Acoustic Analysis and Automatic Recognition of Spontaneous Children's Speech

M. Gerosa[*], *D. Giuliani*[*] *and S. Narayanan*[+]

([*]) Centro per la Ricerca Scientifica e Tecnologica, 38050 Pantè di Povo, Trento, Italy
([+]) Speech Analysis and Interpretation Laboratory, Department of Electrical Engeneering, University of Southern California, Los Angeles, CA 90089, USA
gerosa@itc.it, giuliani@itc.it, shri@sipi.usc.edu

Abstract

This paper presents analyses, and recognition experiments, on spontaneous American English speech collected from children aged from 8 to 13 years. These analyses focused on variations in phone duration and on the scattering of phones in the acoustic space and were aimed at achieving a better understanding of spectral and temporal changes occurring in spontaneous speech produced by children of various ages with a view toward developing robust automatic speech recognition applications. The speech data were partitioned in two subsets depending on the annotated presence/absence of explicit occurrences of spontaneous speech phenomena such as fillers, false starts and other disfluencies. All the analyses carried out, as well as the results of recognition experiments, show a significant difference between these two partitions. In particular, recognition performance for the subset containing annotated spontaneous speech phenomena was significantly worse (by almost 15%) than the one achieved for the other subset. Relative improvements due to acoustic model adaptation and normalization on both data partitions were comparable, underscoring that significant performance degradation happens due to spontaneous speech variability beyond those reflected in segmental spectral characteristics.

Index Terms: speech analysis, automatic speech recognition, children's speech, spontaneous speech

1. Introduction

Automatic speech recognition has a huge potential for use by children. In addition to conventional applications in which speech replaces, or complements, other modalities in human-machine interaction, there are applications such as interactive, computer-based pronunciation or reading tutors, or foreign language learning, in which speech is the key enabling technology.

It is well known that characteristics of speech such as pitch, formant frequencies and segmental durations are related to the age of the speakers [1, 2] and the increased variability in children's speech makes the automatic recognition task inherently more difficult for children than for adults. 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 [3, 4], bringing recognition results near those achieved for adult speakers.

However past efforts focused mainly on read speech, while recognition of spontaneous children's speech remains still a less studied and very difficult task due to high linguistic variability and the presence of spontaneous speech phenomena (notably disfluency phenomena such as hesitations, false starts, breath noise, laughter) and speech extraneous to the main dialog topic [5]. Although spontaneous speech effects are quite common in human communication and may be expected to be prevalent in human machine discourse as people become more comfortable conversing with machines, modeling of speech disfluencies is still an open issue. In addition, spontaneous speech is not only linguistically more variable, but is also characterized by larger acoustic variability compared to read speech.

This paper reports results of analysis and recognition on spontaneous speech collected from children aged between 8 and 13 years. The term "spontaneous speech" represents a broad range of characteristics that manifest themselves at different linguistic levels (including segmental, lexical, syntactic and discourse) and as extra-linguistic aspects, notably disfluencies and markers such as laughter. We also note that the type and extent of these effects may vary both within and across interactions and subjects, and should be appropriately reflected in the analysis. As a step toward that, in this paper, we consider speech data with and without explicitly annotated spontaneous speech phenomena separately in the analysis and recognition experiments.

This paper is organized as follows. The speech corpora used are described in Section 2. Section 3 presents some analyses on temporal and spectral characteristics of spontaneous speech, compared to those of read speech. Section 4 describes the automatic speech recognition experiments that were carried out. Final remarks are given in Section 5 which concludes the paper.

2. Speech Corpora

Two different corpora of children's speech were used in this work: the CID read speech corpus, and the CHIMP spontaneous speech corpus. Details of the two corpora are summarized below.

The CID corpus [6] is an American English read speech database collected from 436 children, aged from 5 to 18 with a resolution of 1 year of age, and from 56 adult speakers. Recordings were made in a sound-treated booth 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. The signals were down-sampled to 16 kHz before being analyzed. Only a subset of this database was used in this work. This subset consists of two repetitions of five phonetically rich sentences from six speakers (3 females and 3 males) for each age in the age range 8-13, for a total of 36 subjects.

