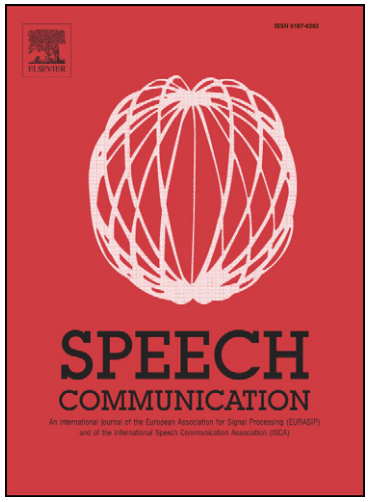
Towards Age-Independent Acoustic Modeling

Matteo Gerosa, Diego Giuliani, Fabio Brugnara



Towards Age-Independent Acoustic Modeling

Matteo Gerosa, Diego Giuliani, Fabio Brugnara

PII: S0167-6393(09)00008-9
DOI: 10.1016/j.specom.2009.01.006
Reference: SPECOM 1777

To appear in: *Speech Communication*

Received Date: 1 March 2007
Revised Date: 30 December 2008
Accepted Date: 23 January 2009

Please cite this article as: Gerosa, M., Giuliani, D., Brugnara, F., Towards Age-Independent Acoustic Modeling, *Speech Communication* (2009), doi: 10.1016/j.specom.2009.01.006

Towards Age-Independent Acoustic Modeling

Matteo Gerosa *, Diego Giuliani and Fabio Brugnara

FBK, Fondazione Bruno Kessler
I-38100 Povo (Trento), Italy
gerosa@itc.it, giuliani@itc.it, brugnara@itc.it

Abstract

In automatic speech recognition applications, due to significant differences in voice characteristics, adults and children are usually treated as two population groups, for which different acoustic models are trained. In this paper, age-independent acoustic modeling is investigated in the context of large vocabulary speech recognition. Exploiting a small amount (9 hours) of children's speech and a more significant amount (57 hours) of adult speech, age-independent acoustic models are trained using several methods for speaker adaptive acoustic modeling. Recognition results achieved using these models are compared with those achieved using age-dependent acoustic models for children and adults, respectively. Recognition experiments are performed on four Italian speech corpora, two consisting of children's speech and two of adult speech, using 64k word and 11k word trigram language models. Methods for speaker adaptive acoustic modeling prove to be effective for training age-independent acoustic models ensuring recognition results at least as good as those achieved with age-dependent acoustic models for adults and children.

Key words: Speaker adaptive acoustic modeling, speaker normalization, vocal tract length normalization, children's speech recognition.

* Corresponding author. Tel: +39 0461 314 561; Fax: +39 0461 314 591.

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 (Burnett and Fanty, 1996; Wilpon and Jacobsen, 1996; Potamianos et al., 1997; Das et al., 1998; Claes et al., 1998; Giuliani and Gerosa, 2003). Characteristics of speech such as pitch, formant frequencies and segmental durations have been shown, in fact, to be related to the age of the speaker (Huber et al., 1999; Lee et al., 1999). For recognition of children's speech, age-specific acoustic models (AMs) trained on speech collected from children of the target age, or age group, should be adopted to ensure good recognition performance (Wilpon and Jacobsen, 1996; Hagen et al., 2003; Nisimura et al., 2004). However, training age-specific acoustic models is costly as it requires collecting an adequate amount of training data for each target age or age group. Furthermore, in languages other than American English, there is a relative scarcity of large, publicly-available corpora of children's speech (Hagen et al., 2003; Batliner et al., 2005). 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 (Potamianos and Narayanan, 2003; Giuliani et al., 2006).

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 and it improves as the children's age increases (Wilpon and Jacobsen, 1996; Li and Russell, 2002; Potamianos and Narayanan, 2003; Hagen et al., 2003). This correlates well with studies showing that intra- and inter-speaker spectral variability decrease as age increases (Lee et al., 1999; Gerosa et al., 2006b). Furthermore, experiments of human perception of speech from children aged 6-11 show that the human word recognition error rate increases as the age of the child decreases (D'Arcy and Russell, 2005). All these results suggest that automatic recognition of children's speech is more difficult than recognition of adult speech especially when addressing younger children.

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 (Li and Russell, 2002; Narayanan and Potamianos, 2002; Potamianos and Narayanan, 2003; Stemmer et al., 2003; Giuliani and Gerosa, 2003; Hagen et al., 2003; Giuliani et al., 2004), however all these issues still require systematic studies.

On the other hand, for adults, variations in voice characteristics due to speaker age are much less evident than for children. As a consequence, for adults, recog-

nition performance is less correlated with age, at least for ASR applications addressing speakers in the age range 18-70 (Wilpon and Jacobsen, 1996).

Due to the significant difference in voice characteristics, in ASR applications adults and children are usually treated as two different population groups, for which different AMs are trained. In this work, we investigate a new approach, considering adults and children between 7 and 13 as a single population of speakers. Age-independent acoustic models are trained by exploiting a small amount (9 hours) of children's speech and a more significant amount (57 hours) of adult speech, for a total of 66 hours of speech. Recognition performance achieved using these models is compared with that achieved using group-specific acoustic models for children and adults. The aim is to train age-independent acoustic models able to perform well on both adult and children's speech.

Using age-independent AMs conventionally trained on a mixture of adult and children's speech results in a performance decrease with respect to using group-specific AMs. In fact, because of the increased inter-speaker acoustic variability caused by the very different characteristics of adult and children's speech, parameters of the age-independent models do not well reflect phonetically relevant acoustic variation present in the training data, providing limited modeling accuracy for each individual speaker. In this work, to cope with the high inter-speaker variability, speaker adaptive acoustic modeling is investigated by adopting three methods: vocal tract length normalization (VTLN) (Welling et al., 1999; Giuliani et al., 2006), speaker adaptive training (SAT) (Gales, 1998), and constrained MLLR based speaker normalization (CMLSN) (Giuliani et al., 2006).

Evaluation of ASR performance is conducted on four Italian read speech corpora, two composed of children's speech and two composed of adult speech, using 64k word and 11k word trigram language models. Two of these are parallel speech corpora consisting of the same set of sentences read by adults and children, respectively. This allows us to compare recognition performance achieved for adults and children on two corpora that are homogeneous in terms of linguistic content and signal quality.

This paper is organized as follows. First, the speech corpora used in this work are described in Section 2. Section 3 summarizes results of analysis of temporal and spectral characteristics of adult and children's speech. Section 4 briefly introduces the adopted speaker adaptive acoustic modeling methods. Recognition experiments are described in Section 5 and final remarks are reported in Section 6 which concludes the paper.

2 Speech corpora

Five Italian speech corpora were used in this work. Three of these corpora consist of children's speech: the ChildIt corpus, the SpontIt corpus and the Tgr-child corpus. The other two corpora consist of adult speech: the IBN corpus and the Tgr-adult corpus.

The ChildIt corpus (Giuliani and Gerosa, 2003) is an Italian task-independent, speech database that consists of clean utterances read by children from 7 to 13, with a mean age of 10 years. About 10 hours of speech were collected from 171 children. Each child read 58 or 65 sentences, depending on his/her grade, selected from electronic texts of literature for children. Each speaker read a different set of sentences. Speech was acquired at 16 kHz, with 16 bit accuracy, using a Shure SM10A head-worn microphone. The ChildIt corpus was partitioned into a training set and a test set for speech recognition experiments.

The SpontIt corpus is a task-independent Italian speech database that consists of clean spontaneous speech from 21 children aged between 8 and 12, with a mean age of 10 years. These 21 speakers were different from the 171 speakers in the ChildIt corpus. Each child was interviewed by an adult about his/her preferred books, TV shows, hobbies, sports, etc. Recordings were performed with a digital audio tape recorder using an head-worn Shure SM10A microphone. Audio signals were then down-sampled from 48 kHz to 16 kHz, with 16 bit accuracy. The SpontIt corpus was used in addition to ChildIt for acoustic model training.

