A Comparison of Read and Spontaneous Children's Speech Recognition

Matteo Gerosa
FBK Fondazione Bruno Kessler
Povo (TN), Italy
gerosa@fbk.eu

Diego Giuliani
FBK Fondazione Bruno Kessler
Povo (TN), Italy
giuliani@fbk.eu

ABSTRACT

This paper presents comparative analyses, and recognition experiments, on read and spontaneous Italian speech collected from children. The presented analyses focus on linguistic variations, variations in phone duration, and the scattering of phones in the acoustic space. The aim of these analyses is to achieve a better understanding of acoustic and linguistic difference between read and spontaneous speech uttered by children in the same age range (9-11).

A recognition system was developed exploiting clean read speech, collected from children aged 7-13, and written texts. Results of phone and word recognition experiments, carried out with this system on read and spontaneous speech, are presented. Results of recognition experiments show that very high recognition performance can be achieved on clean read children's speech (6.9% phone error rare). However, performance drops drastically when the system is applied to spontaneous speech collected from children (27.2% phone error rate).

1. INTRODUCTION

Automatic recognition of children's speech is a difficult task especially when targeting younger children. Characteristics of speech such as pitch, formant frequencies and segmental durations have been shown, in fact, to be related to the age of the speakers. Several studies have shown that intra- and inter-speaker acoustic variability is higher for younger children than for older children and adults [7, 4]. However, in recent years it has been shown that on clean read speech it is possible to achieve for children recognition performance comparable to that achieved for adults [3, 4], at least considering children aged 7 and up.

Recognition of spontaneous children's speech, on the other hand, remains still much more problematic [10, 1]. Spontaneous speech is characterized, in fact, by a higher linguistic variability and the presence of spontaneous speech phenomena such as disfluencies (e.g. hesitations, filled pauses, repeated words, and sentence restarts), extraneous speech (e.g. self-talk) and "noise" events (e.g. loud breath and lip smacks) [10, 5, 1]. Spontaneous speech is not only linguistically more variable, but is also characterized by an higher acoustic variability than read speech [5].

The term "spontaneous speech" represents, however, a broad category that contains several different types of speech. We can assume that speech with different "degrees of spontaneity" presents different characteristics and should be addressed separately. In fact, as people become more comfortable conversing with machines we can expect human-computer interaction to become more similar to human-human interaction and thus more difficult to process for a speech recognition system. For example in [11] it was shown that children's disfluency rates were substantially higher during human-human communication than during human-computer communication.

In this work we coped with speech elicited in a human-human communication task. Children were interviewed by an adult about his/her preferred books, TV shows, hobbies, sports, etc. We investigated how a speech recognition system, developed for the Italian language by exploiting mainly read children speech and written texts, behaves when applied to recognize children's speech collected in this challenging context. Word recognition experiments were carried out exploiting 4-gram language models (LMs) having vocabularies of different size: 10k, 64k and 1210k words. Phone recognition experiments were carried out by exploiting several n-gram phone-based LMs: 7-gram, 5-gram, 3-gram, 2-gram and phone-loop LMs.

The remainder of this paper is organized as follows. The speech corpora used are described in Section 2. Section 3 presents analyses carried out on read and spontaneous speech. Section 4 describes the automatic speech recognition experiments that were carried out and discusses the recognition results achieved. Final remarks are given in Section 5 which concludes the paper.

2. CORPORA

For Acoustic Models (AMs) training we used the ChildIt speech corpus [4], augmented with additional read and spontaneous speech data collected at FBK.

The ChildIt corpus 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. About 10 hours of speech were collected from 171 children. 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.

Additional data used for AM training included 2h:14m of read speech collected from 44 children and 53m of spontaneous speech collected from 6 children.

We employed two test sets: the test set partition of the ChildIt database and the SpontIt corpus, 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. Each of these 21 children were interviewed by an adult about his/her preferred books, TV shows, hobbies, sports, etc.

All training and testing data were acquired with the same head-worn microphone. However, while read speech was acquired at 16 kHz, with 16 bit accuracy, by using the A/D board of a PC, for spontaneous speech recordings were performed by using a digital audio tape recorder and then downsampling audio signals from 48 kHz to 16 kHz, with 16 bit accuracy.

Characteristics of speech corpora used in this work are reported in Table 1.

