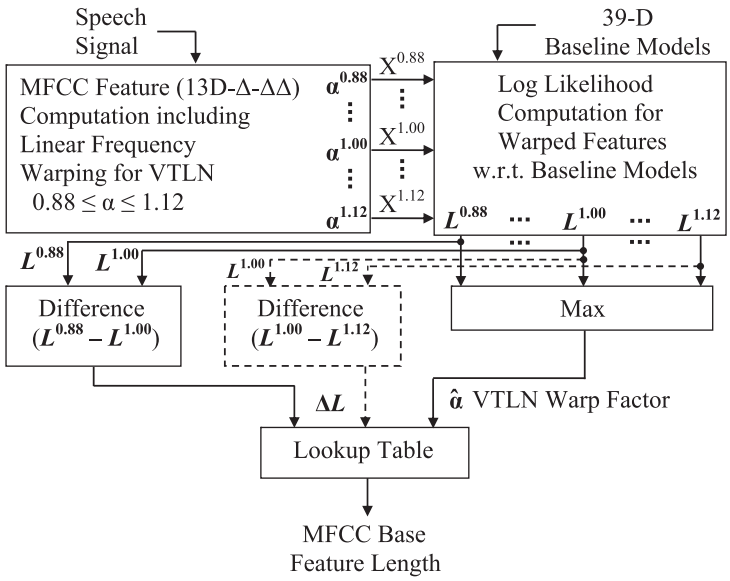

Fig. 7. Flow diagram of the proposed algorithm to determine the appropriate MFCC base feature length for test signal on models trained on adult speech (solid lines) or models trained on child speech (in dashed lines).

https://doi.org/10.1017/ATSIP.2016.16 Published online by Cambridge University Press

Table 12. Performances of test sets on recognition models trained with different training data sets using MFCC features derived using the proposed algorithm referred to as "Proposed" for both connected-digit and continuous speech recognition task.

	Test set WER (%)									
	ADTs		CHTs1		CHTs2		CAMTs		PFTs	
Training data set	Baseline	Proposed	Baseline	Proposed	Baseline	Proposed	Baseline	Proposed	Baseline	Proposed
ADTr	0.43	0.53 (-23%)	11.37	7.09 (38%)	–	–	–	–	–	–
CHTr	13.28	8.70 (35%)	–	–	1.01	0.52 (49%)	–	–	–	–
CAMTr	–	–	–	–	–	–	9.92	10.28 (-3.6%)	56.34	36.21 (36%)
PFTr	–	–	–	–	–	–	68.36	61.43 (10%)	12.41	8.62 (31%)

The numbers given in brackets give the relative improvements obtained using the proposed algorithm over their corresponding baselines.

of adult test sets on matched models show slight degradation using the proposed algorithm as indicated with negative numbers in brackets for two cases in Table 12.

However, their performance degradation, attributed to smaller degree of gross acoustic mismatch between adult training data and adult test data, is insignificant relative to the degree of improvement that is obtained in the recognition performance for child test sets. Further, when tested on models trained on child speech, the proposed algorithm gives consistent improvements for both mismatched adult test set and matched child test set on both digit and continuous speech recognition tasks. Relative improvements of 49 and 31% are obtained over baseline in ASR performances on children's speech. When the adult test set is evaluated using models trained on child speech, relative improvements of 35 and 10% are obtained over baseline on the digit and continuous speech recognition tasks. However, greater improvements are obtained for children's speech than for adults' speech on models trained on child speech. Also, for ASR of adults' speech, the improvements obtained are less than that for ASR of children's speech under mismatched conditions. This is because the variances for the observation densities of the phone models are greater for the poor models trained on child speech than for the models trained on adult speech [2, 8]. This means that the Gaussian densities are more scattered and thus less separable in the acoustic feature space for models trained on child speech.

Thus, the proposed algorithm is effective in reducing the acoustic mismatch between child speech and adult speech, without using any prior knowledge about the speaker of the test utterance, and with the additional advantage of reduced MFCC feature dimensions. Relative improvements obtained using the proposed algorithm in all ASR performances under acoustically mismatched conditions are close to the best performances obtained with constant fixed feature truncation for all test signals. However, these significant improvements in the recognition performances using the proposed algorithm are obtained with an increase in the execution time by a factor of 10 with respect to the default HMM-based ASR algorithm.