The IBN speech corpus was used for training the automatic broadcast news (BN) transcription system developed at ITC-irst [1] for the Italian language (Bertoldi et al., 2001; Brugnara et al., 2002). It's mainly composed of speech from several radio and television news programs. The IBN corpus was partitioned into a training set, consisting of 57h:07m of speech, and a test set, consisting of 6 radio news programs and two television news programs, for a total of 49 minutes and 36 minutes of speech, respectively. The IBN training set includes also two small task-independent corpora called APASCI and SPEEDATA. The APASCI corpus (Angelini et al., 1994) is a task-independent, high quality, acoustic-phonetic Italian database. Recordings were performed in quiet rooms using a digital audio tape recorder and a high-quality close-talk microphone. Audio signals were acquired at 48 kHz sampling frequency and then down-sampled to 16 kHz with 16 bit accuracy. Only a portion of APASCI corpus, consisting of speech from 124 speakers (for an overall duration of 5h:38m), was exploited in this work for speech analysis purposes and acoustic model training. The SPEEDATA corpus (Ackermann et al., 1997) is

[1] Now FBK-irst.

a corpus designed and collected by ITC-irst with criteria very similar to those adopted for APASCI and containing about 5h:48m of speech.

Two parallel corpora, called Tgr-adult and Tgr-child, containing the same set of sentences uttered by adults and children, were also 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 between ages 8 and 12 that were 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. The Tgr-adult corpus is the subset of the IBN test set corresponding to the sentences selected for children.

Tables 1 and 2 summarize the characteristics of the speech corpora used in this work for training and testing, together with some characteristics of the language models (LMs) used.

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.

We have to point out that the set of sentences read by a specific child in the Tgr-child corpus 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. In fact, it was not possible to extract a sufficient number of suitable sentences with an even distribution over adult speakers. This caused a certain difference in experimental conditions. 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 the 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 experimental difference was considered acceptable.

test set	IBN	Tgr-adult	Tgr-child	ChildIt
speaking style	planned/spont.	planned	read	read
signal quality	clean/noisy	clean	clean	clean
sampling freq.	16 kHz	16 kHz	16 kHz	16 kHz
language	Italian	Italian	Italian	Italian
speaker age	Adult	Adult	8-12	7-13
no. speakers	95	76	30	42
no. utterances	1045	570	570	1680
word occurrences	14478	6575	6575	15355
rec. dictionary size	64000	64000	64000	11000
perplexity	204	180	180	900
OOV rate	1.6%	1.0%	1.0%	0.0%

Table 2
Characteristics of speech corpora used for recognition experiments.

3 Age-dependencies in speech acoustic characteristics

A lot of research work has been devoted to the analysis of children's speech in order to achieve a better understanding of its characteristics. In (McGowan and Nittrouer, 1988; Nittrouer and Whalen, 1989; Lee et al., 1999; Narayanan and Potamianos, 2002; Gerosa et al., 2007), it was shown that acoustic and linguistic characteristics of children's speech are widely different from those of adults. Furthermore, these studies also show that characteristics of children's speech vary rapidly as a function of age due to the anatomical and physiological changes occurring during a child's growth and because children become more skilled in coarticulation with age. Below we summarize some of the main results, reported in literature, concerning analysis of adult and children's speech.

Phone duration In literature it is reported that, in planned/read speech, adults and older children tend to show shorter durational patterns than younger children. In (Lee et al., 1999), mean duration of vowels was measured for American English speech uttered by children from 5 to 17 years of age. A significant decrease in duration as age increases was observed up to age 15. In (Gerosa et al., 2007) a similar study was conducted on Italian speech using the ChildIt corpus, achieving similar results (reported here in Figure 1). Observing Figure 1 we can note a decrease of almost 30% in phone duration between age 7 and age 13.

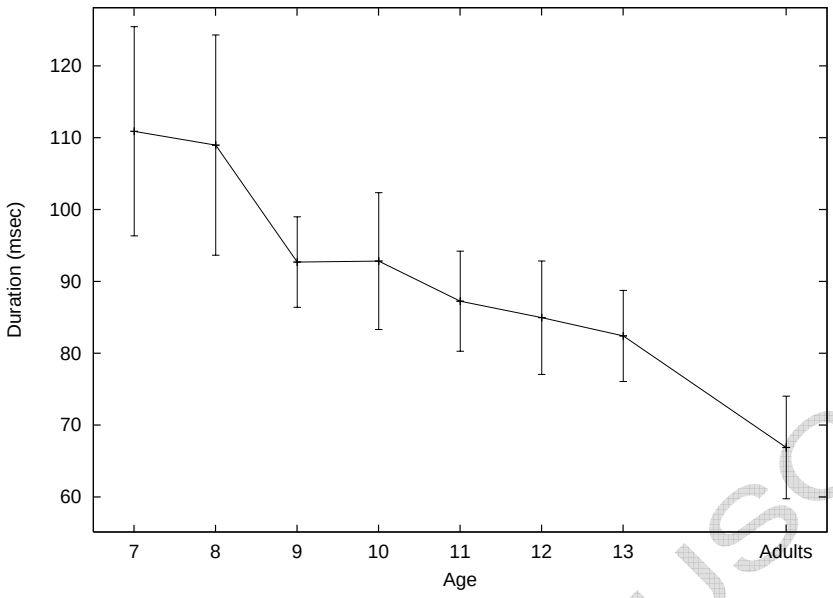


Fig. 1. *Mean duration of phones (msec) per age computed on the ChildIt training set. For comparison purposes, the mean phone duration for adults, computed on the IBN training set, is also reported. Vertical bars denote inter-speaker variability (standard deviation) (Gerosa et al., 2007).*

In (Gerosa et al., 2006a) phone duration was measured on spontaneous speech. The observed decrease in duration with age was smaller than for read speech, but it was still significant.

Age-dependent variation in phone duration introduces variabilities that may affect ASR performance. In the case of adults speakers, it is well known that for speakers speaking much faster than the average of the training population, a low ASR performance is achieved (Pallett et al., 1992; Mirghafori et al., 1996). When testing on children's speech with acoustic models trained on adult speech, the mismatch in duration variation is so large that it may significantly affect recognition performance.

Formant frequencies Studies on morphology and development of the vocal tract (Fitch and Giedd, 1999) reveal that during childhood there is a steady gradual lengthening of the vocal tract as the child grows while a concomitant decrease in formant frequencies occurs (Huber et al., 1999). While for females there is a gradual continuous growth of vocal tract through puberty into adulthood, for males during puberty there is a disproportional growth of vocal tract, which lowers formant frequencies, together with an enlargements of the glottis, which lowers the pitch. As a consequence, adult males show a longer, about 10% on average, vocal tract than adult females.

The above mentioned results are essentially confirmed by analysis of formant

frequency positions carried out in (Lee et al., 1999), for American English speech, and in (Gerosa et al., 2007), for Italian speech. On children's speech, a progressive and significant decrease of frequency position as age increases was observed for the fundamental frequency and the the first three formants. After age 13, formant frequency values were found significantly higher for female speakers than for male speakers. Before this age it is not clear if the difference in formant frequencies between male and female speakers are statistically significant. Pitch and formant frequencies reach adult-like values for children of age 15. This is in contrast with the developmental model presented in (Goldstein, 1980) for which the vocal tract is assumed to continue its growth beyond age 15 until age 20. The source of this discrepancy is still not fully understood. Observed spectral variations explain well why, when an automatic speech recognition system trained on adult speech is used to recognize children's speech, recognition performance decreases drastically (Burnett and Fanty, 1996; Wilpon and Jacobsen, 1996; Claes et al., 1998; Giuliani and Gerosa, 2003).

