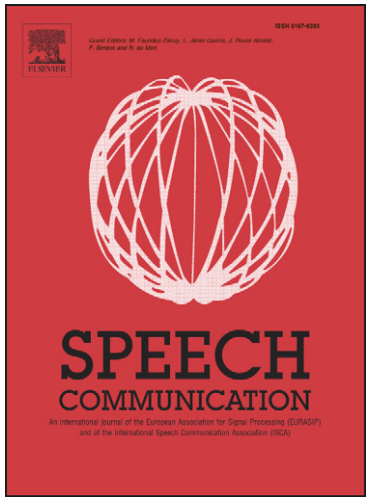
Acoustic Variability and Automatic Recognition of Children's Speech

Matteo Gerosa, Diego Giuliani, Fabio Brugnara



Acoustic Variability and Automatic Recognition of Children's Speech

Matteo Gerosa, Diego Giuliani, Fabio Brugnara

PII: S0167-6393(07)00005-2
DOI: 10.1016/j.specom.2007.01.002
Reference: SPECOM 1605

To appear in: *Speech Communication*

Received Date: 31 March 2006
Revised Date: 22 December 2006
Accepted Date: 11 January 2007

Please cite this article as: Gerosa, M., Giuliani, D., Brugnara, F., Acoustic Variability and Automatic Recognition of Children's Speech, *Speech Communication* (2007), doi: 10.1016/j.specom.2007.01.002

Acoustic Variability and Automatic Recognition of Children's Speech

Matteo Gerosa *, Diego Giuliani and Fabio Brugnara

ITC-irst, Centro per la Ricerca Scientifica e Tecnologica
I-38050 Povo (Trento), Italy
gerosa@itc.it, giuliani@itc.it, brugnara@itc.it

Abstract

This paper presents several acoustic analyses carried out on read speech collected from Italian children aged from 7 to 13 years and North American children aged from 5 to 17 years. These analyses aimed at achieving a better understanding of spectral and temporal changes in speech produced by children of various ages in view of the development of automatic speech recognition applications. The results of these analyses confirm and complement the results reported in the literature, showing that characteristics of children's speech change with age and that spectral and temporal variability decrease as age increases. In fact, younger children show a substantially higher intra- and inter-speaker variability with respect to older children and adults. We investigated the use of several methods for speaker adaptive acoustic modeling to cope with inter-speaker spectral variability and to improve recognition performance for children. These methods proved to be effective in recognition of read speech with a vocabulary of about 11k words.

Key words: Children's speech analysis, automatic speech recognition for children, speaker normalization, speaker adaptive acoustic modeling.

1 Introduction

Speech technology has a huge potential for use by children. In addition to conventional applications in which speech replaces, or complements, other modalities in the human-machine interaction (Gustafson and Sjölander, 2000; Narayanan and Potamianos, 2002; Nisimura et al., 2004), there are applications in which speech is the key enabling technology, including voice interactive computer-based pronunciation or reading tuition and foreign language

* Corresponding author. Tel: +39 0461 314 555; Fax: +39 0461 314 591.

learning (Russell et al., 2000; Eskenazi and Pelton, 2002; Hagen et al., 2003; Banerjee et al., 2003; Mich et al., 2004). For this reason, in recent years an increased attention has been devoted to children as potential users of automatic speech recognition (ASR) technology.

It is well known that acoustic and linguistic characteristics of children's speech are widely different from those of adult speech (Lee et al., 1999; Huber et al., 1999; Arunachalam et al., 2001). For example, children's speech is characterized by higher pitch and formant frequencies with respect to adults' speech. Furthermore, the characteristics of children's speech vary rapidly as a function of age due to the anatomical and physiological changes that occur during a child's growth and because, with age, children become more skilled in articulation.

This is the reason why the performance of an ASR system developed for adult speech decreases drastically when employed to recognize children's speech, especially for younger children (Wilpon and Jacobsen, 1996; Burnett and Fanty, 1996; Potamianos et al., 1997; Das et al., 1998; Claes et al., 1998). Furthermore, recognition performance for children is usually lower than that achieved for adults even when using a recognition system trained on children's speech (Potamianos and Narayanan, 2003; Gerosa et al., 2005). In fact, developmental changes contribute to variation of spectral and temporal parameters of the children's speech signal, resulting in a high inter- and intra-speaker acoustic variability.

Much has been done in the past to analyze the acoustic characteristics of children's and adult speech, with a particular focus on the effect of vocal tract variation on pitch and formant frequency values (Huber et al., 1999; Lee et al., 1999; Whiteside and Hodgson, 2000). Understanding the effects of developmental changes in children's speech can help to devise strategies for dealing with the acoustic mismatch between different age groups. However, almost all these studies were carried out on American English speech with little effort devoted to the analysis of other languages.

In this work, several analyses on children's speech were carried out focusing on aspects that still require further study, including phone duration, inter-speaker spectral variability and intra-speaker spectral and temporal variability (i.e. the variability of segment duration patterns). These analyses were carried out on two corpora of American English and Italian read children's speech, comparing results achieved on children's speech with those achieved on adult speech. The results of the analyses confirmed and complemented the results reported in the literature and improved our understanding of the characteristics of children's speech.

Results of the analyses showed that children in the age range 7-13 are not

a homogeneous group of speakers. Age-dependent variations in formant frequencies introduce, in fact, variability in spectral features across age groups. In ASR applications this variability impacts on recognition performance increasing the word error rate. To cope with age-dependent spectral variability and to improve recognition performance, age-specific acoustic models trained on speech collected from children of the target age or group of ages can be adopted (Wilpon and Jacobsen, 1996; Hagen et al., 2003). However, training age-specific acoustic models is costly as it requires collecting enough speech data for each target age. As a more general solution, to tackle inter-speaker spectral variability in children's speech when training on speech from children of all ages, speaker adaptive acoustic modeling methods can be adopted (Giuliani and Gerosa, 2003; Hagen et al., 2004; Giuliani et al., 2006). In this work, we investigated the use of speaker adaptive acoustic modeling methods, such as vocal tract length normalization (VTLN) (Wegmann et al., 1996; Lee and Rose, 1996; Eide and Gish, 1996), constrained MLLR based speaker normalization (CMLSN) (Giuliani et al., 2006), speaker adaptive training (SAT) (Anastasakos et al., 1996; Gales, 1998) and their combinations. These methods proved to be effective in reducing inter-speaker variability and improved recognition performance on children's speech both in matched conditions, that is training and testing on Italian children aged 7-13, and in unmatched conditions, that is testing on children's speech with models trained on adult speech.

The rest of the paper is organized as follows. The speech corpora used in this work are described in Section 2. Section 3 presents the results of the analyses performed on phone duration, formant patterns and intra-speaker spectral and temporal variability. Methods adopted for speaker adaptive acoustic modeling are presented in Section 4. Recognition experiments are described in Section 5 and final remarks are reported in Section 6, which concludes the paper.

2 Speech Corpora

Several speech corpora were used in this work: three consisting of children's speech and two consisting of adult speech. Main characteristics of the corpora employed are summarized below.

The ChildIt corpus (Giuliani and Gerosa, 2003) is an Italian, task-independent, speech database that consists of clean read speech from children aged from 7 to 13 years, with a mean age of 10 years. Children in ChildIt corpus are evenly distributed by grade, from grade 2 through grade 8. Children in grade 2 were approximately 7 years old while children in grade 8 were approximately 13 years old. Acoustic analyses on the ChildIt corpus were performed on each grade. However, for simplicity, the figures in this paper report the age roughly corresponding to the speaker's grade. 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 concerning 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 corpus was partitioned into a training set, consisting of data from 129 speakers for a total of 7h:47m of speech, and a test set, consisting of data from 42 speakers balanced with respect to age and gender for a total of 2h:29m of speech. This corpus was exploited for speech recognition experiments and for speech analysis purposes.

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 models training.

The CID corpus (Miller et al., 1996) is an American English, task-independent, speech database that consists of read speech from 436 children aged from 5 to 18 and from 56 adults speakers. The data collection was a joint effort of Southwestern Bell Technology Resources and the Central Institute for the Deaf. Recordings were made using a high-fidelity microphone (Bruel & Kjaer model #4179) connected to a real-time waveform digitizer with 20 kHz sampling rate and 16-bit resolution. Audio signals were then down-sampled to 16 kHz before analysis. Only a subset of this database was exploited for acoustic analysis purposes in this work. For this subset, manual segmentation at the phone level was available. This subset consisted of data from five speakers, 3 females and 2 males, for ages 5, 7, 9, 11, 13, 15, 17 and from five adult speakers, for a total of 40 subjects. The speech material analyzed in this paper consisted of repetitions of five phonetically-rich and meaningful sentences. Each sentence was uttered two times by each speaker. Prior to the recording session, target utterances that the speakers, mostly 5 years olds, had difficulty reading were identified and then elicited through imitation of a sample prerecorded by a female speech pathologist.

Two Italian speech corpora collected from adult speakers were also used in this work: the APASCI corpus and the IBN corpus.