D) Combining proposed algorithm with VTLN and/or CMLLR

VTLN and CMLLR are the two effective techniques in the literature that are used to reduce acoustic mismatch between adult speech and child speech [8, 11]. Our proposed algorithm for adaptive MFCC feature truncation also addresses acoustic mismatch and has given significant improvements in performance. It would be interesting to explore whether the improvement obtained by our proposed algorithm is additive to that obtained by VTLN and/or CMLLR or not.

In these experiments, VTLN is performed on an utterance-by-utterance basis on the test speech data. An ML grid search procedure is used to carry out VTLN with frequency warp factors ranging from 0.88 to 1.12 in steps of 0.02 as described in Section IV-B. CMLLR is a feature adaptation technique that estimates a set of linear transformations for the features. The effect of these transformations is to modify the feature vector so as to increase its likelihood with respect to the given model. The transformation matrix used to give a new estimate of the adapted observation $\hat{o}$ is given by:

$$\hat{o} = Ao + b, \tag{3}$$

where o is an $n \times 1$ observation vector, A represents an $n \times n$ transformation matrix, and b represents an $n \times 1$ bias vector. The ML estimates of the affine transformations for adaptation of the features are obtained using the EM algorithm on adaptation data.

The recognition performances for "PFTs" on models trained on "CAMTr" using the default MFCC features and the features derived using the proposed adaptive MFCC feature truncation algorithm both with and without VTLN are given in Table 13. Another 15% relative improvement is obtained over the performance obtained using VTLN on default MFCC features by carrying out VTLN with MFCC features derived using the proposed adaptive truncation algorithm. However, the relative improvement with VTLN

https://doi.org/10.1017/ATSIP.2016.16 Published online by Cambridge University Press

Table 13. Performances of different test sets using the default MFCC features referred to as "Default" and MFCC features derived using the proposed algorithm referred to as "Proposed" both with and without VTLN and/or CMLLR on recognition models trained with different training data sets for continuous speech recognition task. Relative gain in ASR performance obtained with CMLLR over the respective baseline is also given.

	Training: "CAMtr" Test: "PFts"			Training: "PFtr" Test: "PFts"			Training: "PFtr" Test: "CAMts"		
	WER (%)			WER (%)			WER (%)		
Condition	Baseline	with CMLLR	Relative Gain (%)	Baseline	with CMLLR	Relative Gain (%)	Baseline	with CMLLR	Relative Gain (%)
Default	56.34	38.25	32	12.41	10.22	18	68.36	64.92	5
Default + VTLN	26.78	18.63	30	9.06	7.99	12	50.58	42.67	16
Proposed	36.21	23.07	36	8.62	7.54	13	61.43	49.74	19
Proposed + VTLN	22.72	**16.16**	29	7.70	**6.75**	12	47.80	**39.89**	17

The bold numbers highlight the best performances obtained for each test set on recognition models trained on different training data sets using the proposed algorithm in conjunction with VTLN and CMLLR.

is reduced by 15% when the MFCC features derived using the proposed algorithm are used. The ASR performances for "PFts" on models trained on "CAMtr" using CMLLR on the default MFCC features and the features derived using the proposed adaptive truncation algorithm both with and without VTLN are also given in Table 13.

CMLLR is performed for computing speaker-specific transformations as supported in the HTK Toolkit [21]. A relative gain of 40% is obtained in the recognition performance using CMLLR on MFCC features derived using the proposed algorithm over the performance obtained by using CMLLR on the default. MFCC features. In relation to the recognition performances obtained by combined VTLN and CMLLR, a relative improvement of 13% is obtained using the proposed algorithm over that obtained using the default MFCC features.

The additional improvements obtained for "PFts" using the proposed algorithm over those obtained from VTLN and/or CMLLR under mismatched conditions are further validated for recognition of "PFts" and "CAMts" on models trained using "PFtr" as given in Table 13. A similar trend to that obtained for "PFtr" on models trained on "CAMtr" is observed in performance improvements for both "PFts" and "CAMts" on models trained on "PFtr". Relative gains of 15% for VTLN, 26% for CMLLR, and 16% for VTLN with CMLLR are obtained in the ASR performance over baseline for "PFts" using MFCC features derived using the proposed algorithm. Relative gains of 6% for VTLN, 23% for CMLLR, and 7% for VTLN with CMLLR are obtained in the recognition performance over baseline for "CAMts" using MFCC features derived using the proposed algorithm.