Spectral variability It is well known that intra- and inter-speaker variability are an important source of errors in automatic speech recognition systems. In (Lee et al., 1999) it is shown, for American English, that spectral variability in children's vowel sounds is much higher than in adult speech. Increased variability in formant frequencies results in greater overlap among phonemic classes for children than for adult speakers, and makes the speech classification problem inherently more difficult. This study was later extended to consonants in (Gerosa et al., 2006b), obtaining similar results to the ones obtained for vowels. In (Gerosa et al., 2007) the effect of inter-speaker spectral variability was measured on the ChildIt corpus, the distance between probability distributions modeling vowel sounds for different age groups was measured. The average distance for children aged 7-9 was about 25% lower than the average distance measured for adults, indicating more of an overlap between vowel distributions in the acoustic space. The high inter-speaker variability that characterizes children's speech suggests that the use of speaker adaptive acoustic modeling techniques has a great potential for application in cases when training acoustic models on children's speech, or on a mixture of adult and children's speech.

Human perception of children's speech In literature, a decrease of automatic speech recognition performance is usually reported for younger children compared to older ones. This decrease in performance is often ascribed to the increased inter- and intra-speaker acoustic variability caused by the age-related factors discussed in this section.

To ascertain whether similar performance degradation also applies to human

perception of children's speech, in (D'Arcy and Russell, 2005) human perceptual experiments were carried out by considering read clean speech from children between 6 and 11. Results of these experiments show that the human word recognition error rate increases as the age of the child decreases, confirming that recognition of children speech is a challenging task, especially when targeting younger children.

4 Acoustic modeling

This section presents solutions that were investigated to cope with acoustic variability mainly introduced by speaker-specific factors. The investigated solutions include cepstral mean and variance normalization, speaker adaptive acoustic modeling techniques, for acoustic models based on continuous density hidden Markov models (HMMs), and unsupervised speaker adaptation.

4.1 Cepstral mean and variance normalization

In this work, 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, on a speaker-by-speaker basis, 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 as 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 performed by forcing each acoustic feature to have zero mean and unit variance over the speaker's data. The aim was to reduce acoustic variations induced by speaker specific factors, acquisition channel, and environment.

4.2 Speaker adaptive acoustic modeling

Speaker adaptive acoustic modeling aims at reducing or compensating for acoustic variations induced by different characteristics of each training and

testing speaker. In this work we investigate speaker adaptive acoustic modeling through three acoustic feature transformation approaches: VTLN; a variant of the SAT algorithm proposed by Gales (Gales, 1998); and CMLSN, a related method to SAT. Speaker adaptively trained models are typically used in a two-pass decoding scheme in which the output of a first decoding pass, employing conventionally trained acoustic models, is used as a supervision for normalization/adaptation purposes before a second decoding pass.

VTLN Vocal tract length normalization aims at reducing inter-speaker acoustic variability due to vocal tract length (and shape) variations among speakers by warping the frequency axis of the speech power spectrum (Lee and Rose, 1996; Wegmann et al., 1996; Eide and Gish, 1996). In VTLN, typical issues are the estimation of a proper frequency scaling factor for each speaker, or utterance, and the implementation of the frequency scaling during speech analysis. A well known method for estimating the scaling factor is based on a grid search over a discrete set of possible scaling factors by maximizing the likelihood of warped data given a current set of acoustic models (Lee and Rose, 1996). Normalization can be performed on the test data only or both on the training and test data. Frequency scaling is performed by warping the power spectrum during signal analysis or, for filter-bank based acoustic front-end, by changing the spacing and width of the filters while keeping the spectrum unchanged (Lee and Rose, 1996).

In this work, VTLN was performed on a speaker-by-speaker basis both on the training and test data. Frequency warping was implemented by changing the spacing and width of the filters in the mel filter-bank while keeping the speech spectrum unchanged. To cope with the problem of accommodating filters near the band edge, a piecewise linear warping function of the frequency axis of the mel filter-bank was adopted. During training the reference acoustic models for scaling factor selection were speaker independent (SI) triphone HMMs with 1 Gaussian per state, trained on unwarped data. During testing, scaling factor selection was instead performed with respect to the HMMs, trained on warped data, used for the final decoding step. In all cases a grid search over 21 warping factors evenly distributed, with step 0.02, in the range 0.80-1.20, was performed. The training and recognition procedures adopted for implementing VTLN were very similar to those proposed in (Welling et al., 1999) and are described in detail in (Giuliani et al., 2006).

SAT The variant of the SAT algorithm proposed by Gales (Gales, 1998) was used in this work. This variant makes use of an affine transformation, estimated through constrained maximum likelihood linear regression (MLLR), for mapping acoustic observations of each training and testing speaker, instead

of adapting model parameters (Anastasakos et al., 1996). Transformation parameters are estimated with the aim of reducing the acoustic mismatch between speaker's data and the reference models. With this method, a set of SI continuous density HMMs is first fully trained on unnormalized data and then used as seed models. Then, the parameters of speaker-specific affine transformations and the parameters of the Guassian densities are jointly estimated by way of an iterative procedure which alternates estimation of transformations with respect to the current models and estimation of model parameters on the data normalized with the current transformations.

The resulting normalized models are used for decoding on normalized test data. Before decoding, data of each test speaker are normalized through the application of an affine transformation iteratively estimated adopting a procedure similar to the one used in training, except that in this case model parameters are not updated.

CMLSN The CMLSN method performs speaker normalization by transforming the acoustic observation vectors by means of speaker-specific affine transformations, estimated through constrained MLLR. However, in contrast to the variant of SAT proposed by Gales in (Gales, 1998), speaker-specific transformations are estimated with the aim of reducing the acoustic mismatch of the speaker data with respect to a set of target HMMs which is different from the HMM set to be used for recognition. With this method, in fact, the structure of the target and recognition models are determined independently. For example, in this work target models are triphone HMMs with a single Gaussian density per state. However, in (Stemmer et al., 2005) it was shown that a Gaussian mixture model can be effectively used as a target model leading to a text-independent speaker normalization technique.

First, target models are trained on unnormalized data. Data of each training speaker are then normalized by means of an affine transformation estimated through constrained MLLR with respect to the target models. After training data have been transformed, recognition models are trained from scratch by using a conventional training procedure. At the recognition stage, speaker data are normalized with respect to target models before decoding with recognition models trained on normalized data.

A potential advantage over the SAT method (Gales, 1998) is that inter-speaker acoustic variability is reduced before making any decision or performing any training step. In the SAT method, in fact, an adaptive training scheme is added on top of a conventional training procedure. We argue that during the conventional training of the SI seed models, inter-speaker variability has already affected parameter estimation and therefore SAT can only alleviate its effect. Furthermore, the state tying determined during the conventional

training procedure can not be changed during the SAT iterations, while in the CMLSN method state tying is determined by exploiting normalized acoustic data. For a detailed discussion of the differences between the SAT and the CMLSN methods the reader is referred to (Stemmer et al., 2005).

4.3 Unsupervised speaker adaptation

When a two-pass decoding scheme was adopted, unsupervised speaker adaptation was performed by adapting means and variances of Gaussian densities through MLLR (Leggetter and Woodland, 1995). Two regression classes were defined and the associated transformation matrices were estimated through three MLLR iterations exploiting the data of each speaker. Full transformation matrices were used for transforming the means, while diagonal transformation matrices were used for transforming the variances.

5 Recognition experiments

In this section, results of several speech recognition experiments are reported. These experiments concern recognition of adult and children's speech with group-specific acoustic models, trained separately on adult speech and children's speech, and with age-independent acoustic models, trained on a mixture of adult and children's speech.

5.1 Experimental setup

For acoustic models we employed state-tied, cross-word triphone HMMs. Output distributions associated with HMM states were modeled with mixtures of up to 8 diagonal covariance Gaussian densities. In all acoustic model sets trained, "silence" was modeled with a single state HMM. In addition, we trained a number of models for common extra linguistic phenomena, such as human noises (breathing, lip smacks, etc.), non-verbal sounds and filled pauses.

Two language models were estimated and used in speech recognition experiments reported in this paper. For experiments on the IBN, Tgr-child, and Tgr-adult test sets, 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 (Bertoldi et al., 2001). The second language model, used for recognition experiments on the ChildIt test set, was an 11k word 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 perplexity and the out-of-vocabulary (OOV) rate computed on the test sets are reported in Table 2. The high perplexity shown by the 11k word 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 well reflect the statistics of the ChildIt test set, extracted from literature for children.