The APASCI speech corpus (Angelini et al., 1994) is a task-independent, high quality, acoustic-phonetic Italian database. APASCI was developed at ITC-irst and consists of speech data collected from 176 adult speakers, gender balanced. Acquisitions were performed in quiet rooms using a digital audio

tape recorder and a high quality close talk microphone. Audio signals were down-sampled from 48 kHz to 16 kHz with 16 bit accuracy. Only a portion of APASCI corpus, consisting of speech from 124 speakers for a total of 5h:38m, was used in this work. This corpus was exploited for acoustic analysis of adult speech.

The IBN speech corpus is used for training the automatic broadcast news (BN) transcription system developed at ITC-irst for the Italian language (Bertoldi et al., 2001; Brugnara et al., 2002). It is composed of several speech data collections: speech from radio news programs, speech from television news programs and clean read speech from the APASCI and the SPEEDATA corpora. SPEEDATA (Ackermann et al., 1997) is a corpus designed and collected by ITC-irst with criteria very similar to those adopted for APASCI and consisting of about 5h:48m of speech. Recordings from radio and television news programs in the IBN corpus were manually segmented in sentences and only the utterances in clean conditions were included in the training portion of the corpus. The IBN corpus consists of 57h:07m of speech. In this work, the IBN corpus was used to train acoustic models for ASR experiments and for acoustic analysis.

Table 1 summarizes the characteristics of the speech corpora used in this work.

Corpus	ChildIt	SpontIt	CID subset	IBN	APASCI
Language	Italian	Italian	English	Italian	Italian
Speaking mode	Read	Spont.	Read	Read/Spont.	Read
Speaker age	7-13	8-12	5-17/Adult	Adult	Adult
# of speakers	171	21	40	> 1000	124
# of diff. words	11447	2141	27	31400	2191
Recording hours	10h:16m	1h:20m	0h:17m	57h:07m	5h:38m

Table 1
Main characteristics of the speech corpora used in this work.

3 Children's Speech Analysis

This section presents several acoustic analyses on children's speech. These analyses were carried out in order to achieve a better understanding of spectral and temporal changes occurring in speech produced by children of various ages. This section is organized in four parts, presenting analysis on phone duration, intra-speaker variability, characterization of the acoustic space and effects of vocal tract length variations, respectively.

3.1 Phone Duration

In previous studies, younger children have been reported to exhibit longer segment duration patterns compared to older children and adults (Lee et al., 1999; Gerosa et al., 2005). In this work we analyzed phone duration as a function of age on Italian and American English read speech. The mean phone duration was computed first averaging phone duration over all phones of each speaker and then across all speakers in each age group.

For Italian speech, 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. The first set was trained using the ChildIt training set and the SpontIt corpus, while the second set was trained using the IBN training set. Both HMM sets were state-tied context-dependent triphone HMMs with up to 8 Gaussian densities per state. Acoustic features were 13 mel frequency cepstral coefficients (MFCCs) (Davis and Mermelstein, 1980) plus their first and second order time derivatives.

Mispronunciations and time alignment errors may affect the phone duration analysis. However, in the ChildIt corpus mispronounced words were manually annotated and utterances including these words were not included in the speech data used in this work (see Table 1). Furthermore, time alignment errors do not affect significantly the duration analysis presented here, since it depends only on the accurate detection of boundaries toward silence, that is very reliable.

Figure 1 reports the mean phone duration for children, computed on the training portion of the ChildIt corpus (at least 14 speakers per age), and for adults, computed on the training portion of the IBN corpus.

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 13 years old children and adults can be partially explained by the fact that the IBN corpus is formed mostly of speech from professional radio and TV announcers that speak quite fast.

For English speech, duration statistics were computed by exploiting a phone-level manual segmentation obtained in the following way. First, automatic seg-

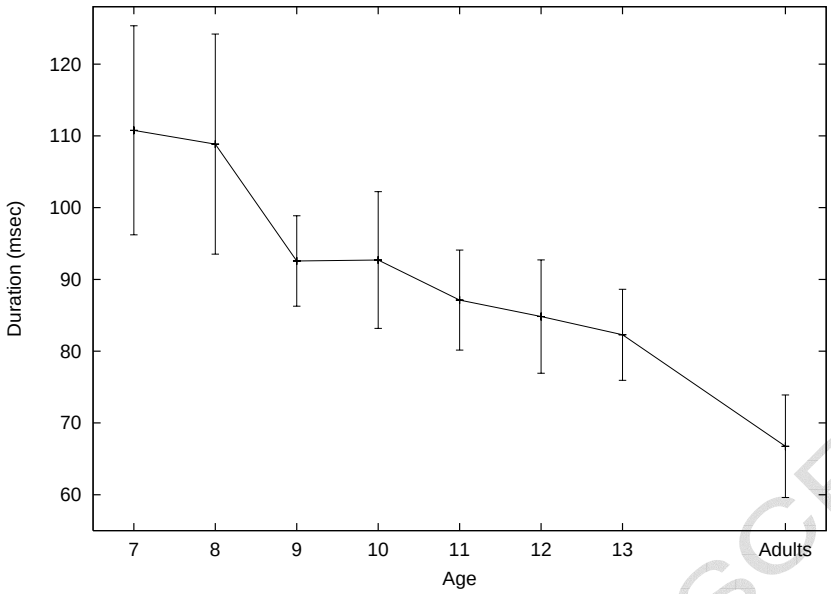


Fig. 1. *Mean duration of phones (msec) per age computed on the ChildIt training set. For comparison purpose, the mean phone duration for adults, computed on the IBN training set, is also reported. Vertical bars denote inter-speaker variability (standard deviation).*

mentation was obtained adopting the same procedure used for Italian speech. 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. Age-dependent HMMs, trained on a subset of the CID corpus, were used. After automatic segmentation, each utterance was analyzed by a native speaker of English with good phonetics knowledge. The annotator modified the boundaries of the automatic phonetic segmentation in order to correct segmentation errors.

Figure 2 reports the mean phone duration for children of different ages and adults, computed in the subset of CID corpus described in Section 2.

As for Italian speech, the mean phone duration varies with age and older age groups exhibit shorter mean phone duration. The significant difference between values obtained for children of age 5 and 7 can be explained by the fact that for children of age 5 speech was elicited through imitation of a sample recorded by an adult, while older children were able to read the set of sentences. Analysis of variance (ANOVA) (Clarke and Cooke, 1998) showed that variation of phone duration with respect to the age of the speakers is significant with $p < .001$ for both American English and Italian speech.

We have to point out that even if the ChildIt and the CID copora were designed with different purposes and concern different languages, variations with age of mean phone duration show a similar trend.

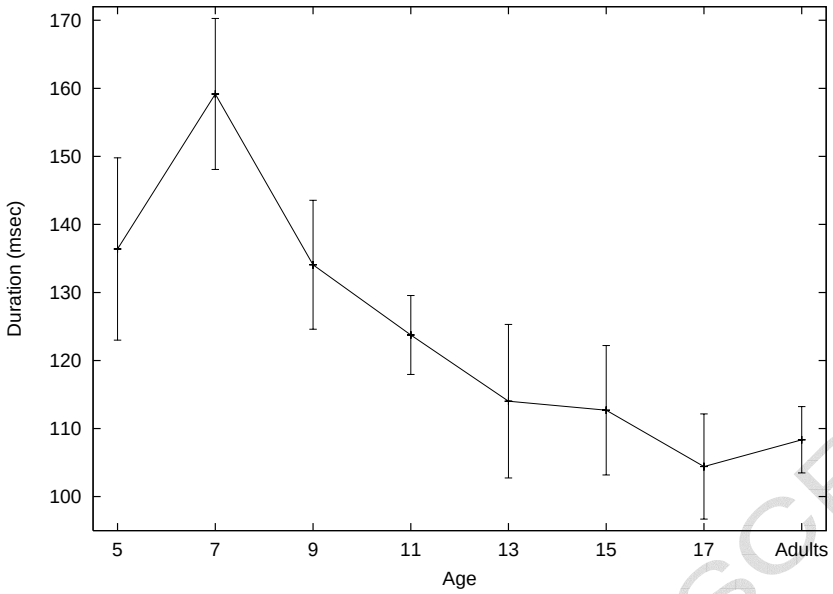


Fig. 2. *Mean phone duration (msec) per age computed on the CID subset. Vertical bars denote inter-speaker variability (standard deviation).*

Age-dependent variation in phone duration introduces variabilities that may affect ASR performance. In case of adults speakers, it is well known in fact that for speakers speaking much faster than the average of the training population low ASR performance is achieved (Mirghafori et al., 1996). The problem is sometimes tackled by training rate-specific acoustic models to describe speech of different rates (Mirghafori et al., 1996; Zheng et al., 2000).

3.2 Intra-speaker Variability

Intra-speaker variability is a measure of the maturity of speech motor control. In (Lee et al., 1999), by exploiting the CID corpus, intra-speaker variability was characterized as the temporal and spectral difference between corresponding vowels in two repetitions of the same sentence. We carried out similar analyses on both vowels and consonants exploiting the CID subset consisting of five phonetically rich sentences for which manual segmentation at the phone-level was available. We considered the two repetitions of the same sentence uttered by a given speaker and measured the spectral and temporal difference between the corresponding realizations of a given phone in the two utterances.