Thus, the improvement obtained using the proposed adaptive MFCC feature truncation algorithm is additive to those obtained with VTLN and/or CMLLR. This is because the proposed algorithm is not constrained to use linear transformations. By reducing the feature dimension for both training and test sets at the same time, SAT is implicitly incorporated in the proposed algorithm and thus does not require the models to be explicitly adapted and retrained.

V. CONCLUSIONS

A novel technique has been developed to reduce the effect of acoustic mismatch between training and test speech on ASR performance. The higher-order MFCCs have been demonstrated to be more affected by acoustic mismatch due to pitch differences than lower-order MFCCs, causing degradation in the ASR performance under mismatched conditions. Based on the correspondence between the MFCC base feature length and the degree of acoustic mismatch for a speech signal, an adaptive algorithm has been proposed for utterance-specific MFCC base feature truncation for each test signal without prior knowledge about the speaker of the test utterance. Using the proposed algorithm, significant improvement is obtained in ASR performance on children's speech under mismatched conditions, with no significant change in ASR performance on adults' speech under matched conditions on both connected-digit recognition and continuous speech recognition tasks. Moreover, these improvements are additive to performance improvements obtained with the traditional VTLN and CMLLR methods used to address acoustic mismatch in ASR. Similar improvements are also obtained on matched child speech and mismatched adult speech with and without VTLN and/or CMLLR.

ACKNOWLEDGEMENTS

This research was originally carried out as a part of the first author's Ph.D. thesis [27] in the Department of Electronics and Electrical Engineering, Indian Institute of Technology Guwahati. Further validation of this research, comparing the cepstral truncation-based algorithm proposed in this paper with a heteroscedastic linear discriminant analysis (HLDA)-based technique for acoustic mismatch reduction, was presented in [28]. The authors gratefully thank Dr. Gordon Ramsay for his valuable suggestions and assistance in writing this manuscript.

https://doi.org/10.1017/ATSIP.2016.16 Published online by Cambridge University Press

REFERENCES

[1] Potamianos, A.; Narayanan, S.: Robust recognition of children's speech. IEEE Trans. Speech Audio Process., 11 (6) (2003), 603–616.

[2] Lee, S.; Potamianos, A.; Narayanan, S.: Acoustics of children's speech: developmental changes of temporal and spectral parameters. J. Acoust. Soc. Amer., 105 (3) (1999), 1455–1468.

[3] Wilpon, J.G.; Jacobsen, C.N.: A Study of Speech Recognition for Children and the Elderly, *ICASSP*, Atlanta, Georgia, 1996, 349–352.

[4] Das, S.; Nix, D.; Picheny, M.: Improvements in Children's Speech Recognition Performance, *ICASSP*, Seattle, Washington, 1998, 433–436.

[5] Gerosa, M.; Giuliani, D.; Brugnara, F.: Towards age-independent acoustic modeling. Speech Commun., 51 (2009), 499–509.

[6] Gerosa, M.; Giuliani, D.; Brugnara, F.: Speaker Adaptive Acoustic Modeling with Mixture of Adult and Children's Speech, *Interspeech*, Lisbon, Portugal, 2005, 2193–2196.

[7] Elenius, D.; Blomberg, M.: Adaptation and Normalization Experiments in Speech Recognition for 4 to 8 Year Old Children, *Interspeech*, Lisbon, Portugal, 2005, 2749–2752.

[8] Gerosa, M.; Giuliani, D.; Brugnara, F.: Acoustic variability and automatic recognition of children's speech. Speech Commun., 49 (10–11) (2007), 847–860.

[9] Narayanan, S.; Potamianos, A.: Creating conversational interfaces for children. IEEE Trans. Speech Audio Process., 10 (2) (2002), 65–78.

[10] Gales, M.J.F.: Maximum likelihood linear transformations for HMM-based speech recognition. Comput. Speech Lang., 12 (2) (1998), 75–98.

[11] Giuliani, D.; Gerosa, M.; Brugnara, F.: Improved automatic speech recognition through speaker normalization. Comput. Speech Lang., 20 (1) (2006), 107–123.

[12] Gustafson, J.; Sjölander, K.: Voice Transformations for Improving Children's Speech Recognition in a Publicly Available Dialogue System, *ICSLP*, Denver, Colorado, 2002, 297–300.