To assess methods for speaker adaptive acoustic modeling, recognition experiments were carried out adopting a two-pass decoding scheme, assuming all the data of each test speaker was available in one block. For this purpose, we exploited the manual annotation of the speaker identity, as well as the manual segmentation in utterances. The decoder was run twice, and the word transcriptions generated with the first decoding pass were used as a supervision for speaker normalization/adaptation purposes before the second decoding pass took place.

5.2 Preliminary experiments

We trained two sets of SI, tied state, cross-word triphone HMMs for children and adults, respectively. AMs for children were trained using the ChildIt training set and the SpontIt corpus — about 9 hours of speech — resulting in about 1700 independent states and 13200 Gaussian densities. Adult HMMs were trained using the IBN training set — about 57 hours of speech — resulting in about 6700 independent states and 53860 Gaussian densities. Table 3 reports the results achieved by performing a single recognition pass with group-specific acoustic models for adults and children.

	Test Set			
HMM set	IBN	Tgr-adult	Tgr-child	ChildIt
Adult HMMs (57h)	16.1	10.4	37.2	41.0
Adult HMMs (9h)	18.8	12.9	32.5	36.8
Child HMMs (9h)	54.4	45.4	14.2	14.4

Table 3
Recognition results (% WER) obtained with acoustic models trained on adults and children by performing a single decoding pass.

Looking at the results on Tgr-adult and Tgr-child corpora, we note that under matched conditions recognition performance for adults is much better than that for children: 10.4% WER compared with 14.2% WER, respectively. However, we point out that much more training data were used for adults than for

children (i.e. 57 hours vs 9 hours) and therefore the performance gap could be partially filled by just having more training data from children.

To measure the influence of the amount of training data, a contrasting experiment was carried out by training a set of adult HMMs using only 9h of speech selected from the IBN training set. Recognition results on the 4 test sets using these AMs are also reported in Table 3, row "Adult HMMs (9h)". On the IBN and Tgr-adult test sets 18.8% and 12.9% WER were achieved, respectively. We can conclude that when using the same amount of training data the difference in performance in matched conditions between adult and children's speech, measured on the Tgr-adult and Tgr-child corpora, is only of about 10% relative - 12.9% WER compared to 14.2% WER. The HMM set trained on 9 hours of adult speech performs better than the HMM set trained on all adult data (57 hours) when used to recognize children's speech. This is probably due to the composition of the training data selected.

Results of Table 3 show that, 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 characteristics of adult and children's voices (Wilpon and Jacobsen, 1996; Potamianos and Narayanan, 2003; Lee et al., 1999; Gerosa et al., 2006b).

We investigated the impact of performing mean and variance normalization of acoustic features by adopting the two acoustic front-ends described in Section 4.1, denoted as MFCC39-MVN13 and MFCC39-MVN39. These experiments were motivated by the fact that the analysis on phone duration, presented in Section 3, revealed that adults and children in the speech corpora used in this work presented 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 (Martinez et al., 1998), therefore performing mean and variance normalization of dynamic features could be useful to compensate for very different speaking rates.

Several sets of HMMs were trained for adults and children, by exploiting the acoustic observations generated by the different acoustic front-ends. Recognition results, achieved with a single decoding pass, are reported in Table 4. As a reference, results obtained using the standard front-end ("MFCC39") are also reported.

We note that mean and variance normalization carried out on all acoustic features ("MFCC39-MVN39") ensures systematic benefits with respect to adopting the standard acoustic front-end. Improvement in recognition performance was validated using the matched-pair sentence test (Gillick and Cox, 1989) to ascertain whether the observed results were inconsistent with the null hypoth-

esis that the output of two systems were statistically identical. Considered significance levels were .05, .01 and .001. In case of unmatched training and testing conditions the improvement is statistically significant for all test sets, with $p < .001$ for the ChildIt, Tgr-child and IBN test sets, and $p < .05$ for the Tgr-adult test set. In the case of matched conditions, the improvement is significant for the IBN and ChildIt test sets(with $p < .05$) and not significant for the Tgr-child and Tgr-adult test sets. Normalizing just the static acoustic features ("MFCC39-MVN13") is less effective and consistent.

Therefore, the following recognition experiments were carried out performing mean and variance normalization on all acoustic features. Word level transcriptions corresponding to recognition results reported in rows "MFCC39-MVN39" were exploited in two-pass recognition experiments as supervision for adaptation/normalization purposes, before performing the second decoding pass.

		Test Set			
HMM set	Feature Set	IBN	Tgr-adult	Tgr-child	ChildIt
Adult	MFCC39	16.1	10.4	37.2	41.0
HMMs	MFCC39-MVN13	15.7	10.2	36.4	42.5
	MFCC39-MVN39	15.6	10.1	33.3	39.9
Child	MFCC39	54.5	45.4	14.2	14.4
HMMs	MFCC39-MVN13	56.0	48.1	14.1	14.0
	MFCC39-MVN39	51.2	44.1	13.8	13.9

Table 4
Recognition results (% WER) obtained with acoustic models trained on adults and children and by adopting different acoustic front-ends.

5.3 Adaptive acoustic modeling

We trained for both adults and children three HMM sets using the VTLN, CMLSN and SAT training methods summarized in Section 4, with the aim of reducing the effect of inter-speaker acoustic variability and improving recognition performance. All HMM sets trained on the same group of speakers had a number of parameters similar to the one of the corresponding baseline HMMs.

In the experiments reported below, in addition to speaker normalization, unsupervised static speaker adaptation of acoustic models was performed before the second decoding pass. Speaker adaptation was performed by adapting means and variances of Gaussian densities through MLLR as described in Section 4.3. Figure 2 reports recognition results obtained on the four test sets by using

acoustic models trained by way of the VTLN, CMLSN and SAT methods. For comparison purposes, results achieved with baseline HMMs and unsupervised static speaker adaptation ("Two-pass Baseline") are also reported.

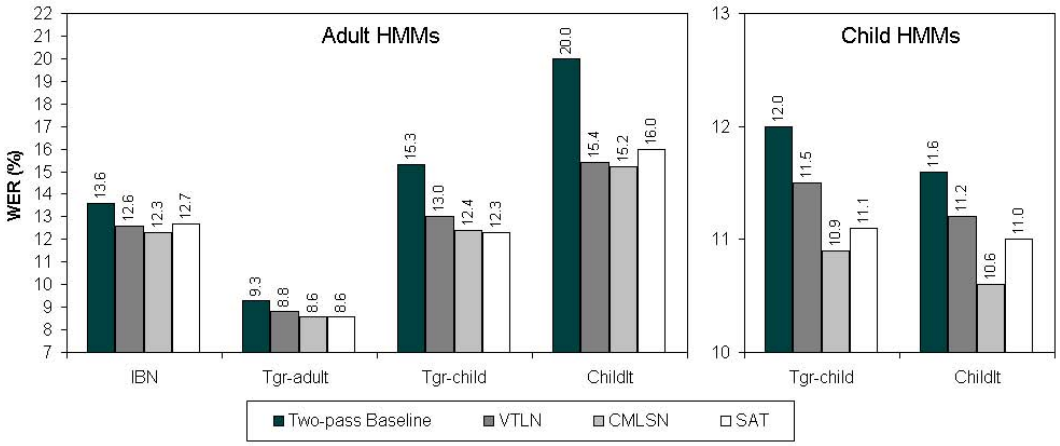


Fig. 2. *Recognition results (% WER) obtained using HMMs trained on adult speech ("Adult HMMs") and children's speech ("Child HMMs") with three speaker adaptive acoustic modeling methods.*

By comparing results achieved by the two-pass baseline with corresponding results reported in Table 4 (rows "MFCC39-MVN39"), it can be noted that the WER reduction achieved by unsupervised AM adaptation is tangible, especially in the case of unmatched conditions. Furthermore, performing unsupervised adaptation of speaker adaptively trained AMs results in improved performance.