To perform the spectral analysis, the speech signal was first blocked into frames of 20 ms duration (with 50% frame overlapping), then each speech frame was parameterized into 12 MFCCs. Cepstral mean subtraction was performed on an utterance-by-utterance basis and each MFCC was scaled with the inverse of its standard deviation computed over all data. The mel cepstrum distance between two speech segments, each corresponding to a phone occurrence, was

computed by first computing the mean MFCC vector for each segment, and then taking the Euclidean distance between the two mean vectors as proposed in (Lee et al., 1999) and (Nakamura et al., 2005).

Figures 3 and 4 show the temporal and spectral difference averaged over all phones of a given speaker and then across all speakers in each age group.

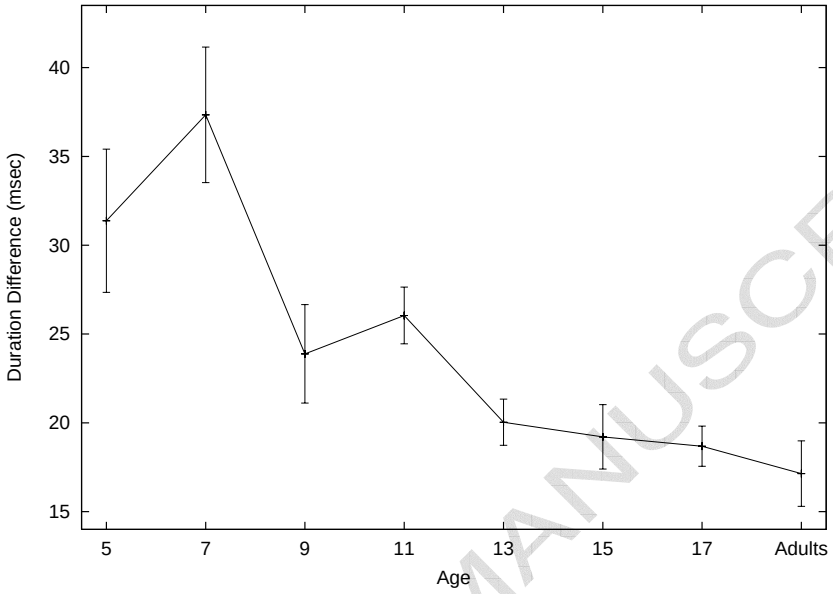


Fig. 3. *Mean duration difference, as a function of age, between corresponding phones in two repetitions of the same sentence by a given speaker in the CID subset. Vertical bars denote inter-speaker variability (standard deviation).*

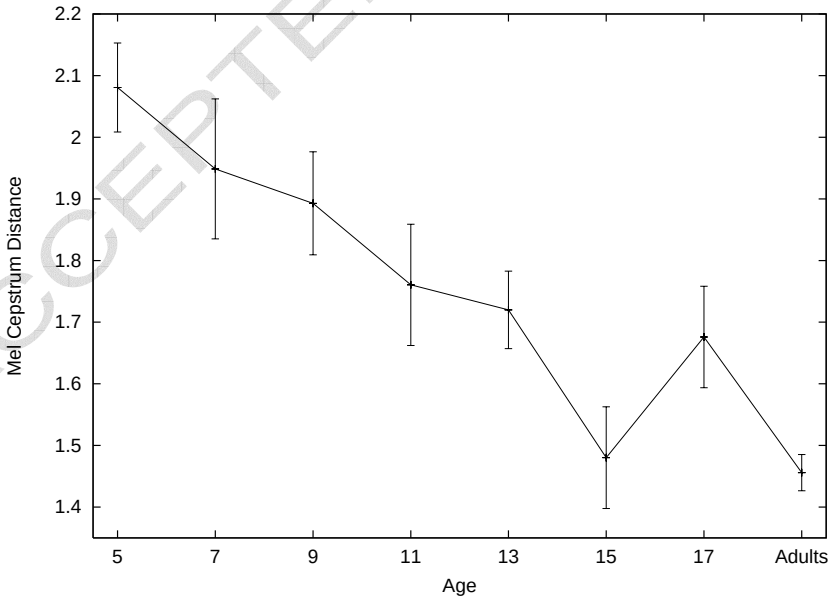


Fig. 4. *Mean mel cepstrum distance, as a function of age, between corresponding phones in two repetitions of the same sentence by a given speaker in the CID subset. Vertical bars denote inter-speaker variability (standard deviation).*

Observing Figures 3 and 4, it is clear that intra-speaker variability, both spectral and temporal, decreases as age increases. Analysis of variance showed that effect of age is significant in both cases with $p < .001$. Imitated speech produced by speakers of age 5 shows smaller temporal variability but higher spectral variability with respect to children of age 7. One possible explanation is that by repeating speech uttered by an adult, children are able to imitate his/her temporal pattern but their articulation control remains still uncertain.

It can be noted that the minimum for spectral variability is observed for children of age 15. This behavior was already observed when analyzing vowels on the same corpus (Lee et al., 1999), however the reason is not clear. This phenomenon could be associated with the learning process or it may be that the articulation control capability peaks during teenage years.

3.3 Characterization of the Acoustic Space

We tried to characterize the acoustic space by measuring the scattering of the observation densities of the phone models. For this purpose we modeled each phone by means of a single Gaussian density and we measured how much Gaussian densities were scattered in the acoustic feature space, when Gaussian parameters were estimated from speech examples collected by a pool of speakers. A statistical measure was used to determine how well phones were scattered in the acoustic space.

Given two phones i and j, modeled by Gaussian distributions, $\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ and $\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$, the distance between them can be measured by means of the Bhattacharyya distance (Fukunaga, 1990) as follows:

$$B(i,j) = \frac{1}{8}\left(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right)^T \left(\frac{\boldsymbol{\Sigma}_i + \boldsymbol{\Sigma}_j}{2}\right)^{-1} \left(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right) + \frac{1}{2}\log \frac{\left|\frac{\boldsymbol{\Sigma}_i + \boldsymbol{\Sigma}_j}{2}\right|}{\sqrt{|\boldsymbol{\Sigma}_i||\boldsymbol{\Sigma}_j|}} \qquad (1)$$

where $\boldsymbol{x}$ is a D-dimensional vector and $\boldsymbol{\mu}_i$, $\boldsymbol{\Sigma}_i$, $\boldsymbol{\mu}_j$ and $\boldsymbol{\Sigma}_j$ are the mean vectors and the covariance matrices of the Gaussian distributions of phones i and j, respectively. The Bhattacharyya distance has been used to measure phone separability and similarity in many works (Mak and Barnard, 1996; Salvi, 2003; Kumar et al., 2005).

Given a set of N Gaussian densities the average Bhattacharyya distance can be defined as follows:

$$AveB = \frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} B(i,j). \quad (2)$$

The average Bhattacharyya distance, $AveB$, can be considered a statistical measure of how scattered the N phones are in the acoustic space. High values of $AveB$ indicate that phone distributions are well scattered in the acoustic space and thus phones should be more easily discriminated, while low values of $AveB$ can be interpreted as an higher superposition of phone distributions and thus the phone discrimination task should be harder.

To estimate the parameters of Gaussian densities associated to phones, we trained a set of context-independent HMMs for each age group and gender. In all cases, a three-state left-to-right topology with a single Gaussian density per state was adopted. Each speech frame was parameterized into a 13-dimensional observation vector composed of 13 MFCCs plus their first and second order time derivatives. Frame energy was represented as the zero order (c_0) MFCC. Cepstral mean subtraction was performed on static features on an utterance-by-utterance basis. For children, HMM training was performed using the ChildIt training set, while, for adults, HMMs were trained using APASCI and SPEEDATA corpora. The ChildIt, APASCI and SPEEDATA corpora consist of read speech collected in controlled environments.

In computing the average Bhattacharyya distance, only Gaussian densities associated to the central states of context-independent HMMs were considered. We assumed that the Gaussian density associated to the central state of an HMM better reflects the acoustic characteristics of the modeled phone than Gaussian densities associated to the initial and final states.

In Figure 5 the average Bhattacharyya distance is reported per age groups and genders. In order to have a more robust estimation of model parameters three age groups were considered: children aged 7-9, 10-11 and 12-13. Only vowels were considered in computing the measures reported in the Figure. It can be noted that the average Bhattacharyya distance among vowel distributions increases with age for both genders showing that vowels distributions are less overlapped in the acoustic spaces of older age groups. This can be interpreted as the effect of a reduction in inter-speaker acoustic variability as age increases (Lee et al., 1999). Analysis of variance showed that the increase in average Bhattacharyya distance with respect to age is significant with $p < .001$, while the difference in the average Bhattacharyya distance due to the speaker gender was found not significant.