Table 1: Main characteristics of speech corpora used for training and testing

	Training	Testing	
		ChildIt	SpontIt
Speaking mode	Read/Spont.	Read	Spont.
Speaker age	7-13	7-13	8-12
# of speakers	179	42	21
# words	72307	15355	9838
Rec. hours	10h:54m	2h:29m	1h:20m

3. ANALYSIS OF THE CORPORA

This section presents several acoustic analyses on children's read and spontaneous speech. These analyses were carried out in order to achieve a better understanding of the differences between speech in the ChildIt and the SpontIt corpora. Analyses were carried out on a subset of the ChildIt and the SpontIt corpora. Since the SpontIt corpus contains speech from only 2 children of age 8 and 1 child of age 12, we limited our analysis to the age range 9-11. A subset of the ChildIt corpus was selected so as to replicate exactly the distribution of speakers by age and gender in SpontIt. Note that children included in the ChildIt and SpontIt corpora are different.

Table 2 shows the distribution of speakers, by age and gender, in the subsets considered.

Table 2: Speaker distribution by age and gender for the ChildIt and SpontIt subsets considered for analysis

	Age 9	Age 10	Age 11	All Ages
Male	2	4	2	8
Female	2	2	6	10
Male+Female	4	6	8	18

3.1 Phone Duration

Mean phone duration was computed 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. Each utterance was time-aligned with the triphone Hidden Markov Model (HMM) concatenation corresponding to the uttered words, allowing insertion of an optional "silence" model between words at the beginning and at the end of the utterance. Segments of signals aligned with the "silence" HMM were not taken into account in computing temporal statistics. Tables 3 reports the mean phone duration computed on ChildIt and SpontIt subsets described above.

Table 3: Mean phone duration (ms) measured on the ChildIt and SpontIt subsets

	ChildIt	SpontIt
Age 9	97.4	95.6
Age 10	92.7	89.2
Age 11	91.1	86.0
Age 9-11	92.9	89.7

It can be noted that there is a small decrease in phone duration as the age increases. The decrease is consistent with the one reported in [4] for the ChildIt corpus for the age range 7-13. It can also be noted that there is only a small difference in mean phone durations computed on read and spontaneous speech.

3.2 Rate of Speech

In addition to phone duration we measured the rate of speech in terms of words per second. Duration statistics were computed from the same segmentation obtained for the phone duration analysis. As before, segments of signals aligned with the "silence" HMM were not taken into account in computing temporal statistics. The computed average number of words per second are reported in Table 4.

Table 4: Average rate of speech (words/sec) measured on the ChildIt and SpontIt subsets

	ChildIt	SpontIt
Age 9	2.24	2.73
Age 10	2.34	2.93
Age 11	2.34	3.00
Age 9-11	2.32	2.90

It can be noted that while there is only a small difference in mean phone duration between read and spontaneous speech, the rate of speech for spontaneous speech is consistently higher than for read speech. This is explained by the fact that words contained in the SpontIt subset are significantly shorter than the ones in the ChildIt subset: the average number of phones per word is 3.8 for SpontIt and 4.6 for ChildIt.

3.3 Characterization of the Acoustic Space

By using the same method presented in [4], we tried to characterize the acoustic space by measuring the scattering of the observation densities of phone models. Each phone is first modeled with a single Gaussian density and then the Bhattacharyya distance [2] is used for measuring how much Gaussian densities are scattered in the acoustic feature 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 Bhattacharyya distance between them is given by:

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

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 [8, 12, 6].

To estimate the parameters of Gaussian densities associated to phones, we trained two sets of acoustic models on the ChildIt and SpontIt subsets, respectively. Each set of HMMs was formed by phone models having a three state left-to-right topology and output distributions associated to states modeled by a single Gaussian density. Each speech frame was parametrized into a 39-dimensional observation vector composed of 13 mel frequency cepstral coefficients (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.

To measure how scattered Gaussian densities are in the acoustic feature space, for each set of models, the average Bhattacharyya distance was computed by considering only Gaussian densities associated to the central states of HMMs representing vowels [4]. Table 5 reports the average Bhattacharyya distance for vowels on the ChildIt and SpontIt subsets.

Table 5: Average Bhattacharyya distance between vowels computed on the ChildIt and SpontIt subsets

ChildIt	SpontIt
5.11	3.57

The average Bhattacharyya distance computed on the HMM set trained on read speech is greater than the distance computed on HMMs trained using spontaneous speech. These values are compatible with the ones computed on the ChildIt training set and on adult speech reported in [4]. Similar results are reported in [9], where it is shown a significant reduction of the cepstral space of spontaneous speech with respect to that of read speech.