When comparing speaker adaptive acoustic modeling methods, we can see that the CMLSN method outperforms the VTLN method in both matched and unmatched conditions. In matched conditions this improvement is statistically significant for the IBN (with $p < .01$), Tgr-child (with $p < .05$), and ChildIt (with $p < .01$) test sets, while it is not significant for the Tgr-adult test set. In unmatched conditions (see "Tgr-child" and "ChildIt" bars in the "Adult HMMs" part of the Figure) the difference in performance between the CMLSN and the VTLN methods is significant for the Tgr-child test set ($p < .05$) and not significant for the ChildIt test set.

SAT and CMLSN give similar results on Tgr-adult and Tgr-child corpora, while on IBN and ChildIt test sets using CMLSN ensures an improvement in performance. This improvement was found significant with $p < .01$ for the IBN test set, with $p < .001$ for the ChildIt test set in unmatched conditions, and with $p < .01$ for the ChildIt test set in matched conditions. We point out that the recognition results achieved on the Tgr-child and ChildIt test sets using models trained on adult data using VTLN, CMLSN, and SAT are still worse than those achieved by the baseline models trained on children's speech.

Finally, it can be noted that, when the CMLSN method is adopted, in matched conditions the gap between recognition performance achieved on the Tgr-adult test set, 8.6% WER, and the Tgr-child test set, 10.9% WER, is about 27% relative.

5.4 Age-independent acoustic modeling

The use of age-independent acoustic models was investigated by training on a mixture of speech data collected from adults and children of different ages. Our goal was to develop a set of acoustic models able to ensure good recognition performance for adult and child speakers. The use of adult speech for reinforcing the training data in the case of a dearth of children's speech has been investigated in several papers (Wilpon and Jacobsen, 1996; Steidl et al., 2003). However, simply adding adult data to the child training data always resulted in a degradation in recognition performance for child speakers due to the acoustic mismatch between the voices of adults and those of children.

In this work, we started from an experimental condition in which much more training data were available for adults than for children. However we didn't want to sacrifice any training data available for adults in order to balance training data between adults and children, as we were also interested in maintaining good recognition performance for adults. Moreover this represents a very common experimental condition, as for all languages in the past much more data for adults were collected than for children.

Using 66 hours of speech, a cross-word triphone HMM set with 7320 tied states and about 58800 Gaussian densities was trained. In addition to these baseline models, three HMM sets were trained on the same data, using the VTLN, CMLSN, and SAT training procedures. Figure 3 reports recognition results achieved on the four test sets with the models trained on adult and children's speech. As before, at the recognition stage, in addition to speaker normalization, unsupervised static MLLR model adaptation was performed before the second decoding pass.

By comparing results reported in Figure 3 with those reported in Figure 2, it is clear that simply training with mixed data results in a decrease in performance for both adults and children (compare results for "Two-pass Baseline"). On the other hand, speaker adaptive acoustic modeling with mixed data proves to be very effective. In fact, recognition results achieved on the IBN, Tgr-adult, Tgr-child and ChildIt test sets using age-independent HMMs trained adopting the CMLSN method, 12.5%, 8.5%, 10.2%, 10.7% WER respectively, are similar to those achieved on the same test sets with the group-specific HMMs for adults and children reported in Figure 2, 12.3%, 8.6%, 10.9% and

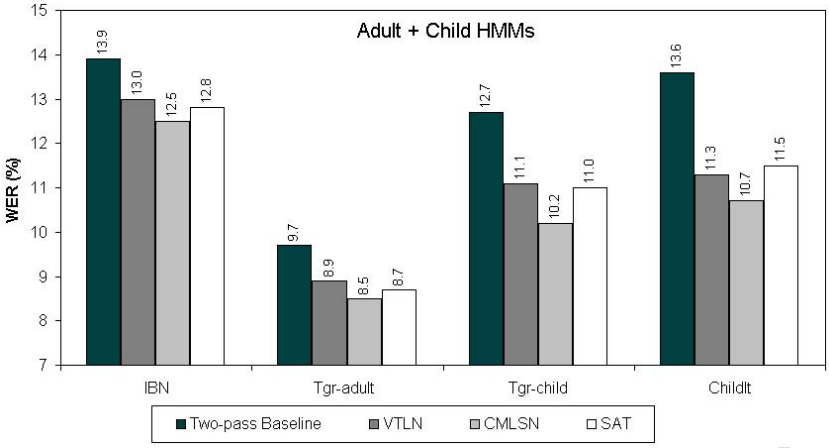


Fig. 3. *Recognition results (% WER) obtained using HMMs trained on adult and children's speech data, with and without speaker adaptive acoustic modeling methods.*

10.6% WER, respectively.

5.5 Combination of methods

We investigated the improvement that could be further gained by combining methods for speaker adaptive acoustic modeling. In particular we investigated the combination of VTLN with SAT, and of VTLN with CMLSN. In fact, while in CMLSN and SAT no assumption is made about the nature of the acoustic mismatch between the data of different speakers, the VTLN method is conceived to reduce spectral differences induced specifically by variations in vocal tract length. So, these methods could have some additive effects.

Figure 4 reports the results achieved with the second decoding pass using HMMs trained by applying the two combinations of speaker adaptive acoustic modeling methods ("VTLN+SAT" and "VTLN+CMLSN"). For comparison purposes, results achieved with two-pass baseline systems are also reported. Results on all the four test sets confirm the effectiveness of speaker adaptive acoustic modeling methods. For example by considering group-specific models for adults, the baseline system guarantees a 9.3% WER on the Tgr-adult test set. In contrast, we can achieve an 8.2% WER by cascading methods for speaker adaptive acoustic modeling ("VTLN+CMLSN"). By considering group-specific models for children, the baseline system guarantees a 12.0% WER on the Tgr-child test set, compared with a 10.5% WER achieved by cascading methods for speaker adaptive acoustic modeling ("VTLN+CMLSN").

By comparing results reported in Figure 4 with those reported in Figure 2 and Figure 3, it can be noted that results achieved by cascading the VTLN