The CHIMP corpus [5] is an American English spontaneous speech database collected from children aged between 8 and 14 years. This corpus represents spoken dialog interactions collected in a Wizard of Oz (WoZ) experiment using a popular interactive computer game "Where in the U.S.A. is Carmen Sandiego?" designed for children aged eight and older. Data from a total of 160 children and 7 adults were collected. High-quality audio recordings were collected using a close-talking head-mounted micro-

This work was supported in part by the National Science Foundation.

phone (Sennheiser HMD 410) and a far-field desktop microphone (Sennheiser K6-C with a cardioid ME64 capsule). Only the signals collected with the head-mounted microphone from 144 speakers were used in this work. Manual annotation of spontaneous phenomena, like hesitations, filled pauses and non verbal sounds such as laughter and cough, was available on the whole corpus.

The CHIMP corpus was partitioned into test and training sets, summarized in Table 1. The test set, that consists of speech from 6 children (3 females and 3 males) for each age in the age range 8-13, was further partitioned in two subsets: the utterances without any explicitly annotated spontaneous speech phenomena were grouped in a subset called "T1" while the utterances containing annotated spontaneous speech phenomena were grouped in a subset called "T2". Both subsets contained utterances from all the 36 speakers in the CHIMP test set.

Partition	Training	T1+T2	T1	T2
Speaking style	spontaneous			
Signal quality	clean			
Language	American English			
Speaker age	8-14	8-13	8-13	8-13
# speakers	108	36	36	36
# hours	8h:07m	2h:38m	1h:55m	0h:43m
# utterances	29756	9563	7286	2277
# words	89893	34036	24985	9051

Table 1: *Partitioning of the CHIMP speech corpus.*

3. Spontaneous Speech Analysis

This section presents several acoustic analyses on read and spontaneous American English speech collected from children aged from 8 to 13 years. These analyses were carried out in order to obtain a better understanding of the spectral and temporal differences between read and spontaneous speech of children in this age range.

3.1. Phone Duration

Herein, we analyzed phone duration as a function of age on read and spontaneous speech. The mean phone duration was computed by first averaging phone duration over all phones of each speaker and then across all speakers in each age group.

Duration statistics were computed by exploiting a phone-level segmentation produced automatically by forced-alignment. 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.

Figure 1 reports the mean phone duration as a function of age for children, computed on the subset of CID corpus described in Section 2 (6 speakers for each age) and on the CHIMP training set. As expected [1, 2], mean phone duration varies with age and older children exhibit shorter mean phone durations. However, it can be noted that mean phone durations measured on spontaneous speech is much smaller than the ones measured on read speech for all the corresponding age groups. In addition, the effect of age is much less evident in case of spontaneous speech. In fact, while in the case of read speech the mean phone duration for children of age 13 is about 20% smaller than for children of age 8 (138 msec vs. 112 msec), in the case of spontaneous speech this difference was only 10% (99 msec vs. 89 msec). We have to point out that the mean phone durations computed on the CID subset are likely affected by reading ability.

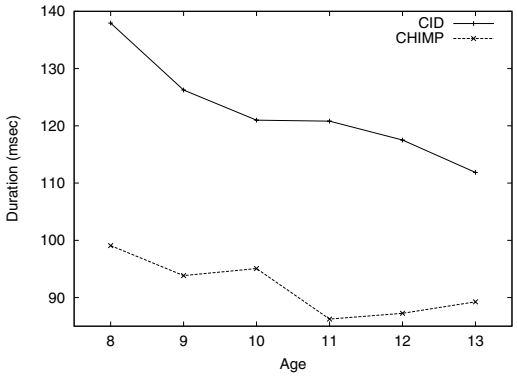

Figure 1: *Mean duration of phones (msec) per age computed on a subset of the CID corpus and on the CHIMP training set.*

3.2. Characterization of the Acoustic Space

In this work we attempted to characterize the acoustic space measuring the scattering of the Gaussian densities modeling phones. For this purpose, we modeled each phone by means of a single Gaussian density and we measured how much these 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 as follows:

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

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.