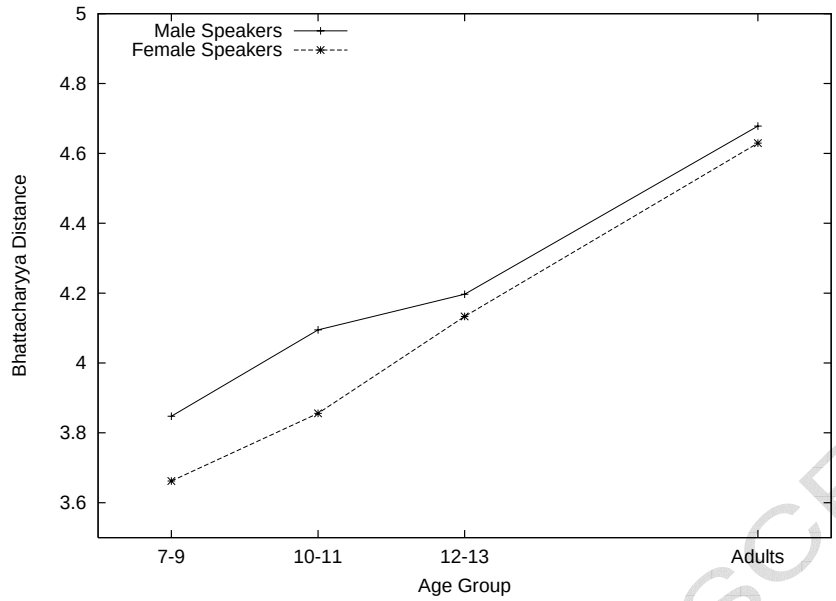


Fig. 5. *Average Bhattacharyya distance across vowel sounds per gender and age groups.*

3.4 Effects of Vocal Tract Length Variations

The correspondence between vocal tract morphology and speech acoustics predicted by the acoustic theory of speech (Wakita, 1977), has been studied in several works on vocal tract length (Fitch and Giedd, 1999) and formant patterns in speech (Huber et al., 1999; Lee et al., 1999). Anatomical measurements presented in (Fitch and Giedd, 1999) document the changes in vocal tract anatomy occurring during growth and maturity. Measurements 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 is reported in (Huber et al., 1999; Lee et al., 1999).

For children up to age 11 no significant difference in vocal tract length is observed between males and females of the same age, however formant frequencies of females tend to be higher than those of males of the same age. 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 enlargement of the glottis, which lowers the pitch. Adult males show a longer, about 10% on average, vocal tract than adult females. These anatomical measurements correlates well with the acoustic data and the formant pattern analysis presented in (Huber et al., 1999; Lee et al., 1999), where it is reported a steady gradual decreasing of formant frequencies with age for boys and girls up to the age of 11. After this age, for females a gradual decrease in formant frequencies is

still observed until age 15, when formant frequencies become similar to those of women. For males, beyond age 11, a substantial lowering in formant frequency is observed until the age 15, when formants frequencies become similar to those of men. This is largely explained by the disproportional growth of the vocal tract occurring during puberty (about age 11-15) in male subjects (Fitch and Giedd, 1999). After age 15, males show a substantial longer vocal tract and lower formant frequencies than females.

The above mentioned results from the literature are essentially confirmed by acoustic measurements we carried out during this work. We measured formant frequency values on the ChildIt corpus and on the APASCI corpus, exploiting the phone level automatic segmentation obtained as described in Section 3.1. Segments corresponding to vowel sounds were extracted and their mean formant frequencies were estimated using the Praat software tool (Boersma and Weenink, 2001). In Figure 6 the mean frequencies, for each age and gender, of the fundamental frequency (F0) and of the first, second and third formants (F1, F2, F3) are reported. Frequency values were first averaged across all vowels of a given speaker and then across all speakers in each age group. It can be noted that the mean frequency values decrease with age, as expected, and that F1, F2 and F3 values are higher for female speakers than for male speakers.

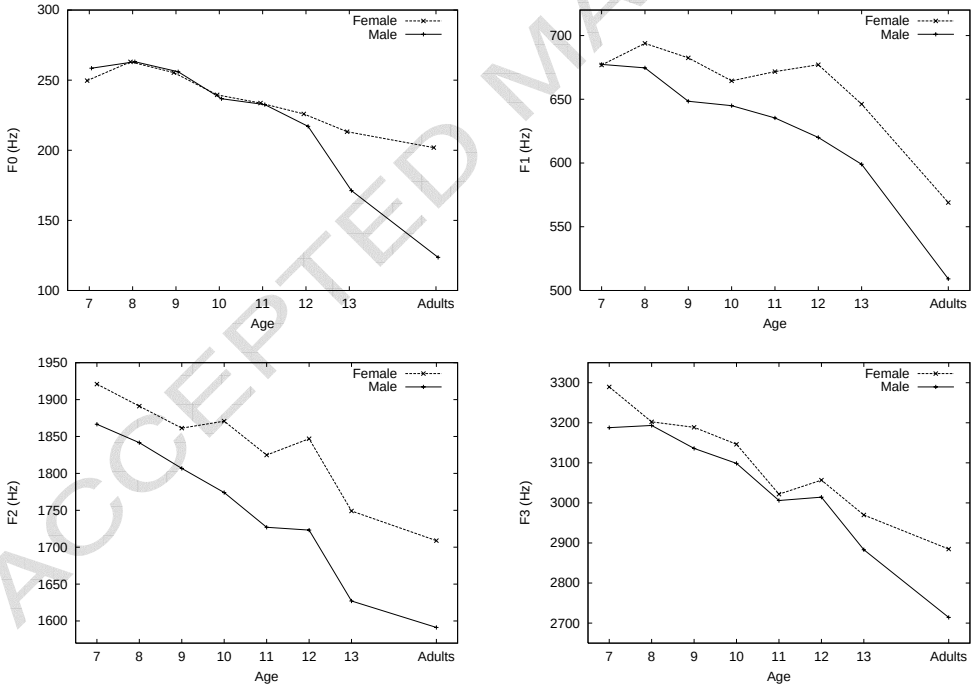


Fig. 6. *Mean frequency values, per age and gender, of the fundamental frequency (F0) and the first three formants (F1, F2, F3).*

Age-dependent variation in formant frequencies introduces variability in the spectral features across age groups, this concurs to explain the degradation in performance that can be observed when an ASR system trained on speakers

of a certain age group is tested on speakers of a different age group (Wilpon and Jacobsen, 1996; Hagen et al., 2003). As will be described in Section 4, a common practice in automatic speech recognition is to try to compensate for spectral differences caused by differences in vocal tract length (and shape) by warping the frequency axis of the speech power spectrum of each speaker (Lee and Rose, 1996).

To measure the spectral mismatch between adult and children's speech caused by differences in vocal tract length, we computed the mean warping factor for speakers in the ChildIt corpus with respect to two HMM sets trained on adult speech. These two HMM sets were trained on the IBN training set, using speech from male speakers only ("Adult male HMMs") and from female speakers only ("Adult female HMMs"), respectively. For both males and females triphone HMMs, with a single Gaussian density per state, were trained. Warping factors, that are scaling factors to be applied to the frequency axis of the speech power spectrum, were determined according to the procedure summarized in Section 4.1. A grid search over 28 possible warping factors, evenly distributed (with step 0.02) in the range 0.66-1.20, was performed for each speaker in order to maximize the likelihood of the speaker's data with respect to a reference model set.

In Figure 7 the mean warping factor, averaged across speakers in each age, is reported for children in the ChildIt corpus (at least 10 speakers for each age and gender) and, for comparison purpose, for adult speakers in the test portion of the IBN corpus.

A mean warping factor of 1.0 denotes that there is, on average, no spectral mismatch between the training and testing populations. An almost linear variation of the warping factor with respect to children's age can be observed. On average, significantly lower warping factors are reported for younger children than for the older ones. Analysis of variance showed that this variation is significant in all cases reported ($p < .001$). Furthermore, girls show a constant lower warping factor with respect to boys of the same age. It can be noted that for each age and gender the mean warping factor estimated with respect to reference HMMs trained on adult female voices is closer to 1.0 than the corresponding mean warping factor estimated with respect to reference HMMs trained on adult male voices. This confirms that voices of children tend to be more similar to the voices of women than to those of men. In general, mean warping factors per age and gender reported in Figure 7 are compatible with results on formant pattern analysis reported in the literature for this range of ages (Huber et al., 1999; Lee et al., 1999) and with measurements of formant frequencies we carried out on the ChildIt corpus, reported in Figure 6.