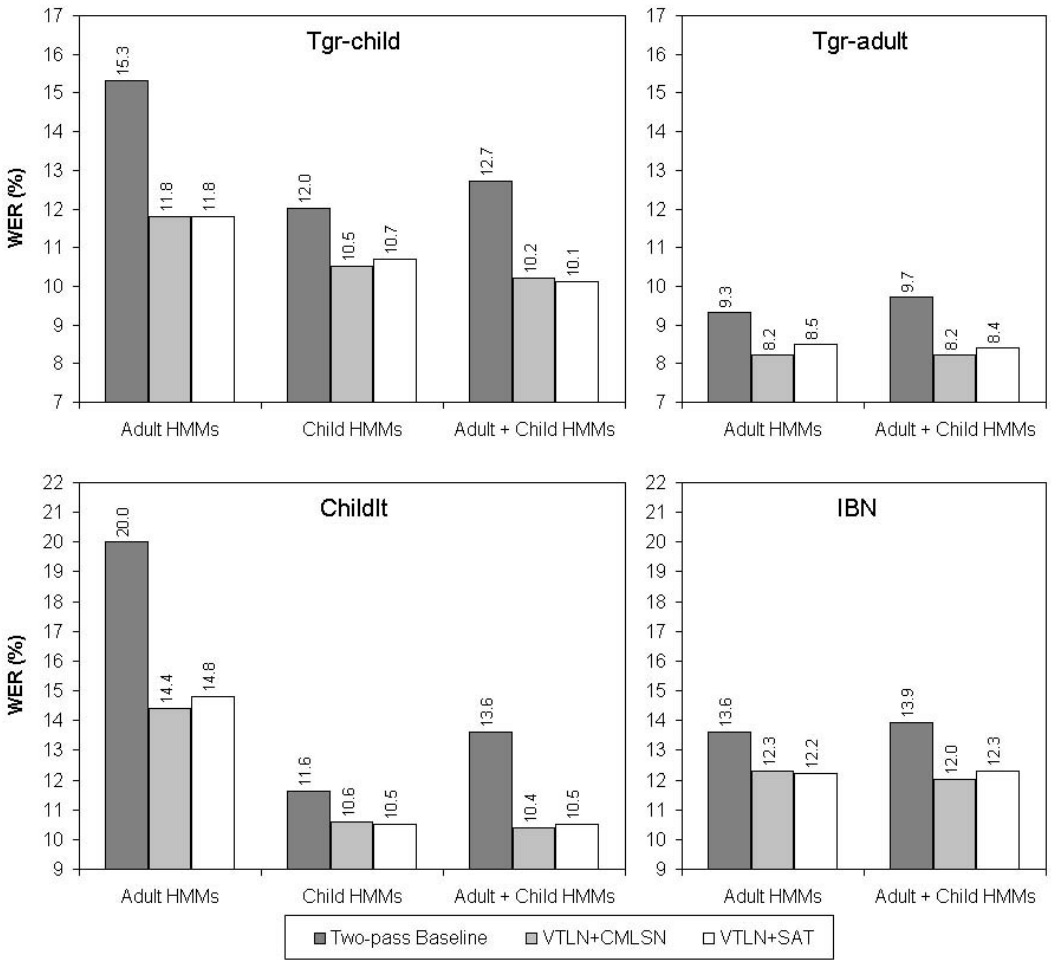


Fig. 4. *Recognition results (% WER) on the 4 test sets obtained using HMMs trained on adult speech, on adult and children's speech, and on children's speech with and without speaker adaptive acoustic modeling methods.*

method with the CMLSN and SAT methods are always equal to or better than those achieved using one of the normalization methods alone.

Most importantly, age-independent HMMs, trained by cascading methods for speaker adaptive acoustic modeling, ensure performance aligned with that achieved by age-specific acoustic models trained with the same methods (compare results for "VTLN+CMLSN" in Figure 4).

6 Conclusions

In this paper, age-independent acoustic modeling has been addressed with the aim of developing acoustic models able to perform well on speakers of all ages.

Recognition experiments were carried out exploiting four test sets composed of Italian speech, including two parallel test corpora made of repetitions of the same sentences read by adults and 8-12 year-old children. This allowed direct comparison between results achieved on adult and children's speech in the context of a large-vocabulary (64k words) speech transcription task.

Speaker adaptive acoustic modeling was investigated through the use of the VTLN, CMLSN, and SAT methods and their combinations. These methods proved to be effective when used to train group-specific acoustic models for adults and children. In matched conditions, we obtained WER relative reductions, with respect to the baseline systems, between 8.6% and 12.5%.

More importantly, speaker adaptive acoustic modeling proved to be effective when applied to train age-independent acoustic models by exploiting speech from adult and child speakers. Recognition results achieved using age-independent acoustic models were, in fact, aligned with those achieved using the group-specific HMMs for adults and children. Developing acoustic models able to perform well on speakers of all ages can be important in application scenarios in which speech recognition technology is applied to recognize speech from speakers of unknown age such as automatic transcription of audio-visual documents (for example, TV programs) and voice interactive services over the telephone line. It may also reduce the need of collecting large amounts of training data from speakers of each age group.

Furthermore, on the parallel corpora consisting of the same sentences read by adults and children, the WER achieved for children, 10.2%, was only 24% (relative) higher than the WER achieved for adult, 8.2%, thus demonstrating that for the age-range considered, 8-12 years, large vocabulary recognition of read children's speech is a feasible task.

7 Acknowledgments

This work was partially financed by the European Union under the projects PF-STAR (grant IST-2001-37599, http://pfstar.itc.it) and TC-STAR (grant FP6-506738, http://www.tc-star.org).

References

Ackermann, U., Angelini, B., Brugnara, F., Federico, M., Giuliani, D., Gretter, R., and Niemann, H. (1997). Speedata: A Prototype for Multilingual Spoken Data-Entry. In *Proc. of EUROSPEECH*, pages 1807–1810, Rhodes, Greece.

Anastasakos, T., McDonough, J., Schwartz, R., and Makhoul, J. (1996). A Compact Model for Speaker-Adaptive Training. In *Proc. of ICSLP*, pages 1137–1140, Philadelphia, PA.

Angelini, B., Brugnara, F., Falavigna, D., Giuliani, D., Gretter, R., and Omologo, M. (1994). Speaker Independent Continuous Speech Recognition Using an Acoustic-Phonetic Italian Corpus. In *Proc. of ICSLP*, pages 1391–1394, Yokohama, Japan.

Batliner, A., Blomberg, M., D'Arcy, S., Elenius, D., Giuliani, D., Gerosa, M., Hacker, C., Russell, M., Steidl, S., and Wong, M. (2005). The PF-STAR Children's Speech Corpus. In *Proc. of INTERSPEECH*, pages 2761–2764, Lisboa, Portugal.

Bertoldi, N., Brugnara, F., Cettolo, M., Federico, M., and Giuliani, D. (2001). From Broadcast News to Spontaneous Dialogue Transcription: Portability Issues. In *Proc. of ICASSP*, volume 1, pages 37–40, Salt Lake City, UT.

Brugnara, F., Cettolo, M., Federico, M., and Giuliani, D. (2002). Issues in automatic transcription of historical audio data. In *Proc. of ICSLP*, pages 1441–1444, Denver, CO.

Burnett, D. C. and Fanty, M. (1996). Rapid Unsupervised Adaptation to Children's Speech on a Connected-Digit Task. In *Proc. of ICSLP*, volume 2, pages 1145–1148, Philadelphia, PA.

Claes, T., Dologlou, I., ten Bosch, L., and Compernolle, D. V. (1998). A Novel Feature Transformation for Vocal Tract Length Normalization in Automatic Speech Recognition. *IEEE Trans. on Speech and Audio Processing*, 6(6):549–557.

D'Arcy, S. and Russell, M. (2005). A Comparison of Human and Computer Recognition Accuracy for Children's Speech. In *Proc. of INTERSPEECH/ICSLP*, pages 2197–2199, Lisboa, Portugal.

Das, S., Nix, D., and Picheny, M. (1998). Improvements in Children's Speech Recognition Performance. In *Proc. of ICASSP*, pages 433–436, Seattle,WA.

Eide, E. and Gish, H. (1996). A Parametric Approach to Vocal Tract Lenght Normalization. In *Proc. of ICASSP*, pages 346–349, Atlanta, GA.

Fitch, W. T. and Giedd, J. (1999). Morphology and development of the human vocal tract: A study using magnetic resonance imaging. *Journal of Acoust. Soc. Amer.*, 106(3):1511–1522.

Gales, M. J. F. (1998). Maximum likelihood linear transformations for HMM-based speech recognition. *Computer Speech and Language*, 12:75–98.

Gerosa, M., Giuliani, D., and Brugnara, F. (2007). Acoustic variability and automatic recognition of children's speech. *Speech Communication*, 49:847–869.

Gerosa, M., Giuliani, D., and Narayanan, S. (2006a). Acoustic analysis and automatic recognition of spontaneous children's speech. In *Proc. of INTERSPEECH/ICSLP*, Pittsburgh, PA.

Gerosa, M., Lee, S., Giuliani, D., and Narayanan, S. (2006b). Analyzing Children's Speech: An acoustic Study of Consonants and Consonant-Vowel Transition. In *Proc. of ICASSP*, pages 393–396, Toulouse, France.

Gillick, L. and Cox, S. J. (1989). Some Statistical Issues in the Comparison of Speech Recognition Algorithms. In *Proc. of ICASSP*, pages I–532–535, Glasgow, UK.

Giuliani, D. and Gerosa, M. (2003). Investigating Recognition of Children Speech. In *Proc. of ICASSP*, volume 2, pages 137–140, Hong Kong.

Giuliani, D., Gerosa, M., and Brugnara, F. (2004). Speaker Normalization through Constrained MLLR Based Transforms. In *Proc. of INTERSPEECH*, pages 2893–2897, Jeju Island, Korea.

Giuliani, D., Gerosa, M., and Brugnara, F. (2006). Improved automatic speech recognition through speaker normalization. *Computer Speech and Language*, 20(1):107–123.

Goldstein, U. G. (1980). An Articulatory Model for the Vocal Tract of Growing Children. *Ph. D. Thesis, MIT, Cambridge (MA)*.