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 a greater overlap of the phone distributions and thus the phone discrimination task should be more difficult.

To estimate the parameters of Gaussian densities associated to phones, we trained context-independent (CI) HMMs adopting in all cases a three-state left-to-right topology with a single Gaussian density per state. Each speech frame was parameterized into a 13-dimensional observation vector composed of 13 mel frequency cepstral coefficients (MFCCs) plus their first and second order time derivatives. In computing the average Bhattacharyya distance, only Gaussian densities associated to the central states of context-independent HMMs corresponding to vowel sounds were considered. For the experiments reported below, we trained a set of CI HMMs on the T1 subset, on the T2 subset and on the whole

CHIMP test set ("T1+T2"). For comparison purpose, CI HMMs were trained also on the CID subset. Table 2 reports the average Bhattacharyya distance for both spontaneous and read speech.

Test set	T2	T1+T2	T1	CID
Bhattacharyya distance	2.12	2.56	2.76	3.93

Table 2: *Average Bhattacharyya distance across vowel sounds computed on the CID subset and the CHIMP test set partitions.*

The average Bhattacharyya distance computed on HMM sets trained on read speech is greater than the distance computed on HMMs trained using spontaneous speech. In addition, it can be noted that speech in the T2 subset presents a lower distance than speech in the T1 subset. Similar results were reported in [7], where it is shown that the cepstral distribution of spontaneous speech is significantly reduced with respect to that of read speech. However, we have to point out that the high average Bhattacharyya distance obtained for the CID subset may also be due to the limited number of different words uttered.

3.3. Acoustic-space reduction ratio between T1 and T2 subsets

To complement the measures of average Bhattacharyya distance carried out in the previous Section, we tried to quantitatively analyze the reduction of the acoustic space between spontaneous speech in subset T2 with respect to that in subset T1. To do this we adopted the method proposed in [7], by exploiting the same CI HMMs used to compute the average Bhattacharyya distance in the previous Section. First, for each phoneme p, the distance of the mean vector of the Gaussian distribution modeling its central state and the center of the distributions of all phonemes was calculated. Then the ratio between the distance for the HMMs trained on the T2 subset and the distance for the HMMs trained on the T1 subset was calculated as follows:

$$red_p = \frac{||\boldsymbol{\mu}_p^{T2} - Av[\boldsymbol{\mu}^{T2}]||}{||\boldsymbol{\mu}_p^{T1} - Av[\boldsymbol{\mu}^{T1}]||}, \quad (3)$$

where $\boldsymbol{\mu}_p^{T2}$ is the mean vector of phoneme p estimated on subset T2, $\boldsymbol{\mu}_p^{T1}$ is the mean vector of phoneme p estimated on subset T1 and $Av[\cdot]$ indicates the average computed over all the phonemes.

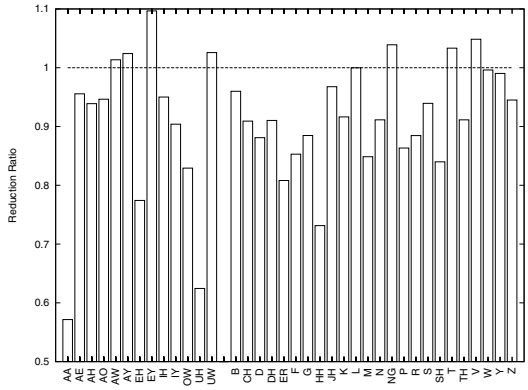

Figure 2: *Reduction ratio between T1 and T2 subsets estimated for each phone.*

Figure 2 shows reduction ratios between T1 and T2 for the basic units that compose the CMU dictionary phone set, used in this work. The units /OY/ and /AX/ which rarely occurred in the utterances considered were not considered in this analysis. It can be noted that the great majority of vowels and consonants show a reduction in cepstrum space when parameters are estimated on the T2 subset with respect to when parameters are estimated on T1 subset. The mean reduction ratios for vowels and consonants was comparable – 0.89 and 0.92, respectively.