[13] Garau, G.; Renals, S.: Combining spectral representations for large vocabulary continuous speech recognition. IEEE Trans. Audio Speech Lang. Process., 16 (2008), 508–518.

[14] Singer, H.; Sagayama, S.: Pitch Dependent Phone Modelling for HMM Based Speech Recognition, vol. **1**. ICASSP, San Francisco, CA, 1992, 273–276.

[15] Sinha, R.; Ghai, S.: On the use of Pitch Normalization for Improving Children's Speech Recognition, *Interspeech*, Brighton, UK, 2009, 568–571.

[16] Rabiner, L.; Juang, B.-H.: Fundamentals of Speech Recognition, 1st ed., *Prentice-Hall PTR*, Upper Saddle River, NJ, 1993.

[17] Leonard, R.: A Database for Speaker-independent Digit Recognition, *ICASSP*, San Diego, CA, 1984, 42.11.1–42.11.4.

[18] Robinson, T.; Fransen, J.; Pye, D.; Foote, J.; Renals, S.: WSJCAM0: A British English Speech Corpus for Large Vocabulary Continuous Speech Recognition, *ICASSP*, Detroit, Michigan, 1995, 81–84.

[19] Batliner, A. et al.: The PFSTAR Children's Speech Corpus, *Interspeech*, Lisbon, Portugal, 2005, 2761–2764.

[20] Sjölander, K.; Beskow, J.: Wavesurfer – An Open Source Speech Tool, *ICSLP*, Beijing, China, 2000, 464–467.

[21] Young, S. et al.: The HTK Book Version 3.4, *Cambridge University Engineering Department*, Cambridge, UK, 2006.

[22] Hirsch, H.; Pearce, D.: The AURORA Experimental Framework for the Performance Evaluation of Speech Recognition Systems Under Noisy Conditions, *ISCA IITRW ASRU*, Paris, France, 2000, 181–188.

[23] Burnett, D.; Fanty, M.: Rapid Unsupervised Adaptation to Children's Speech on a Connected-digit Task, vol. **2**, *ICSLP*, Philadelphia, PA, 1996, 1145–1148.

[24] Cabral, J.P.; Oliveira, L.C.: Pitch-synchronous Time-scaling for Prosodic and Voice Quality Transformations, *Interspeech*, Lisbon, Portugal, 2005, 1137–1140.

[25] Ghai, S.; Sinha, R.: Exploring the effect of differences in the acoustic correlates of adults' and children's speech in the context of automatic speech recognition. EURASIP J. Audio Speech Music Process., Article ID 318785 (2010), 15. doi: 10.1155/2010/318785.

[26] Lee, L.; Rose, R.C.: A frequency warping approach to speaker normalization. IEEE Trans. Speech Audio Process., 6 (1) (1998), 49–60.

[27] Ghai, S.: Addressing Pitch Mismatch for Children's Automatic Speech Recognition. Unpublished Ph.D. thesis, Department of Electronics and Electrical Engineering, Indian Institute of Technology Guwahati, India, 2011.

[28] Kathania, H.K.; Ghai, S.; Sinha, R.: Soft-Weighting Technique for Robust Children Speech Recognition under Mismatched Condition, *INDICON*, Mumbai, India, 2013.

Shweta Ghai received a B. Tech. degree in Electronics and Communication Engineering from Kurukshetra University, India in 2006 and a Ph.D. degree in Electronics and Electrical Engineering from the Indian Institute of Technology Guwahati, India in 2011. She is currently a post-doctoral fellow in the Department of Pediatrics at Emory University School of Medicine, Atlanta, USA. Her main research interests are children's speech analysis and speech recognition, which she is currently applying to study vocal and spoken communication development in infants and toddlers at risk of autism spectrum disorders.

Rohit Sinha received the M. Tech. and Ph. D. degrees in Electrical Engineering from the Indian Institute of Technology, Kanpur, in 1999 and 2005, respectively. From 2004 to 2006, he was a post-doctoral researcher in the Machine Intelligence Laboratory at Cambridge University, Cambridge, UK. Since 2006 he has been with Indian Institute of Technology Guwahati, Assam, India, where he is currently a full professor in the Department of Electronics and Electrical Engineering. His research interests include fast speaker adaptation in context of automatic speech recognition, audio segmentation, signal enhancement/denoizing, speaker and language recognition, and spectrum sensing.

https://doi.org/10.1017/ATSIP.2016.16 Published online by Cambridge University Press