Speaker Adaptive Acoustic Modeling with Mixture of Adult and Children's Speech

M. Gerosa[+], D. Giuliani[*] and F. Brugnara[*]*

([+]) International Graduate School, University of Trento, 38050 Pantè di Povo, Trento, Italy
([*]) Centro per la Ricerca Scientifica e Tecnologica, 38050 Pantè di Povo, Trento, Italy
{gerosa,giuliani,brugnara}@itc.it

Abstract

In this paper, speaker adaptive acoustic modeling is investigated in the context of large vocabulary speech recognition by training acoustic models with adult speech, children's speech and a mixture of adult and children's speech.

By exploiting a limited amount (9 hours) of children's speech and a more significant amount (57 hours) of adult speech, group-specific acoustic models for children and adults, respectively, were trained using several methods for speaker adaptive acoustic modeling. In addition, age-independent acoustic models were trained by exploiting adult and children's speech. Recognition experiments were performed on three speech corpora, two consisting of children's speech and one of adult speech, using 64k word and 11k word trigram language models.

Methods for speaker adaptive acoustic modeling proved to be effective, in particular for training acoustic models on a mixture of adult and children's speech, ensuring recognition performance aligned with that achieved with group-specific models for adults and children. A 10.2% word error rate was achieved on speech collected from children in the age range 8-12, compared with the 8.2% word error rate achieved for adults uttering the same texts.

1. Introduction

It is well known that when an automatic speech recognition system trained on adult speech is employed to recognize children's speech, performance decreases drastically, especially for younger children [1, 2, 3]. Characteristics of speech such as pitch, formant frequencies and segmental durations are, in fact, related to the age of the speakers [4]. For recognition of children's speech, age-specific acoustic models trained on speech collected from children of the target age, or group of ages, should be adopted [1, 2, 5]. However, training age-specific acoustic models is costly as it requires collecting an adequate amount of training data for each target age or group of ages. In languages other than American English [5], there is a relative scarcity of large, publicly-available corpora of children's speech. Therefore, as a first approximation, children are often treated as an homogeneous population group and group-specific acoustic models are trained with speech from children of all ages [1, 2, 6].

However, even in the case of adequate amounts of age-specific training data, recognition performance reported for children is usually significantly lower than that reported for adults [1, 2] and it improves as the children's age increases [1, 2, 5]. This correlates well with studies showing that intra- and inter-speaker spectral variability decrease as age increases [4] and confirm that recognition of children's speech is more difficult than recognition of adult speech especially when targeting younger children [7].

In recent years, research issues, such as vocal tract length normalization, speaker adaptive training, language modeling and pronunciation variation modeling have been investigated for improving children's speech recognition [7, 2, 3, 5, 6], however all these issues still require systematic studies.

In this work, group-specific acoustic models for children and adults, respectively, were trained using several methods for speaker adaptive acoustic modeling. A limited amount (9 hours) of children's speech was exploited for training acoustic models for children while a more significant amount of adult speech (57 hours) was available for training acoustic models for adults. In addition, age-independent acoustic models were trained by exploiting adult and children's speech.

Acoustic modeling was investigated in the context of a large vocabulary speech recognition task by exploiting two parallel speech corpora consisting of the same set of sentences read by adults and children, respectively. This allowed us to compare recognition performance achieved for adults and children. In addition, a third test set, composed of read children's speech, was exploited to further validate the results. Results showed that speaker adaptive acoustic modeling methods were effective in training acoustic models on a mixture of adult and children's speech, ensuring recognition performance aligned with that achieved with group-specific models.

The paper is organized as follows: first, the speech corpora used in this work are described in Section 2. Section 3 presents an analysis on phone duration in children's speech as a function of age. The experimental set-up is then presented in Section 4. Section 5 briefly introduces the speaker adaptive acoustic modeling methods adopted. Recognition experiments are described in Section 6. Final remarks are reported in Section 7 which concludes the paper.