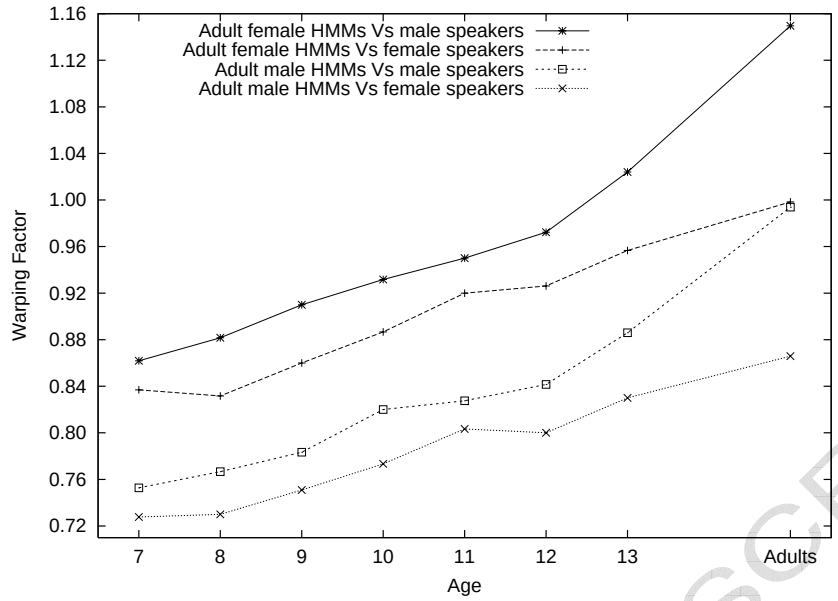


Fig. 7. *Mean warping factor, per age and gender, estimated with respect to two reference adult HMM sets, the first one trained on adult males ("Adult male HMMs") and the second one trained on adult females ("Adult female HMMs").*

4 Speaker Normalization

As we have seen in Section 3, children are not a homogeneous group of speakers due to the changes occurring with age, and even when considering a particular age group children's speech is characterized by a higher acoustic variability than adult speech. Therefore, it is expected that speaker adaptive acoustic modeling methods have a high potential of application in the context of children's speech recognition. In fact, 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, speaker adaptive acoustic modeling was investigated through VTLN, SAT and CMLSN methods.

4.1 VTLN

VTLN 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 the frequency warping approach to speaker normalization, 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). 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 maintaining the spectrum unchanged (Lee and Rose, 1996).

Let us consider a set Λ of speaker independent (SI) HMMs trained with speech data from a pool of training speakers. The optimal scaling factor $\hat{\alpha}$ for an utterance $\boldsymbol{x} = \boldsymbol{x}_1, ..., \boldsymbol{x}_T$, with $\boldsymbol{x}_t$ denoting the acoustic observation at time t, can be determined according to the maximum likelihood criterion as follows (Lee and Rose, 1996):

$$\hat{\alpha} = \underset{\alpha}{\operatorname{argmax}} P(\boldsymbol{x}^{\alpha}|\boldsymbol{w}, \Lambda) \tag{3}$$

where $\boldsymbol{x}^{\alpha} = \boldsymbol{x}_1^{\alpha}, ..., \boldsymbol{x}_T^{\alpha}$ denotes the acoustic observation sequence obtained by applying the scaling factor α to the frequency axis of the speech power spectrum and $\boldsymbol{w}$ denotes the sequence of uttered words. The optimal scaling factor $\hat{\alpha}$ can be determined for each utterance or for a group of utterances uttered by a given speaker. Ideally, the effect of adopting a scaling factor selected in this way for each utterance or speaker is that of normalizing speech data with respect to the average vocal tract length of the training population of the model set Λ, thus reducing inter-speaker acoustic variability.

In this work, frequency warping was always implemented by changing the spacing and width of the filters in the mel filter-bank while maintaining the speech spectrum unchanged. To cope with the problem of accommodating filters near the band edge, a piece-wise linear warping function of the frequency axis of the mel filter-bank was adopted. During training the reference acoustic models for scaling factor selection, carried out on a speaker-by-speaker basis, were SI triphone HMMs with 1 Gaussian per state and trained on unwarped data. Differently, during testing scaling factor selection was performed with respect to the HMMs, trained on warped data, used for the final decoding step. Furthermore word level transcription of incoming utterances were generated by a preliminary decoding step carried out with models trained on unnormalized data. During both training and testing 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 follow closely those proposed in (Welling et al., 1999) and are described in detail in (Giuliani et al., 2006).

4.2 SAT

Speaker adaptive training aims at compensating for inter-speaker acoustic variability present in the training set by means of speaker-specific transformations. In the original formulation (Anastasakos et al., 1996) it involved maximum likelihood linear regression (MLLR) adaptation of the means of output distributions of continuous density HMMs. The resulting HMMs exhibit usually smaller variances and lead to significantly higher likelihood. The use of these models in combination with speaker adaptation techniques can result in an improvement of recognition performance.

The variant of the SAT scheme developed by Gales (Gales, 1998) was used in this work. This variant makes use of an affine transformation, estimated through constrained MLLR, for mapping acoustic observations of each training and testing speaker, instead of adapting model parameters. Transformation parameters are estimated with the aim of reducing the acoustic mismatch between speaker data and the reference models. With this method, a set of SI 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 means 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. 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.

4.3 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, differently from 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's data with respect to a set of target HMMs which is different from the HMM set to be used for recognition. Target models are, in fact, triphone HMMs, having a single Gaussian density per state with diagonal covariance matrix, trained on unnormalized data. These models are used as target models during both training and recognition.

For each speaker, estimation of transformation parameters is carried out within

the Expectation-Maximization (EM) framework, which requires the maximization of the following auxiliary function in order to increase the likelihood of the transformed data:

$$\mathcal{Q} = -\frac{1}{2}\sum_{\boldsymbol{x}\in X}\sum_{g=1}^{G}\sum_{t=1}^{T(\boldsymbol{x})}\gamma_t(g)\Big(-\log(|\boldsymbol{A}|^2)+ \\ +(\boldsymbol{A}\boldsymbol{x}_t+\boldsymbol{b}-\boldsymbol{\mu}_g)^*\boldsymbol{\Sigma}_g^{-1}(\boldsymbol{A}\boldsymbol{x}_t+\boldsymbol{b}-\boldsymbol{\mu}_g)\Big) \quad (4)$$

where X denotes the set of available utterances, g denotes a Gaussian density of the target HMMs, t denotes a time frame and $\boldsymbol{A}$ and $\boldsymbol{b}$ represent the matrix and the offset vector of the constrained transformation. $\gamma_t(g)$ represents the conditional probability of Gaussian density g at time t given the observation sequence and the current set of model parameters. A re-estimation procedure for $\boldsymbol{A}$ and $\boldsymbol{b}$ was proposed in (Gales, 1998) under the assumption of diagonal covariance matrices. Maximization of Eq. (4) is carried out with respect to the triphone HMM concatenation corresponding to word level transcriptions of the available utterances.

Once training data have been transformed, acoustic models to be used for recognition, that is HMMs having output distributions modeled with Gaussian mixtures, are trained from scratch exploiting normalized data. During recognition, for each test speaker, word level transcription of incoming utterances, needed for the estimation of the transformation parameters, are generated by a preliminary decoding step carried out with models trained on unnormalized data. Then transformation parameters are estimated with respect to the same target models used during training and finally speaker's data are transformed and decoded with the models trained on normalized data. Details about training and testing procedures can be found in (Giuliani et al., 2006).

4.4 Combination of Techniques

With SAT and CMLSN no assumption is made about the nature of the acoustic mismatch between the speaker's data and the target HMMs. On the contrary the VTLN method is tailored to reduce spectral differences induced by variations in vocal tract by warping the frequency axis of the power spectrum. Therefore, it can be expected that the VTLN method can be applied in combination with either of the CMLSN or SAT procedure with good results (Giuliani et al., 2006). In fact, after VTLN is performed, the two methods can be applied with the aim of reducing the residual acoustic mismatch induced by sources of speaker individualities different from vocal tract variations (e.g. accent or dialect) or/and by acquisition conditions (e.g. environment and type of microphone). During recognition, as the most computationally expensive

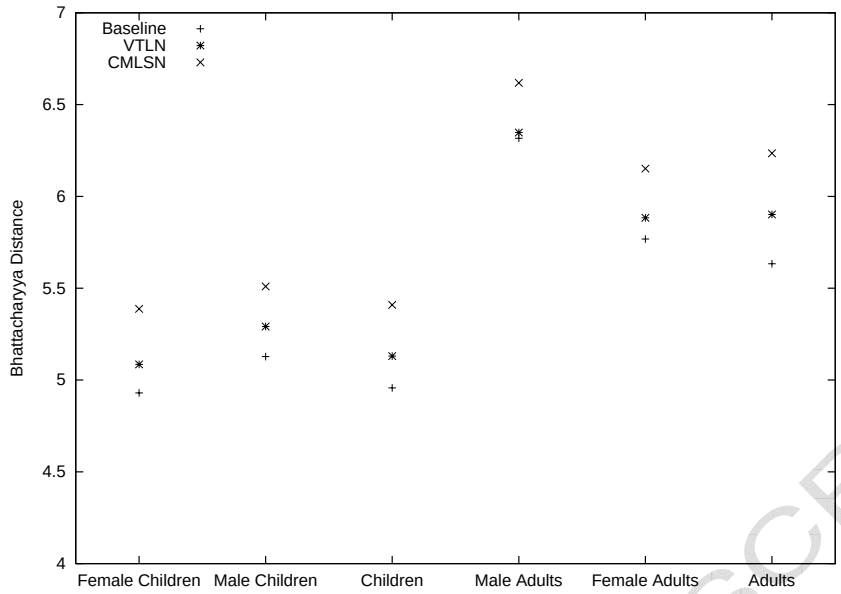


Fig. 8. *Average Bhattacharyya distance across all phones per gender and age groups, computed on the baseline models and on models trained using the VTLN and CMLSN procedures.*

stage is the generation of the transcription of test data, the computational cost of these combinations of methods is not much higher than the cost of the application of only one of them.