Hagen, A., Pellom, B., and Cole, R. (2003). Children's Speech Recognition with Application to Interactive Books and Tutors. In *Proc. of IEEE Automatic Speech Recognition and Understanding (ASRU) Workshop*, St. Thomas, US Virgin Islands.

Huber, J. E., Stathopoulos, E. T., Curione, G. M., Ash, T. A., and Johnson, K. (1999). Formants of children women and men: The effect of vocal intensity variation. *Journal of Acoust. Soc. Amer.*, 106(3):1532–1542.

Lee, L. and Rose, R. C. (1996). Speaker Normalization Using Efficient Frequency Warping Procedure. In *Proc. of ICASSP*, pages 353–356, Atlanta, GA.

Lee, S., Potamianos, A., and Narayanan, S. (1999). Acoustic of children's speech: Developmental changes of temporal and spectral parameters. *Journal of Acoust. Soc. Amer.*, 105(3):1455–1468.

Leggetter, C. J. and Woodland, P. C. (1995). Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Computer Speech and Language*, 9:171–185.

Li, Q. and Russell, M. (2002). An Analysis of the Causes of Increased Error Rates in Children's Speech Recognition. In *Proc. of ICSLP*, pages 2337–2340, Denver, CO.

Martinez, F., Tapias, D., and Alvarez, J. (1998). Towards Speech Rate Independence in Large Vocabulary Continuous Speech Recognition. In *Proc. of ICASSP*, volume 2, pages 725–728, Seattle,WA.

McGowan, R. S. and Nittrouer, S. (1988). Differences in fricative production between children and adults: Evidence from an acoustic analysis of /f/ and /s/. *Journal of Acoust. Soc. Amer.*, 83(1):229–236.

Mirghafori, N., Fosler, E., and Morgan, N. (1996). Towards Robustness to Fast Speech in ASR. In *Proc. of ICASSP*, pages 335–338, Atlanta, GA.

Narayanan, S. and Potamianos, A. (2002). Creating Conversational Interfaces for Children. *IEEE Trans. on Speech and Audio Processing*, 10(2):65–78.

Nisimura, R., Lee, A., Saruwatari, H., and Shikano, K. (2004). Public Speech-Oriented Guidance System with Adult and Child Discrimination Capability. In *Proc. of ICASSP*, volume 1, pages 433–436, Montreal, Canada.

Nittrouer, S. and Whalen, D. H. (1989). The perceptual effects of child-adult differences in fricative-vowel coarticulation. *Journal of Acoust. Soc. Amer.*, 86(4):1266–1276.

Pallett, D. S., Fiscus, J. G., and Garofolo, J. S. (1992). Resource Management Corpus: September 1992 Test Set Benchmark Test Results. In *Proceedings of the 1992 ARPA's Continuous Speech Recognition Workshop*, Stanford, CA.

Potamianos, A. and Narayanan, S. (2003). Robust Recognition of Children's Speech. *IEEE Trans. on Speech and Audio Processing*, 11(6):603–615.

Potamianos, A., Narayanan, S., and Lee, S. (1997). Automatic Speech Recognition for Children. In *Proc. of EUROSPEECH*, pages 2371–2374, Rhodes, Greece.

Steidl, S., Stemmer, G., Hacker, C., Nöth, E., and Niemann, H. (2003). Improving Children's Speech Recognition by HMM Interpolation with an Adults' Speech Recognizer. In *Pattern Recognition, 25th DAGM Symposium*, pages 600–607.

Stemmer, G., Brugnara, F., and Giuliani, D. (2005). Using Simple Target Models for Adaptive Training. In *Proc. of ICASSP*, volume 1, pages 997–1000, Philadelphia, PA.

Stemmer, G., Hacker, C., Steidl, S., and Nöth, E. (2003). Acoustic Normalization of Children's Speech. In *Proc. of EUROSPEECH*, pages 1313–1316, Geneva, Switzerland.

Wegmann, S., McAllaster, D., Orloff, J., and Peskin, B. (1996). Speaker Normalisation on Conversational Telephone Speech. In *Proc. of ICASSP*, pages I–339–341, Atlanta.

Welling, L., Kanthak, S., and Ney, H. (1999). Improved methods for vocal tract normalization. In *Proc. of ICASSP*, volume 2, pages 761–764, Phoenix, AZ.

Wilpon, J. G. and Jacobsen, C. N. (1996). A Study of Speech Recognition for Children and Elderly. In *Proc. of ICASSP*, pages 349–352, Atlanta, GA.