2. Speech corpora

For acoustic model training three speech corpora were used. Two of them, the ChildIt and the SpontIt corpora, consist of children's speech while the third one, the IBN corpus, consists of adult speech. Tables 1 and 2 summarize the characteristics of the speech corpora used in this work for training and testing.

training set	IBN	ChildIt	SpontIt
speaking style	planned/spont.	read	spont.
signal quality	clean	clean	clean
sampling frequency	16 kHz	16 kHz	16 kHz
language	Italian	Italian	Italian
speaker age	>20	7-13	8-12
no. speakers	>1000	129	21
recording hours	57h:07m	7h:47m	1h:20m

Table 1: *Characteristics of speech corpora used for acoustic model training.*

The ChildIt corpus is an Italian task-independent speech

This work has been partially funded by the European Union under the project TC-STAR (grant FP6-506738, http://www.tc-star.org) and by the Autonomous Province of Trento (Italy) under the project PEACH (Fondo Unico Program).

database that consists of read speech collected from children. Data collection was performed in several schools located in the North of Italy involving a total of 171 children, from grade 2 through grade 8, evenly distributed by grade and gender. Children in grade 2 were approximately 7 years old while children in grade 8 were approximately 13 years old. Speech data from 129 children were used for training.

The SpontIt corpus is an Italian task-independent speech database that consists of spontaneous speech from 21 children aged between 8 and 12, with a mean age of 10 years. The ChildIt and the SpontIt corpora together provided about 9 hours of children's speech to be used for training acoustic models.

The IBN corpus consists mainly of recordings of Italian radio and TV news programs which were manually segmented, annotated and transcribed [8].

test set	Tgr-adult	Tgr-child	ChildIt
speaking style	planned	read	read
signal quality	clean	clean	clean
sampling freq.	16 kHz	16 kHz	16 kHz
language	Italian	Italian	Italian
speaker age	>20	8-12	7-13
no. speakers	76	30	42
no. utterances	570	570	1680
word occurrences	6575	6575	15355
language model	trigram		trigram
rec. dictionary size	64000		11000
perplexity	180		900
OOV rate	1.0%		0.0%

Table 2: *Characteristics of speech corpora used for recognition experiments.*

Two parallel corpora, called Tgr-adult and Tgr-child, containing the same set of sentences uttered by adults and children, were designed for testing. By exploiting manual segmentation and word transcription, the sentences in the IBN test set suitable to be read by children were identified and grouped into lists of about 20 sentences each. These sentences were selected among those judged well pronounced by transcribers of the IBN corpus and characterized by good acoustic conditions. Each of the 30 children, aged from 8 to 12, involved in the data collection was asked to read one of these lists. Children were allowed to repeat the same sentence more than once, and just the last repetition was stored.

In practice, the Tgr-adult corpus is a subset of the IBN test set (which was not used in this work) corresponding to the sentences selected for children. We have to point out that the set of sentences read by a specific child in Tgr-child was usually pronounced by several speakers in the IBN test set, as is evident by the number of speakers in the two corpora reported in Table 2. This caused a certain mis-alignment in experimental conditions in some recognition experiments. In practice, in experiments in which, at the recognition stage, the system is adapted to the incoming test data, the amount of data available plays a role in the grade of adaptation achieved. For each adult speaker in Tgr-adult corpus, system adaptation was performed on all the speech available in the IBN test set while performance was reported only for utterances included into the Tgr-adult corpus described above. However, this mis-alignment was accepted because it was not possible to extract a sufficient number of suitable sentences with an even distribution over adult speakers.

In addition to the two Tgr corpora the test portion of the ChildIt corpus was also used, which contains data from 42 children, 3 males and 3 females for each grade.