4.5 Effect of the VTLN and CMLSN Methods on the Acoustic Space

We tried to characterize the effect of the VTLN and CMLSN methods on the acoustic space by measuring the scattering of Gaussian densities modeling phones in the normalized feature space using the average Bhattacharyya distance as described in Section 3.3. We trained context-independent HMMs with one Gaussian density per state on the original data and on data normalized through the VTLN and CMLSN methods. Different sets of models were trained for adults and children, and for males and females of the two population groups. HMMs training was performed exploiting the training sets available for children and adults described in Section 2.

Figure 8 reports the average Bhattacharyya distance per age group and gender computed on baseline HMMs and on HMMs trained by adopting the VTLN and CMLSN training procedures. As expected, the average Bhattacharyya distance computed on HMM sets trained on adult speech is higher than the distance computed on HMMs trained using children's speech. Furthermore, it is clear that with both normalization techniques the average Batthacharyya distance increases, supporting the hypothesis that in the normalized acoustic

space phones are better scattered and thus phone discrimination should be easier. In all cases, the average Batthacharyya distance increases more for models trained using the CMLSN procedure than for models trained with the procedure based on VTLN.

5 Recognition Experiments

Several recognition experiments were carried out to investigate the impact of the increased variability in children's speech on ASR performance.

In all the experiments carried out acoustic models were state-tied, cross-word triphone HMMs. In particular, a Phonetic Decision Tree (PDT) was used for tying the states of triphone HMMs (Young et al., 1994). Output distributions associated with HMM states were modeled with mixtures with up to 8 diagonal covariance Gaussian densities.

A set of 48 phonetic units, 7 vowels and 41 consonants, corresponding to the Italian phonemes, were modeled. "Silence" was modeled with a single state HMM. In addition, a number of models for common extra linguistic phenomena, such as human noises (e.g. breathing and lip smacks), non-verbal sounds and filled pauses, were trained.

Each speech frame was parameterized into a 39-dimensional observation vector composed of 13 MFCCs plus their first and second order time derivatives. Cepstral mean subtraction (CMS) was performed on static features on an utterance-by-utterance basis.

For "batch" recognition experiments, where data of each speaker were assumed available in a single block, a different acoustic feature normalization was adopted. Mean and variance normalization was performed on all 39 acoustic features, on a speaker-by-speaker basis, forcing each acoustic feature to have zero mean and unit variance. Preliminary experiments showed that this kind of normalization ensures systematic benefits with respect to the CMS adopted in the standard acoustic front-end, especially in case of unmatched training and testing conditions.

The language model used was an 11k word trigram language model estimated on a corpus of newspaper articles. The recognition vocabulary was composed of the words occurring in the training and test set of the ChildIt corpus. The perplexity of the 11k word LM on the ChildIt test set was 900. This high perplexity is explained by the fact that the n-gram statistics estimated on the training text corpus, composed of newspaper articles, did not reflect well the word distribution in the ChildIt corpus, which was made of texts extracted

from literature for children. In addition, most of the sentences in the ChildIt corpus are short (with an average sentence length of 7 words and a minimum sentence length of 4 words). In perplexity computation, this results in a high number of unseen trigrams and bigrams at the sentence boundaries and in the use of many back off probabilities.

5.1 Baseline for Children's Speech Recognition

We trained a set of SI cross-word triphone HMMs using the ChildIt training set and the SpontIt corpus (about 9 hours of speech). The total number of independent states was 1700 for a total of about 13200 Gaussian densities. We evaluated recognition performance on the ChildIt test set and used this result as a reference for all the experiments reported below.

A second set of SI cross-word triphone HMMs was trained adopting the VTLN training procedure described in Section 4.1, with warping factor selection performed on a speaker-by-speaker basis. During the decoding stage the warping factor selection was instead carried out on an utterance-by-utterance basis. Table 2 reports the word error rates (WERs) achieved on the ChildIt test set by using baseline models and models trained on warped data.

HMM set	Baseline	VTLN
Child HMMs	14.4	12.7

Table 2
Recognition results (% WER) obtained on the ChildIt test set using models trained on children's speech with and without adopting VTLN.

Adopting VTLN in training and test reduces the WER with respect to the baseline system by 1.7%, from 14.4% to 12.7%, that corresponds to a relative reduction in WER of 11.8%.

Figure 9 shows the WER as a function of age in the ChildIt test set, achieved using the acoustic models trained on children speech, with and without applying VTLN. While the mean WER reported in Table 2 is 14.4%, there is a large difference in performances achieved for children in different ages. It can be noted that recognition results achieved on children with ages in the middle of the age range considered are better than those achieved on younger and older children. This can be explained by the fact that acoustic characteristics of voices of children in the middle age groups are better represented in the training set. Applying the VTLN method gives recognition results consistently better for all ages. However, the performance trend is still similar to that achieved using the baseline models.

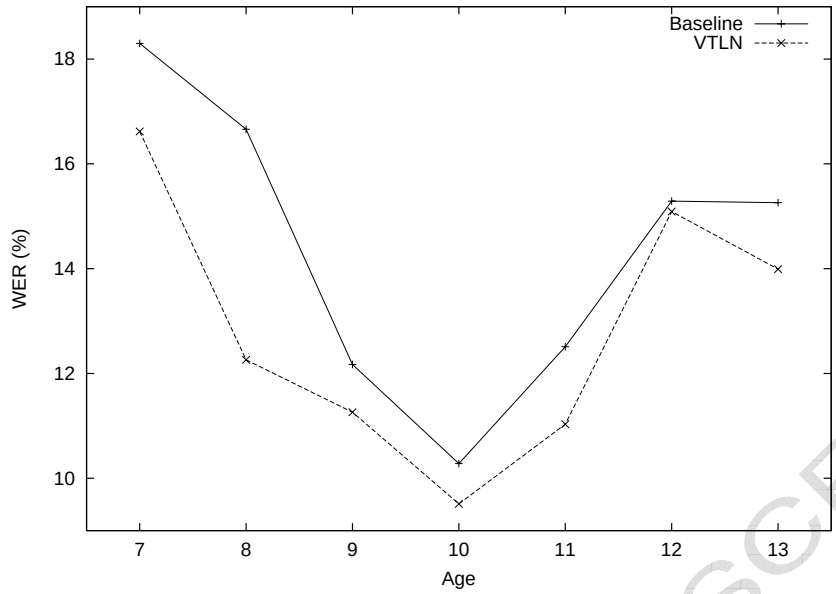


Fig. 9. *Recognition results (% WER) on the ChildIt test set as a function of age by using HMMs trained on children's speech with and without applying VTLN.*

5.2 Experiments in Unmatched Conditions

A set of recognition experiments was carried out with the aim of measuring how much recognition performance decreases, within our experimental framework, under unmatched training and testing conditions. By using adult speech from the IBN corpus, three different HMM sets were trained: the first one using speech from adult male speakers in the IBN training set, the second one using speech from adult female speakers in the IBN training set and, finally, the third one using the whole IBN training set. This resulted in three cross-word triphone HMM sets having 40000, 29560 and 53860 Gaussian densities, respectively.

Table 3 reports the WERs achieved on the ChildIt test set by using model trained on adult speech ("Adult HMMs"), on adult male speech only ("Adult male HMMs") and adult female speech only ("Adult female HMMs"). Recognition results, in column "Baseline", show that WER achieved by using HMMs trained on adult male speech is more than twice higher than WER achieved by using HMMs trained on adult female speech (72.1% vs 31.2%). This is consistent with results of analysis reported in Section 3, where it is shown that spectral characteristics of children's voices are more similar to those of female voices than those of male voices.

Figure 10 shows the WER as a function of age on the ChildIt test set, achieved using acoustic models trained on adult speech. As we can see, while the mean WER reported in Table 3 is 41%, there is a large difference in performance

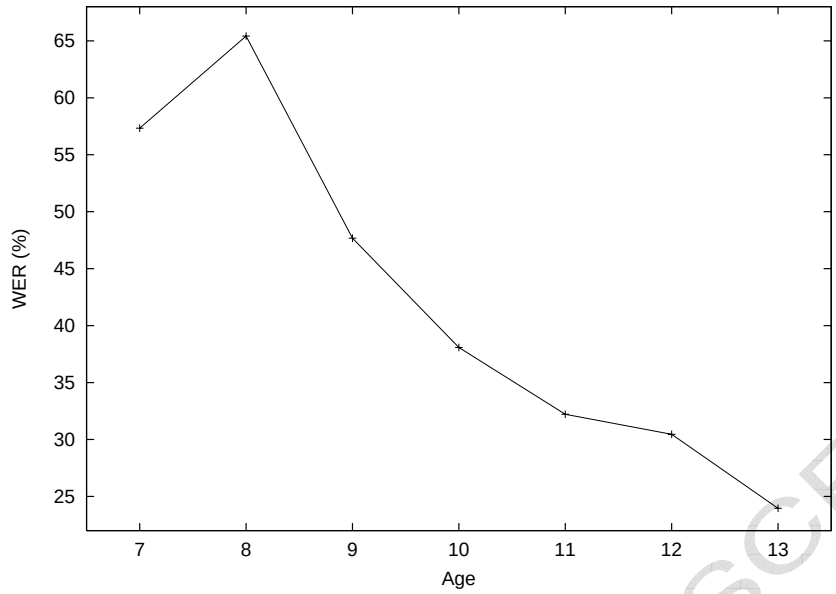


Fig. 10. *Recognition results (% WER) on the ChildIt test set as a function of age by using HMMs trained on adult speech.*

achieved for younger children and older children. WER for children aged 7-9 years is more than 95% higher than that achieved for children aged 13 years. This preformance trend is in agreement with the effect of developmental changes on acoustic parameters discussed in Section 3.4.