4. Recognition Experiments

A set of recognition experiments was carried out with the aim of investigating how much the extent of spontaneous effects of children's speech (as deemed by the human annotations) can influence the robustness of speech recognition performance.

State-tied, cross-word triphone HMMs were adopted for acoustic modeling. In particular, a Phonetic Decision Tree (PDT) was used for tying the states of triphone HMMs. Output distributions associated with HMM states were modeled with mixtures with up to 8 diagonal covariance Gaussian densities.

The CMU dictionary phone set, composed of 39 basic units, was adopted for transcription and annotation of all English corpora considered in our experiments. "Silence" was modeled with a single state HMM. In addition a number of models for common non-verbal phenomena were trained.

Each speech frame was parameterized into a 39-dimensional observation vector composed of 13 mel frequency cepstral coefficients (MFCCs) plus their first and second order time derivatives. Cepstral mean subtraction was performed on static features on an utterance-by-utterance basis. For the experiments considered, a unigram language model implemented with a word-loop FSN having 600 words was deemed adequate and used. Unigram probabilities were estimated as the relative frequency of each word in the training portion of the CHIMP corpus (gains due to adaptive and higher order LMs for this data were reported in [5]).

4.1. Baseline Results

We trained a set of cross-word triphone HMMs using the CHIMP training set. For this set of models, used as a baseline, the PDT tying scheme resulted in about 1000 independent states for a total of about 8000 Gaussian densities.

Table 3 reports recognition results achieved on T1 and T2 test sets by using baseline models. For comparison the results on the whole CHIMP test set ("T1+T2") are also reported.

	Test Set		
HMM set	T1+T2	T1	T2
Baseline	34.8	30.1	48.0

Table 3: *Recognition results (% WER) obtained on the CHIMP test sets.*

The Table shows that the difference in performance between the two subsets T1 and T2 is significant. In fact, the WER, achieved by using the baseline system on T2 test set (48.0%) is about 60% higher than WER achieved on T1 test set (30.1%).

We tried to correlate these recognition results with the spontaneous effects shown by the utterances of each speaker. For this purpose we first computed the WER for each speaker in the whole CHIMP test set and then we computed the correlation between the WER and some clues revealing the extent of spontaneous effects present in the speech of a particular speaker. The three clues that we analyzed for each test speaker were the percentage of spontaneous sentences uttered (in the games they played), the number of different words used, and the number of hesitations and filled pauses normalized by the number of games played by each speaker. For each of these characteristics we estimated the correlation coefficient with respect to the WER. Each characteristic presented a positive correlation with WER: the most correlated was the number of hesitations/filled pauses (+0.65), followed by the percentage of sentences with annotations marking spontaneous effects (+0.60). In comparison, the number of different words was correlated to a lower degree (+0.5) with WER.

4.2. Speaker Adaptive Acoustic Modeling

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

In particular, the training and recognition procedures adopted for implementing VTLN follow closely those proposed in [8] and are described in detail in [4]. For implementing SAT, we adopted the variant of the SAT algorithm developed by Gales [9], which makes use of an affine transformation, estimated through constrained MLLR, for mapping acoustic observations of each training and testing speaker. CMLSN is a speaker normalization method which performs speaker normalization by transforming the acoustic observation vectors by means of speaker-specific constrained MLLR transformations. Details about this method can be found in [4].

Three HMM sets were trained using the SAT, VTLN and CMLSN training procedures on the CHIMP training set. In this case, we assumed that the data of each test speaker were available in block for multiple processing. The decoder was run twice, and the output of the first decoding step was exploited as a supervision for system adaptation/normalization before the second decoding step took place. In addition to speaker normalization, unsupervised static speaker adaptation of acoustic models was performed by adapting means and variances of Gaussian densities through MLLR before the second decoding step. Two regression classes were defined and the associated transformation matrices were estimated through three MLLR iterations exploiting the data of each speaker.

Table 4 reports recognition results achieved on the CHIMP test sets by using baseline models ("2-pass Baseline") and models trained using the VTLN, CMLSN and SAT methods.