3. Mean phone duration

In the literature it is reported that adults and older children tend to show shorter durational patterns than younger children [4]. In this work phone duration was analyzed as a function of age. The mean phone duration was computed first averaging the duration of phones across all phones of each speaker and then across speakers in each grade. Duration statistics were computed by exploiting a phone-level segmentation produced automatically. Each utterance was time-aligned with the HMM concatenation corresponding to the uttered words allowing insertion of an optional "silence" model between words and at the beginning and the end of the utterance.

Segments of signals aligned with the "silence" HMM were not taken into account in computing temporal statistics. Two group-specific sets of triphone HMMs were used for children and adults. Figure 1 reports the mean phone duration for children, computed on the training set of the ChildIt corpus (at least 14 speakers per grade), and for adults, computed on the training set of the IBN corpus.

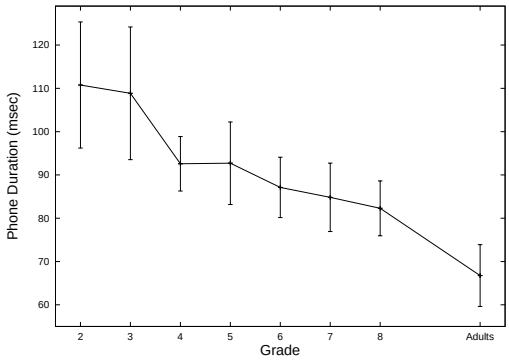

Figure 1: *Mean duration of phones (msec) per grade computed on the ChildIt training set. For comparison purpose, the mean phone duration for adults, as computed on the IBN training set, is also reported. Vertical bars denote inter-speaker variability (standard deviation).*

Mean phone duration varies with age and older age groups exhibit shorter mean phone durations. However, we have to point out that the mean phone durations reported here are likely affected by reading ability and length of sentences (much shorter for younger children). Furthermore, the significant difference in mean phone duration between children in grade 8 and adults can be explained by the fact that the IBN corpus is formed mostly of speech from professional radio and TV announcers speaking quite fast.

4. Experimental set-up

For recognition experiments we used the ITC-irst HMM software package employing state-tied, cross-word triphone HMMs. In particular, a Phonetic Decision Tree (PDT) was used for tying the states of triphone HMMs. Output distributions associated with HMM states were modeled with mixtures with up to 8 diagonal covariance Gaussian densities.

In all acoustic model sets trained, "silence" was modeled with a single state HMM. In addition a number of models for common non-verbal phenomena (15 when training on adult speech and 5 when training on children's speech) were trained.

Each speech frame was parameterized into a 39-dimensional observation vector composed of 13 mel frequency cepstral coefficients (MFCCs) plus their first and second order time derivatives. Cepstral mean subtraction was performed on static features on an utterance-by-utterance basis. In the following, this acoustic front-end is denoted as MFCC39.

Two additional acoustic front-ends were considered by performing mean and variance normalization in two different ways. In one case, after generating the MFCCs, mean subtraction and variance normalization was performed before computing first and second order time derivatives. We will denote this set of acoustic features with MFCC39-MVN13. Alternatively, mean

and variance normalization was applied to all 39 unnormalized acoustic features. We will denote this latter set of acoustic features as MFCC39-MVN39. Mean and variance normalization was always performed on a speaker-by-speaker basis by assuming data from each speaker available in a single block and forcing each acoustic feature to have zero mean and unit variance.

Two language models (LMs) were estimated and used in speech recognition experiments reported in this paper.

For experiments on the two Tgr corpora the language model was the 64k word trigram language model adopted by the broadcast news transcription system developed at ITC-irst for the Italian language [8]. The second language model, used for recognition experiments on the ChildIt test set, was an 11k trigram language model estimated on a corpus of newspaper articles. The word dictionary was composed of the words occurring in the training and test sets of the ChildIt corpus.

The perplexities and the out-of-vocabulary (OOV) rates computed on the test sets are reported in Table 2. The high perplexity shown by the 11k trigram language model on the ChildIt test set is explained by the fact that the statistics estimated on the training text corpus, composed of newspaper articles, do not reflect well the statistics of the ChildIt test set, which was extracted from literature for children.