A simple approach to compensate the mismatch between children's voices and acoustic models trained on adult speech is to introduce a fixed scaling, constant over all speakers, in the frequency axis of the speech spectrum (Das et al., 1998; Claes et al., 1998; Giuliani and Gerosa, 2003). Figure 11 reports WERs achieved on the ChildIt training set as function of the warping factor adopted. We considered 13 warping factors, evenly distributed (with step 0.02) in the range 0.76-1.00, where a warping factor of 1.00 corresponds to the case of no warping applied. As in the VTLN experiments, speaker-independent frequency warping was implemented by changing the spacing and width of the filters in the mel filter-bank while maintaining the speech spectrum unchanged.

Looking at results reported in Figure 11, we can see that a warping factor of 0.86 is the best suited to compensate for the existing mismatch between children's speech and HMMs trained on adult voices. The best warping factor for models trained on adult female speakers is 0.92, while the one for models trained on adult male speakers is 0.80. Recognition results obtained on the ChildIt test set with the estimated optimal speaker-independent warping factors are reported in Table 3 in the column "SI Warping".

Recognition performance improvement is significant in all cases even if still far from that achieved under matched conditions (training and testing on

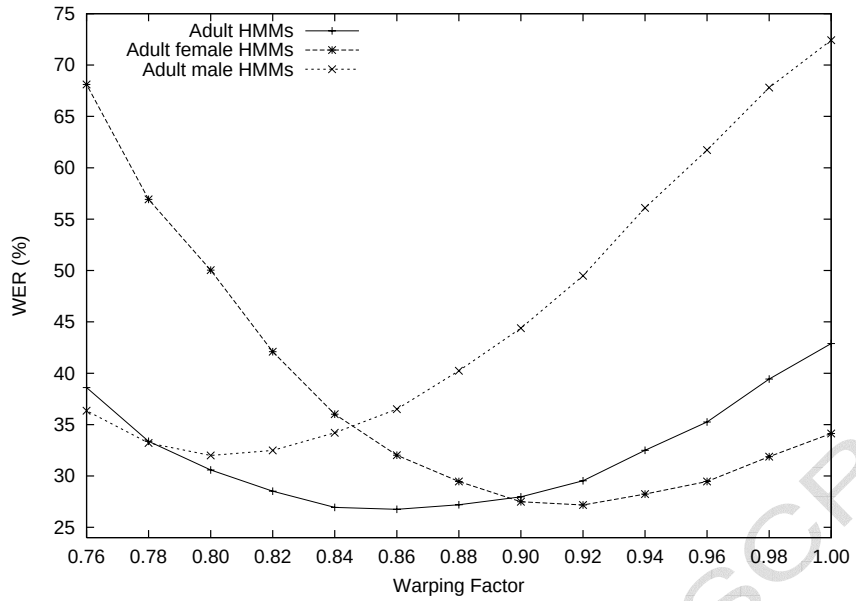


Fig. 11. *Recognition results (% WER) on the ChildIt training set as function of the warping factor by using HMMs trained on adult males and females ("Adult HMMs"), adult males ("Adult male HMMs") and adult females ("Adult female HMMs").*

HMM set	Baseline	SI Warping	UD Warping	SI Adaptation
Adult HMMs	41.0	23.2	22.4	14.8
Adult male HMMs	72.1	28.6	27.3	14.8
Adult female HMMs	31.2	23.5	21.2	14.4

Table 3
Recognition results (% WER) obtained on the ChildIt test set using models trained on adult speech, performing SI warping of children's speech spectrum ("SI Warping"), adopting an utterance-dependent warping factor ("UD Warping") and adapting to children's speech the HMMs trained on adult speech ("SI Adaptation").

children). Performance achieved by models trained on adults of both genders and by models trained on adult females are very similar and still much better than that achieved by models trained on adult males.

Instead of applying a speaker-independent frequency scaling factor, the warping factor can be selected on an utterance-by-utterance basis adopting the VTLN testing procedure. Baseline acoustic models, trained on unwarped adult data, were exploited for the preliminary decoding step, warping factor selection and final recognition step. Recognition results obtained using the utterance-dependent warping factors are reported in Table 3 in the column "UD Warping". It can be noted that using an utterance-dependent warping factor gives performance only slightly better than using a speaker-independent frequency

warping factor.

Finally, we tried to improve recognition performance by using the data in the training set for children (about 9h of speech) to perform acoustic model adaptation on the three HMM sets trained with adult data. Means and variances of Gaussian densities were adapted through 4 iterations of MLLR exploiting a regression class tree for dynamic allocation of regression classes according to the amount of adaptation data (Leggetter and Woodland, 1995). The minimum occupancy threshold was fixed to 1000, roughly 10 seconds of speech. For each regression class a full transformation matrix was estimated for Gaussian mean adaptation while a diagonal transformation matrix was used for variance adaptation. Recognition results are reported in Table 3 in column "SI Adaptation". As expected performance improves substantially for all the three set of models. Adapting HMMs trained on adult females results in best recognition performance with a 14.4% WER which is the same results achieved with HMMs trained on the same children's data (see Table 2). We have to point out that MLLR adaptation does not adapt state transition probabilities and mixture weights.

5.3 Speaker Adaptive Acoustic Modeling

This section reports results of recognition experiments that were carried out in batch mode, assuming all the data of each test speaker available in block for multiple processing. For this purpose, the manual annotation of the speaker identity, in addition to the manual segmentation into separate utterances, was exploited. The decoder was run twice, and the output of the first decoding step was exploited as a supervision for performing system adaptation before the second decoding step took place. Unsupervised speaker adaptation was performed by adapting means and variances of Gaussian densities through MLLR. 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. In the experiments reported below the word transcriptions generated with the first decoding step were exploited as a supervision also for performing speaker normalization.

For both adults and children, we trained three normalized HMM sets using the VTLN, CMLSN and SAT training procedures summarized in Section 4, with the aim of reducing the effect of inter-speaker acoustic variability and improving recognition performance.

Table 4 reports recognition results obtained on the ChildIt test set by adopt-

ing the three above mentioned speaker adaptive acoustic modeling methods and two combinations of methods, "VTLN+SAT" and "VTLN+CMLSN". Note that both in case of baseline models and of models trained with speaker adaptive methods, unsupervised static MLLR speaker adaptation of acoustic models was performed before the second decoding step.

HMM set	Baseline	VTLN	CMLSN	SAT	VTLN + CMLSN	VTLN + SAT
Adult HMMs	20.0	15.4	15.2	16.0	14.4	14.8
Child HMMs	11.6	11.2	10.6	11.0	10.6	10.5

Table 4
Recognition results (% WER) obtained on the ChildIt test set using HMMs trained on adult speech and children's speech with several speaker adaptive acoustic modeling methods.

It can be noted that the CMLSN method outperforms the VTLN and SAT methods in both matched and unmatched conditions. This improvement in performance was validated using the matched-pair sentence test (Gillick and Cox, 1989) to ascertain whether the observed results were inconsistent with the null hypothesis that the output of two systems were statistically identical. Considered significance levels were .01, .005 and .001. In matched condition (see row "Child HMMs") the improvements achieved with the CMLSN method with respect to the VTLN and SAT methods are statistically significant with $p < .005$ and $p < .01$, respectively. In unmatched conditions (see row "Adult HMMs") the CMLSN method ensures better result than the SAT method, with $p < .001$, while the difference in performance between the CMLSN and the VTLN method is not statistically significant.

By using the CMLSN method in matched conditions a 8.6% WER relative reduction was achieved with respect to the baseline system, from 11.6% to 10.6% WER, while in unmatched conditions a 23.0% WER relative reduction was achieved, from 20.0% to 15.2%. We have to point out that the recognition results achieved using models trained on adult data using VTLN, CMLSN and SAT (15.4%, 15.2% and 16.0% WER, respectively) are still worse than those achieved by the baseline models trained on children's speech (11.6% WER). When cascading the VTLN method with the CMLSN and SAT methods, results achieved are always equal or better than those achieved using one of the normalization method alone. It is to point out that while the difference in WER achieved with baseline models for adults and children is around 72.0% relative (20.0% vs 11.6%), using the VTLN+CMLSN method the difference is reduced to 36.0% relative (14.4% vs 10.6%).

6 Conclusions

In this paper, results of several acoustic analyses on children's and adult read speech were presented. These analyses focused on phone duration, effects of vocal tract length variation and intra-speaker spectral and temporal variability.