3.4 Analysis of the Linguistic Content

In view of speech recognition experiments we analyzed the linguistic content of the ChildIt and SpontIt test sets. Table 6 reports some statistics computed on speech from all children included in the two test sets. We can see that the 2 test sets have very different characteristics as expected due to the different nature of the two corpora. The number of different words used in spontaneous speech is much lower than in read speech. The average number of words per sentence is significantly higher in the SpontIt corpus. The number of truncated words and spontaneous phenomena is, not surprisingly, higher in the SpontIt test set. We point out that in the reported statistics only some types of spontaneous speech phenomena are counted such as filled pauses and non-verbal sounds.

Table 6: Statistics on the ChildIt and SpontIt test sets

	ChildIt	SpontIt
# Words	15355	9838
# Unique words	4910	2065
# Words / # Unique words	3.1	4.8
# Sentences	2219	1054
# Words per sentence	6.9	9.3
# Phones per word	4.6	3.8
# Truncated words	18	164
# Spontaneous phenomena	9	381

Table 7 reports the out-of-vocabulary (OOV) rate computed on the two test sets assuming recognition vocabularies of different sizes. A text corpus of 600M words, mainly composed of newspaper articles, was exploited for LMs training. Vocabularies of different size were obtained sorting by frequency the words in the training corpus and selecting the N most frequent words, with N equal to 2k, 5k, 10k, 20k and 64k. Interestingly, for vocabulary sizes of up to 20k words the OOV rates computed on the ChildIt test set are much higher than the ones computed on the SpontIt corpus. We can conclude that in the spontaneous speech interactions analyzed children prefer to use more common and shorter words than what is generally found in the literature for children from which we extracted the texts composing the ChildIt corpus.

The OOV rates computed with the 64k word vocabulary are similar for the SpontIt and the ChildIt test sets. With the full 1200k word vocabulary the OOV rate for the read speech test set is close to zero (0.2%), while in the SpontIt corpus the OOV rate is 2.1%. This difference is mainly caused by the high number of truncated words, which are not included in the recognition vocabulary, contained in the SpontIt corpus.

Table 7: OOV rate on the ChildIt and SpontIt test sets with recognition vocabularies having different size. Perplexity obtained using the 1200k word *4*-gram LM is also reported in the last row

	ChildIt	SpontIt
OOV 2k	36.6% (5626)	23.7% (2337)
OOV 5k	27.9% (4296)	17.1% (1684)
OOV 10k	20.6% (3170)	12.0% (1182)
OOV 20k	14.1% (2170)	8.1% (797)
OOV 64k	4.8% (717)	4.2% (419)
OOV 1210k	0.2% (30)	2.1% (206)
PP 1210k	875	622

The high perplexity (PP) reported in Table 7 for the two test sets can be explained by the fact that the *4*-gram statistics estimated on the training text corpus, which is mainly composed of newspaper articles, do not reflect well the statistics of the ChildIt and SpontIt test sets.

4. RECOGNITION EXPERIMENTS

Each speech frame was parametrized into a 39-dimensional observation vector composed of 13 MFCCs plus their first and second order time derivatives as specified above.

Acoustic models were state-tied, cross-word triphone HMMs. In particular, a phonetic decision tree was used for tying the states of triphone HMMs. Output distributions associated with HMM states were modeled with mixtures having up to 32 diagonal covariance Gaussian densities, for a total of about 21000 Gaussian densities in the HMM set. Models were conventionally trained on the speech data described in Table 1.

Several *4*-gram LMs were estimated on the 600M word corpus by considering vocabularies of different sizes. Phone based LMs, with different n-gram orders, were also estimated on the same corpus after mapping words into sequences of phone labels: in particular we estimated 7-gram, 5-gram, 3-gram and 2-gram LMs.

4.1 Phone Recognition Results

Table 8 reports phone recognition results obtained employing phone based LMs with different n-gram orders. We can note that with a simple phone-loop finite state network a Phone Error Rate (PER) of 19.9% is achieved on read speech while a 40.6% PER is achieved on spontaneous speech. This drop in performance can be explained by the fact that the system was mainly trained on read speech and that recognition of spontaneous speech is inherently more difficult than recognition of read speech.

On the other hand, as the n-gram order increases, the relative difference in performance between read and spontaneous speech increases: the PER for spontaneous speech is about 100% higher than the PER obtained for read speech using a phone loop, and about 300% higher using a 7-gram LM. This is probably caused by disfluency phenomena, like repetitions, revisions and restarts, that generate phone sequences hardly ever seen in the LM training data.