5. Speaker adaptive acoustic modeling

Speaker adaptive modeling aims at reducing or compensating for acoustic variations induced by different characteristics of each training and testing speaker. In this work, speaker adaptive acoustic modeling was investigated through vocal tract length normalization (VTLN), speaker adaptive training (SAT) and constrained MLLR based speaker normalization (CMLSN).

VTLN aims at reducing inter-speaker acoustic variability due to vocal tract length variation among speakers [9]. The training and recognition procedures adopted for implementing VTLN in this work follow closely those proposed in [10]. However, some changes were introduced. The frequency warping process was implemented by changing the spacing and the width of the filters in the mel filter-bank while keeping the speech spectrum unchanged [9]. In addition, during recognition a speaker-specific warping factor was estimated instead of selecting the warping factor on an utterance-by-utterance basis.

Speaker adaptive training aims at compensating for inter-speaker acoustic variability present in the training set by means of speaker-specific transformations, estimated through maximum likelihood linear regression (MLLR), of the means of output distributions of continuous density HMMs [11]. The variant of the SAT algorithm developed by Gales [12] was used in this work. This variant makes use of an affine transformation for transforming acoustic observations of each training and testing speaker, instead of modifying model parameters. With this method, during training, a set of SI models, fully trained on unnormalized data, is assumed as seed models, and parameters of Gaussian densities of these models are iteratively re-estimated by applying the estimated transformations on training data. During the recognition stage, transformation parameters are again iteratively estimated with respect to the models to be used for decoding.

CMLSN is a speaker normalization method which performs speaker normalization by transforming the acoustic observation vectors by means of speaker-specific constrained MLLR transformations. However, differently from the variant of SAT proposed by Gales in [12], speaker-specific transformations are estimated with the aim of reducing the acoustic mismatch of the speaker's data with respect to a set of target HMMs which is different from the HMM set to be used for recognition. This is done both during training and decoding stages. Details about this technique can be found in [6].

For each of the three methods described above, word transcriptions for test utterances of each speaker were provided by a preliminary decoding step carried out with baseline models trained on unnormalized data. Data of each speakers were assumed available in one block for multiple processing. The combinations of the VTLN method described above with the SAT and the CMLSN methods were also investigated.

6. Recognition experiments

In this Section, results of several experiments are reported concerning recognition of adult and children's speech with acoustic models trained on adult speech (Adult HMMs), children's speech (Child HMMs) and a mixture of adult and children's speech (Adult+Child HMMs).

6.1. Acoustic feature normalization

A common practice in automatic transcription experiments is that of performing mean and variance normalization of acoustic features. We investigated two possibilities - performing mean and variance normalization on a speaker-by-speaker basis of: 1) the 13 MFCCs; 2) all the 39 acoustic features. These experiments were motivated by the fact that analysis on phone duration, presented in Section 3, revealed that adults and children's speech in the corpora used in this work was characterized by a very different mean phone duration. It can be hypothesized that the effect of the speaking rate is mostly concentrated on the first and second order time derivatives of the MFCCs [13], therefore performing mean and variance normalization of dynamic features could be useful to compensate for very different speaking rates.

Recognition results, in terms of word error rate (WER), are reported in Table 3. By considering first results obtained with the standard acoustic front-end (MFCC39), we note that with group-specific acoustic models under matched conditions (that is, training and testing on the same population group), performance for adults is much better than that for children, 10.4%WER to be compared with 14.2% WER. However, we have to point out that there are much more training data for adults than for children and therefore the performance gap could be partially filled by just having more training data. As expected, under unmatched conditions (for example, in the case of children's speech recognized with acoustic models trained on adult speech), recognition results are much worse than those achieved under matched conditions. This is mainly due to different vocal characteristics of adults and children [1, 2, 4].