	Test Set		
HMM set	T1+T2	T1	T2
2-pass Baseline	31.5	26.4	45.8
VTLN	31.0	25.5	43.9
CMLSN	29.9	25.0	43.3
SAT	30.1	25.3	43.3

Table 4: *Recognition results (% WER) obtained on the CHIMP test sets using HMMs trained with and without speaker adaptive acoustic modeling methods.*

All the results reported in Table 4 were achieved with the second decoding pass after performing MLLR model adaptation. When recognizing with HMMs trained using speaker adaptive acoustic modeling procedures, the difference in performance achieved on the two subsets remained almost the same with all the three normalization methods. With the CMLSN method, the relative reduction in WER on T2 is 5.5% compared to 4.1% on T1 (compare values in rows "CMLSN" and "2-pass baseline"). There is limited effectiveness of speaker adaptive acoustic modeling techniques on the CHIMP test set. The reason for this has not been fully understood. However, one possible explanation is that the high error rate in the CHIMP test set is due to linguistic/acoustic factors different from inter-speaker acoustic variability, and thus speaker adaptive acoustic modeling techniques have only limited impact in this case.

5. Conclusions

In this paper, analyses on spontaneous children's speech were presented. These analyses focused on phone duration and on the scattering of phones in the acoustic space. It was found that phone duration for spontaneous speech is significantly lower than that for read speech uttered by children of the same age. Phone duration decreases as age increases for both read and spontaneous speech: however the decrease observed between speech uttered by children of age 8 and age 13 for read speech is almost twice as that observed for spontaneous speech.

Measurements on scattering of phones in the acoustic space confirm those reported in [7] for adult speech. Phone distributions estimated on read speech are more scattered in the acoustic space than phone distributions estimated on spontaneous speech and thus for the latter the phone discrimination task should be more difficult. All the analyses carried out show a significant difference between the two partitions of the CHIMP test set considered, T1 and T2. This difference is reflected in the recognition results obtained for the two subsets: in fact WER for T2 is 60% higher than WER for T1 (48.0% WER vs. 30.1% WER). However, it can be assumed that this difference in performance is likely to be caused more by the presence of spontaneous speech phenomena that are at a level beyond that manifested in the segmental acoustic differences between the two subsets.

Finally, we have to point out that speaker normalization techniques showed only a limited effectiveness on the CHIMP test set (and comparable performance on data with and without predominance of spontaneous effects as given by human annotations). Robust ASR solutions should hence consider models beyond the conventional acoustic adaptation and normalization; these are topics for our future work.

6. References

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

[2] M. Gerosa, S. Lee, D. Giuliani, and S. Narayanan, "Analyzing Children's Speech: An acoustic Study of Consonants and Consonant-Vowel Transition," in *Proc. of ICASSP*, Toulouse, France, May 2006, pp. 393–396.

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

[4] D. Giuliani, M. Gerosa, and F. Brugnara, "Improved automatic speech recognition through speaker normalization.," *Computer Speech and Language*, vol. 20, no. 1, pp. 107–123, Jan. 2006.

[5] S. Narayanan and A. Potamianos, "Creating Conversational Interfaces for Children," *IEEE Trans. on Speech and Audio Processing*, vol. 10, no. 2, pp. 65–78, Feb. 2002.

[6] J. D. Miller, S. Lee, R. M. Uchanski, A. H. Heidbreder, B. B. Richman, and J. Tadlock, "Creation of Two Children's Speech Databases," in *Proc. of ICASSP*, Atlanta, GA, May 1996, pp. 849–852.

[7] M. Nakamura, K. Iwano, and S. Furui, "Analysis of Spectral Space Reduction in Spontaneous Speech and Its Effects on Speech Recognition Performances," in *EUROSPEECH*, Lisbon, Portugal, Sept. 2005, pp. 3381–3384.

[8] L. Welling, S. Kanthak, and H. Ney, "Improved methods for vocal tract normalization," in *Proc. of ICASSP*, Phoenix, AZ, Apr. 1999, vol. 2, pp. 761–764.

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