Speech analyses were carried out on read Italian speech collected from children 7-13 years old and on read American English speech collected from children 5-17 years old. Measurements of fundamental and formant frequencies, carried out on Italian speech, and of phone duration, carried out on both Italian and American English speech, showed that spectral and temporal characteristics of children's speech vary with age, as it was expected from results reported in the literature for American English speech. In particular for the age range considered we observed a significant decrease of pitch and formant frequencies as a child's age increases. A similar trend was observed for phone duration. Also intra-speaker temporal and spectral variability, measured on American English speech, decreases as age increases.

Results of the analyses carried out confirmed that children are not a homogeneous group of speakers. In addition, measures of intra- and inter-speaker spectral variability showed that spectral variability of children's speech is higher than that of adult speech even when a specific age group is considered. This raises challenging issues in the development of highly effective acoustic models for ASR applications.

In this work, to cope with variations in spectral parameters induced by developmental changes, speaker adaptive acoustic modeling was investigated through the use of the VTLN, CMLSN and SAT methods. These methods proved to be effective when used to train acoustic models on adult and children's speech. In particular, the CMLSN method always outperformed the other methods used in this work.

The best recognition results were achieved by cascading the VTLN method with the CMLSN method. Using this combination of methods, we obtained a relative WER reduction, with respect to the baseline systems, of 8.6% and 28.0%, respectively, for matched and unmatched conditions. However, even with the adoption of speaker adaptive acoustic modeling techniques, recognition results achieved using HMMs trained on children's speech were significantly better than those achieved using HMMs trained on adult speech.

The improvement in recognition performance achieved with the adopted speaker adaptive acoustic modeling methods opens new prospectives for developing of general acoustic models able to perform well on speakers of all ages, thereby reducing the need for age-specific acoustic models.

Acknowledgements

This work was partially financed by the Autonomous Province of Trento (Italy) under the project PEACH (Fondo Unico Program).

References

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

Anastasakos, T., McDonough, J., Schwartz, R., Makhoul, J., Oct. 1996. A Compact Model for Speaker-Adaptive Training. In: Proc. of ICSLP. Philadelphia, PA, pp. 1137–1140.

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

Arunachalam, S., Gould, D., Andersen, E., Byrd, D., Narayanan, S., Sept. 2001. Politeness and Frustration Language in Child-Machine Interactions. In: Proc. of EUROSPEECH. Aalborg, Denmark, pp. 2675–2679.

Banerjee, S., Beck, J. E., Mostow, J., Sept. 2003. Evaluating the Effect of Predicting Oral Reading Miscues. In: Proc. of EUROSPEECH. Geneva, Switzerland.

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

Boersma, P., Weenink, D., 2001. Praat, a system for doing phonetics by computer. Glot International 5 (9/10), 341–345.

Brugnara, F., Cettolo, M., Federico, M., Giuliani, D., Sept. 2002. Issues in automatic transcription of historical audio data. In: Proc. of ICSLP. Denver, CO, pp. 1441–1444.

Burnett, D. C., Fanty, M., 1996. Rapid Unsupervised Adaptation to Children's Speech on a Connected-Digit Task. In: Proc. of ICSLP. Vol. 2. Philadelphia, PA, pp. 1145–1148.

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

Clarke, G. M., Cooke, D., 1998. A Basic Course in Statistics. Arnold, chapter 22, pages 520-546.

Das, S., Nix, D., Picheny, M., May 1998. Improvements in Children's Speech Recognition Performance. In: Proc. of ICASSP. Seattle,WA.

Davis, S., Mermelstein, P., 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. on Acoustics, Speech and Signal Processing 28, 357–366.

Eide, E., Gish, H., May 1996. A Parametric Approach to Vocal Tract Lenght Normalization. In: Proc. of ICASSP. Atlanta, GA, pp. 346–349.

Eskenazi, M., Pelton, G., Sept. 2002. Pinpointing pronunciation errors in children's speech: examining the role of the speech recognizer. In: PMLA. Aspen Lodge, CO, pp. 48–52.

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

Fukunaga, K., 1990. Introduction to Statistical Pattern Recognition, 2nd Edition. Academic Press, New York.

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., Brugnara, F., Sep. 2005. Speaker Adaptive Acoustic Modeling with Mixture of Adult and Children's Speech. In: Proc. of INTERSPEECH/EUROSPEECH. Lisboa, Portugal, pp. 2193–2196.

Gillick, L., Cox, S. J., May 1989. Some Statistical Issues in the Comparison of Speech Recognition Algorithms. In: Proc. of ICASSP. Glasgow, pp. I–532–535.

Giuliani, D., Gerosa, M., Apr. 2003. Investigating Recognition of Children Speech. In: Proc. of ICASSP. Vol. 2. Hong Kong, pp. 137–140.

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

Gustafson, J., Sjölander, K., Oct. 2000. Voice transformations for improving children's speech recognition in a publicly available dialogue system. In: Proc. of ICSLP. Beijing, China, pp. 297–300.

Hagen, A., Pellom, B., Cole, R., Dec. 2003. Children's Speech Recognition with Application to Interactive Books and Tutors. In: Proc. of the ASRU Workshop. St. Thomas Irsee, US Virgin Islands.

Hagen, A., Pellom, B., Vuuren, S. V., Cole, R., May 2004. Advances in Children's Speech Recognition within an Interactive Literacy Tutor. In: Proc. of HLT/NAACL. Boston, MA.

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

Kumar, S. C., Mohandas, V. P., Li, H., Sept. 2005. Multilingual Speech Recognition: A Unified Approach. In: Proc. of INTERSPEECH/EUROSPEECH. Lisboa, Portugal, pp. 3357–3360.

Lee, L., Rose, R. C., May 1996. Speaker Normalization Using Efficient Frequency Warping Procedure. In: Proc. of ICASSP. Atlanta, GA, pp. 353–356.

Lee, S., Potamianos, A., Narayanan, S., March 1999. Acoustic of children's speech: Developmental changes of temporal and spectral parameters. Jour-

nal of Acoust. Soc. Amer. 105 (3), 1455–1468.

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

Mak, B., Barnard, E., Oct. 1996. Phone Clustering Using the Bhattacharyya Distance. In: Proc. of ICSLP. Philadelphia, PA, pp. 2005–2008.

Mich, O., Giuliani, D., Gerosa, M., 2004. Parling: a CALL System for Children. In: Proc. of InSTIL/ICALL. Venice, Italy, pp. 169–172.

Miller, J. D., Lee, S., Uchanski, R. M., Heidbreder, A. H., Richman, B. B., Tadlock, J., May 1996. Creation of Two Children's Speech Databases. In: Proc. of ICASSP. Atlanta, GA, pp. 849–852.

Mirghafori, N., Fosler, E., Morgan, N., 1996. Towards Robustness to Fast Speech in ASR. In: Proc. of ICASSP. Atlanta, GA, USA, pp. 335–338.

Nakamura, M., Iwano, K., Furui, S., Sept. 2005. Analysis of Spectral Space Reduction in Spontaneous Speech and its Effects on Speech Recognition Performances. In: Proc. of EUROSPEECH. Lisbon, Portugal, pp. 3381–3384.

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

Nisimura, R., Lee, A., Saruwatari, H., Shikano, K., May 2004. Public Speech-Oriented Guidance System with Adult and Child Discrimination Capability. In: Proc. of ICASSP. Vol. 1. Montreal, Canada, pp. 433–436.

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

Potamianos, A., Narayanan, S., Lee, S., Sept. 1997. Automatic Speech Recognition for Children. In: Proc. of EUROSPEECH. Rhodes, Greece, pp. 2371–2374.

Russell, M. J., Series, R. W., Wallace, J. L., Brown, C., Skilling, A., Apr. 2000. The STAR system: an interactive pronunciation tutor for young children. Computer Speech and Language 14 (2), 161–175.

Salvi, G., Aug. 2003. Accent clustering in Swedish using the Bhattacharyya distance. In: Proc. of the 15th ICPhS International Congress of Phonetic Sciences. Barcelona, Spain.

Wakita, H., 1977. Normalization of wovels by vocal tract length and its application to vowel identification. IEEE Trans. on Acoustics, Speech and Signal Processing 25, 183–192.

Wegmann, S., McAllaster, D., Orloff, J., Peskin, B., May 1996. Speaker Normalisation on Conversational Telephone Speech. In: Proc. of ICASSP. Atlanta, pp. I–339–341.

Welling, L., Kanthak, S., Ney, H., Apr. 1999. Improved methods for vocal tract normalization. In: Proc. of ICASSP. Vol. 2. Phoenix, AZ, pp. 761–764.

Whiteside, S. P., Hodgson, C., 2000. Speech patterns of children and adults elicited via a picture-naming task: an acoustic study. Speech Communication 32, 267–285.

Wilpon, J. G., Jacobsen, C. N., May 1996. A Study of Speech Recognition for

Children and Elderly. In: Proc. of ICASSP. Atlanta, GA, pp. 349–352.

Young, S. J., Odell, J. J., Woodland, P. C., 1994. Tree-based state tying for high accuracy acoustic modelling. In: HLT '94: Proceedings of the workshop on Human Language Technology. pp. 307–312.

Zheng, J., Franco, H., Weng, F., Sankar, A., Bratt, H., 2000. Word-level rate-of-speech modeling using rate-specific phones and pronunciations. In: Proc. of ICASSP. Vol. 3. pp. 1775–1778.