When training age-independent models with an unbalanced amount of adult and children's speech (57 hours of adult speech plus 9 hours of children's speech for a total of 66 hours), recognition results were worse than those achieved with group-specific models, especially for children: 20.6% WER was achieved to be compared with 14.2% WER. This means that simply mixing unbalanced amounts of training data is not an effective approach for training age-independent acoustic models.

		Test Set	
HMM set	Feature Set	Tgr-adult	Tgr-child
Adult HMMs	MFCC39	10.4	37.2
	MFCC39-MVN13	10.2	36.4
	MFCC39-MVN39	10.1	33.3
Child HMMs	MFCC39	45.4	14.2
	MFCC39-MVN13	48.1	14.1
	MFCC39-MVN39	44.1	13.8
Adult + Child HMMs	MFCC39	11.0	20.6
	MFCC39-MVN13	10.4	20.4
	MFCC39-MVN39	10.5	17.9

Table 3: *Performance (% WER) obtained on the Tgr-adult and Tgr-child test sets with acoustic models trained on adults and children and by adopting different acoustic front-ends.*

In Table 3 we can note that mean and variance normalization of all acoustic features (MFCC39-MVN39) ensures sys-

tematic benefits with respect to adopting the standard acoustic front-end (MFCC39), especially in the case of unmatched training and testing conditions and of training with a mixture of adult and children's speech. Normalizing just the static acoustic features (MFCC39-MVN13) seems to be less effective and consistent. Therefore, the experiments reported in the following were carried out performing mean and variance normalization on all acoustic features.

6.2. Adaptive training experiments

Speaker adaptive acoustic modeling was carried out by training acoustic models with VTLN, CMLSN and SAT methods. Additional speaker adaptive methods resulted from cascading the VTLN method with the SAT and the CMLSN methods.

Word transcriptions of test utterances, needed for normalization/adaptation purposes, were provided by preliminary decoding steps with the single pass baseline systems, the reference performance of which are reported in Table 3 (rows MFCC39-MVN39).

In addition, unsupervised static speaker adaptation was always performed by adapting means and variances of Gaussian densities through MLLR before performing the second decoding step. Two regression classes were defined and the associated transformation matrices were estimated through three MLLR iterations. Recognition results are reported in Table 4.

		Test Set		
		Tgr-adult	Tgr-child	ChildIt
Adult HMMs	Baseline	9.3	15.3	20.0
	VTLN	8.8	13.0	15.4
	CMLSN	8.6	12.4	15.2
	SAT	8.6	12.3	16.0
	VTLN+SAT	8.5	11.8	14.8
	VTLN+CMLSN	8.2	11.8	14.4
Child HMMs	Baseline	//	12.0	11.6
	VTLN	//	11.5	11.2
	CMLSN	//	10.9	10.6
	SAT	//	11.1	11.0
	VTLN+SAT	//	10.7	10.5
	VTLN+CMLSN	//	10.5	10.6
Adult + Child HMMs	Baseline	9.7	12.7	13.6
	VTLN	8.9	11.1	11.3
	CMLSN	8.5	10.2	10.7
	SAT	8.7	11.0	11.5
	VTLN+SAT	8.4	10.1	10.5
	VTLN+CMLSN	8.2	10.2	10.4

Table 4: *Performance (% WER) obtained on the Tgr-adult and Tgr-child test sets using HMMs trained on adult speech, children's speech and a mixture of adult and children's speech with several speaker adaptive acoustic modeling methods. For comparison purposes, results on ChildIt corpus are also reported.*

Results on the two Tgr test sets show that methods for speaker adaptive acoustic modeling are effective. While CMLSN and SAT methods outperform the VTLN method, best results are achieved by cascading the VTLN method with the CMLSN and SAT methods. By considering group-specific models for adults, on adult speech the baseline system ensures a 9.3% WER, while with speaker adaptive acoustic modeling an 8.2% WER is achieved. By considering group-specific models for children, the baseline system ensures a 12.0% WER compared with a 10.5% WER achieved with speaker adaptive acoustic modeling. When training with a mixture of adult and children's speech, speaker adaptive acoustic modeling ensures performance aligned (compare results in rows VTLN+CMLSN) with that achieved for adults and children by using group-specific acoustic models: 8.2% WER for adults and 10.2% WER for children.