Figure 1

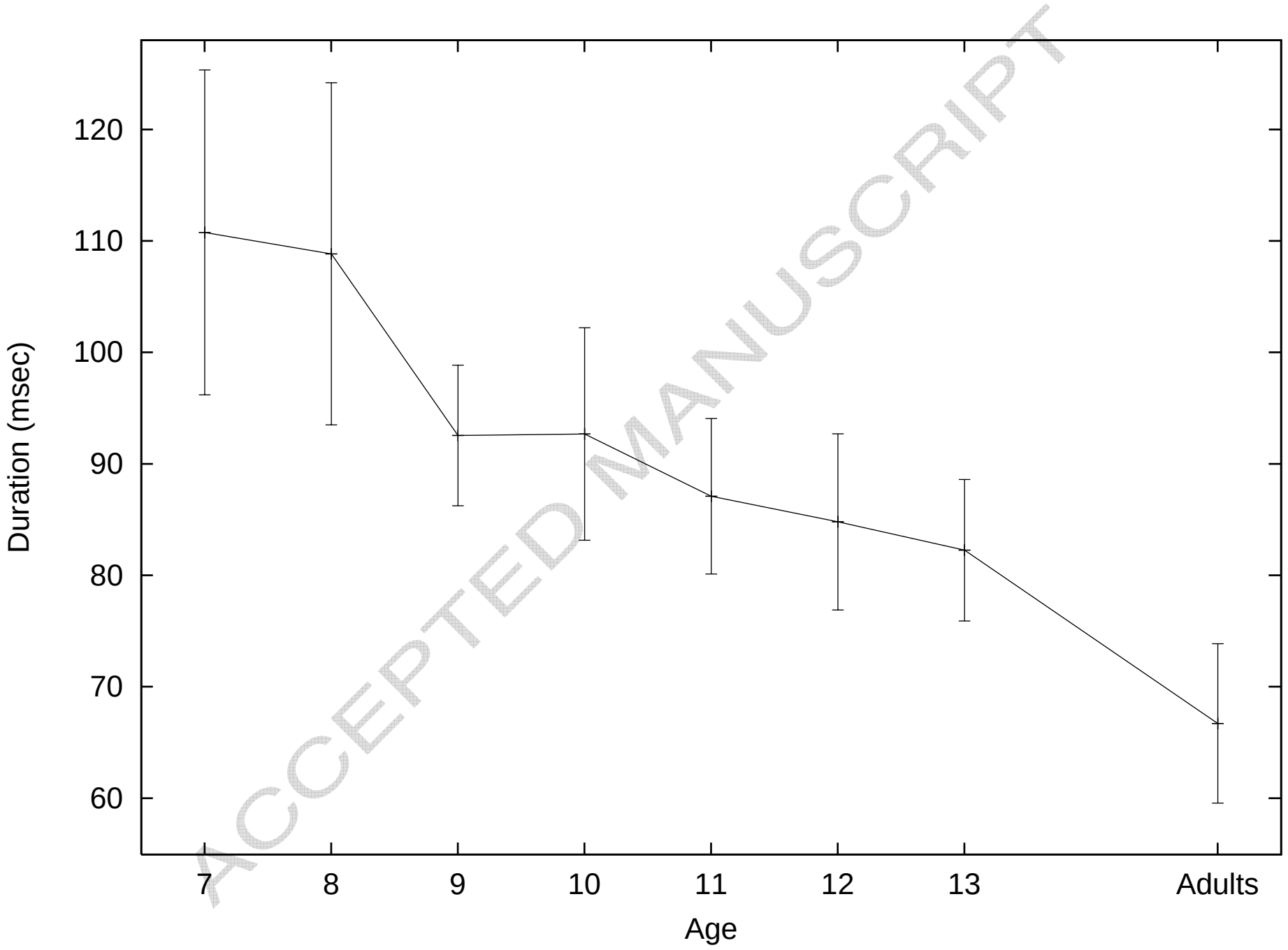


Figure 2

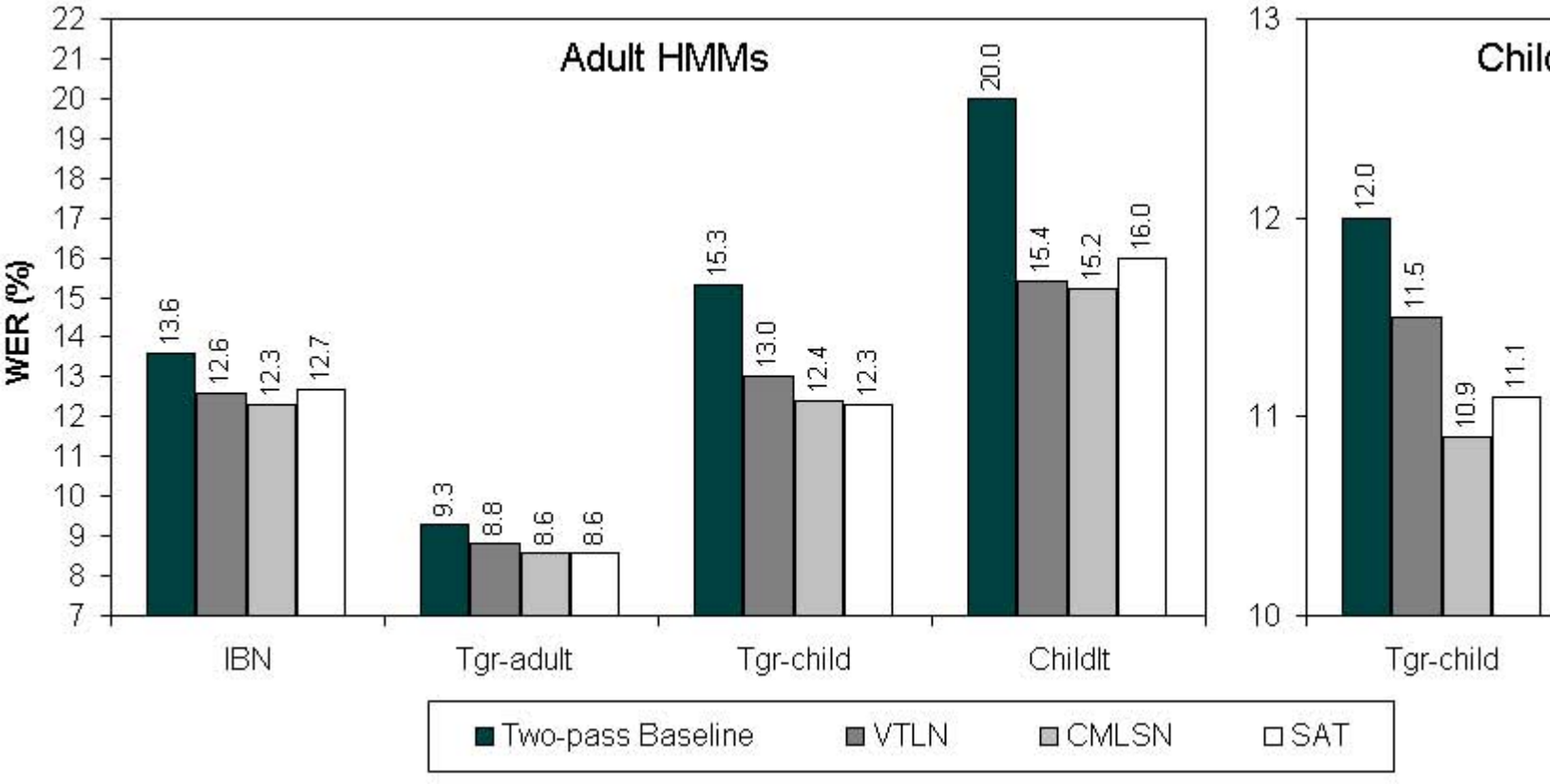


Figure 3

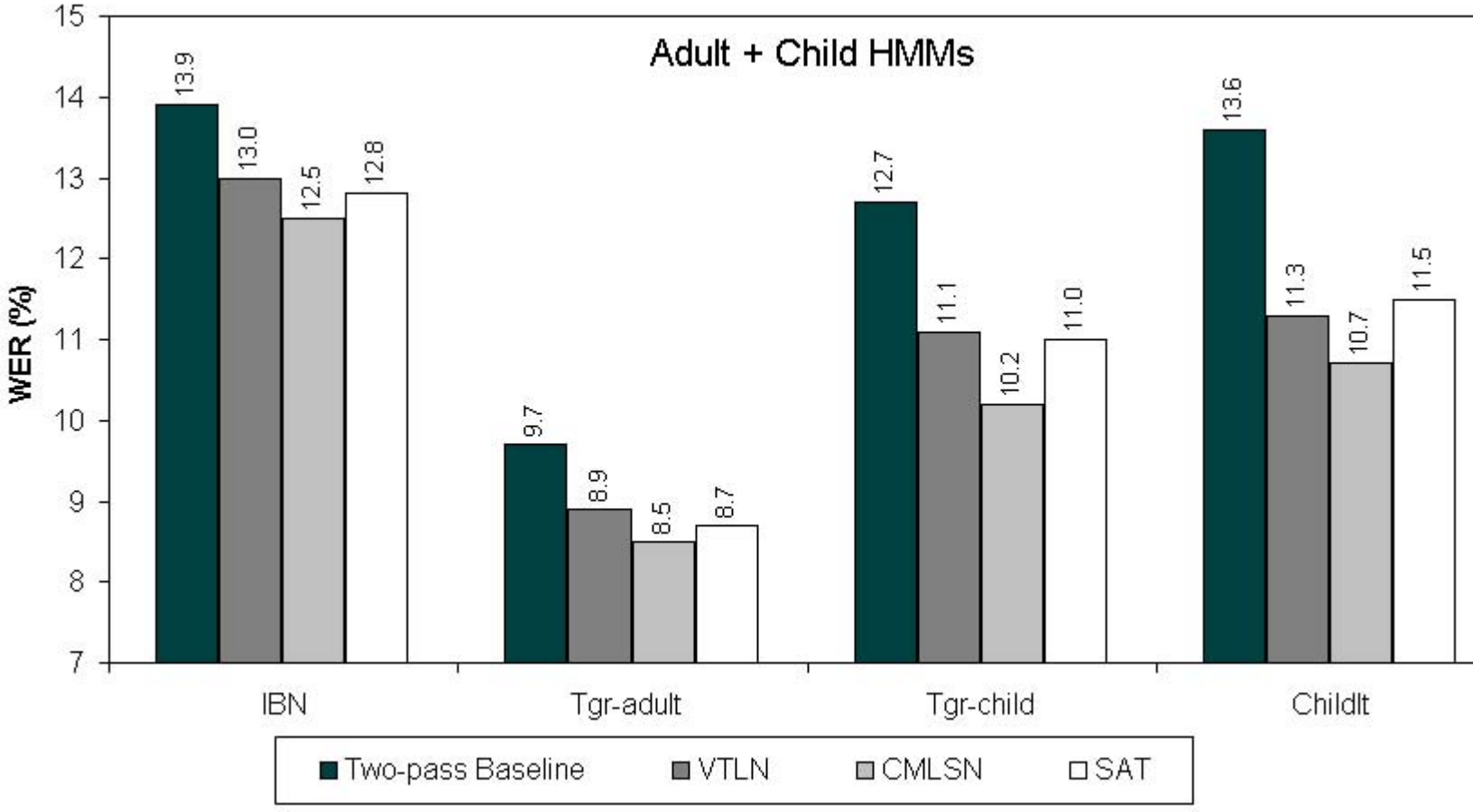


Figure 4