Results on the ChildIt test set confirm the effectiveness of speaker adaptive acoustic modeling methods. By considering models trained on children's speech, the baseline system ensures a 11.6% WER, while with speaker adaptive acoustic modeling a 10.5% WER is achieved. Furthermore, when training HMMs on adult and children's data a 10.4% WER (row VTLN+CMLSN) is obtained adopting speaker adaptive acoustic modeling.

7. Conclusions

In this work, speaker adaptive acoustic modeling was investigated by using for training children's speech, adult speech and a mixture of children's and adult speech.

Results on automatic transcription tasks with a large vocabulary showed that mean and variance normalization of acoustic features together with speaker adaptive acoustic modeling allowed the development of age-independent acoustic models ensuring performance for adults and children aligned with that provided by group-specific acoustic models. This also opens new perspectives to raise recognition performance for children aged from 8 to 12 to a similar level of that for adults.

8. References

[1] J.G. Wilpon and C.N. Jacobsen, "A Study of Speech Recognition for Children and Elderly," in *Proc. of ICASSP*, Atlanta, GA, May 1996, pp. I–349–352.

[2] A. Potamianos and S. Narayanan, "Robust Recognition of Children's Speech," *IEEE Trans. on Speech and Audio Processing*, vol. 11, no. 6, pp. 603–615, Nov. 2003.

[3] D. Giuliani and M. Gerosa, "Investigating Recognition of Children's Speech," in *Proc. of ICASSP*, Hong Kong, China, April 2003, pp. II–137–140.

[4] S. Lee, A. Potamianos, and S. Narayanan, "Acoustic of children's speech: Developmental changes of temporal and spectral parameters," *Journal of Acoust. Soc. Amer.*, vol. 105, no. 3, pp. 1455–1468, March 1999.

[5] A. Hagen, B. Pellom, and R. Cole, "Children's Speech Recognition with Application to Interactive Books and Tutors," in *Proc. of ASRU Workshop*, St. Thomas, USA, December 2003.

[6] D. Giuliani, M. Gerosa, and F. Brugnara, "Speaker Normalization through Constrained MLLR Based Transforms," in *Proc. of INTERSPEECH/ICSLP*, Jeju Island, Korea, Oct. 2004, pp. 2893–2897.

[7] Q. Li and M. Russell, "An Analysis of the Causes of Increased Error Rates in Children's Speech Recognition," in *Proc. of ICSLP*, Denver,CO, Sep. 2002.

[8] N. Bertoldi, F. Brugnara, M. Cettolo, M. Federico, and D. Giuliani, "From broadcast news to spontaneous dialogue transcription: Portability issues," in *Proc. of ICASSP*, Salt Lake City, UT, 2001, vol. 1, pp. 37–40.

[9] L. Lee and R.C. Rose, "Speaker Normalization Using Efficient Frequency Warping Procedure," in *Proc. of ICASSP*, Atlanta, GA, May 1996, pp. I–353–356.

[10] L. Welling, S. Kanthak, and H. Ney, "Improved Methods for Vocal Tract Normalization," in *Proc. of ICASSP*, Phoenix, AZ, April 1999, vol. 2, pp. 761–764.

[11] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, "A Compact Model for Speaker-Adaptive Training," in *Proc. of ICSLP*, Philadelphia, PA, Oct. 1996, pp. 1137–1140.

[12] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, pp. 75–98, 1998.

[13] F. Martnez, D. Tapias, and J. Alvarez, "Towards Speech Rate Independence in Large Vocabulary Continuous Speech Recognition," in *Proc. of ICASSP*, Seattle,WA, May 1998, vol. 2, pp. 725–728.