Table 8: Phone recognition results (% PER) on the ChildIt and SpontIt test sets using phone based LMs with different n-gram orders

	ChildIt	SpontIt
7-grams	6.90%	27.19%
5-grams	10.12%	29.79%
3-grams	15.24%	33.81%
2-grams	16.55%	34.68%
phone-loop	19.94%	40.57%

The 6.9% PER achieved on read speech by using the 7-gram LM is comparable with the PER we usually obtain on adult speech under similar experimental conditions.

4.2 Word Recognition Results

Table 9 reports word recognition results using *4*-gram LMs having vocabularies of different size.

Increasing the size of the recognition vocabulary leads to a large decrease in Word Error Rate (WER) on the read speech test set. This is mainly due to the decrease in OOV rate (from 20.6% for 10k words vocabulary to 0.2% for 1210k word vocabulary). For spontaneous speech there is little gain in increasing the recognition vocabulary: in fact the difference in WER is even lower than the decrease in OOV rate (from 12.0% for 10k word vocabulary to 2.1% for 1210k word vocabulary).

Table 9: Word recognition results (% WER) obtained on the ChildIt and SpontIt test sets using 4-gram LMs having vocabularies of different size

	ChildIt	SpontIt
10k	47.09%	58.62%
64k	19.84%	52.55%
1210k	13.77%	51.69%

The difference in performance achieved on the two test sets increases with the recognition vocabulary size. This is an effect similar to the one observed when increasing the n-gram order in phone recognition experiments and similar considerations can be done. The LM was, in fact, estimated on written texts and thus it is inadequate for modeling disfluency phenomena like repetitions, revisions and restarts.

5. CONCLUSIONS

In this paper we have presented analyses, and recognition experiments, on read and spontaneous speech collected from Italian children. The spontaneous speech corpus was particularly challenging, being composed of speech collected during human-human interactions (i.e. children interviewed by an adult).

The phone duration analyses performed have shown phone duration values consistent with those reported in literature for the age range considered (9-11). Interestingly, on the corpora considered, there was only a small difference in average phone duration between read and spontaneous speech. Characterization of the acoustic space based on the average Bhattacharyya distance between Gaussian distributions modeling vowel sounds, on the other hand, has shown a clear difference between read and spontaneous speech. This result suggests that vowel sounds in the spontaneous speech corpus are less scattered in the acoustic space and therefore more confusable than in the read speech corpus, making the ASR task more difficult.

Results of recognition experiments have shown that a children's speech recognition system developed exploiting read speech and written texts can ensure very high recognition performance on clean read speech. This is especially true for phone recognition experiments (6.9% PER) while for word recognition experiments the performance is probably limited by the mismatch between the LM and the linguistic content of the utterances to be recognized. On the other hand, performance drops drastically when the system is applied to spontaneous speech collected from children interviewed by an adult. This drop in performance was higher than expected, however we were not able to find explanations other than the higher intrinsic difficulty of recognizing this kind of speech.

6. REFERENCES

[1] T. Cincarek, I. Shindo, T. Toda, H. Saruwatari, and K. Shikano. Development of Preschool Children Subsystem for ASR and Q&A in a Real-environment Speech-oriented Guidance Task. In *Proc. of INTERSPEECH*, pages 1469–1472, Antwerp, Belgium, 2007.

[2] K. Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press, New York, second edition, 1990.

[3] M. Gerosa, D. Giuliani, and F. Brugnara. Speaker Adaptive Acoustic Modeling with Mixture of Adult and Children's Speech. In *Proc. of INTERSPEECH*, pages 2193–2196, Lisboa, Portugal, Sep. 2005.

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

[5] M. Gerosa, D. Giuliani, and S. Narayanan. Acoustic analysis and automatic recognition of spontaneous children's speech. In *Proc. of INTERSPEECH*, Pittsburgh, PA, Sep. 2006.

[6] S. C. Kumar, V. P. Mohandas, and H. Li. Multilingual Speech Recognition: A Unified Approach. In *Proc. of INTERSPEECH*, pages 3357–3360, Lisboa, Portugal, Sept. 2005.

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

[8] B. Mak and E. Barnard. Phone Clustering Using the Bhattacharyya Distance. In *Proc. of ICSLP*, pages 2005–2008, Philadelphia, PA, Oct. 1996.

[9] M. Nakamura, K. Iwano, and S. Furui. Analysis of Spectral Space Reduction in Spontaneous Speech and its Effects on Speech Recognition Performances. In *Proc. of INTERSPEECH*, pages 3381–3384, Lisbon, Portugal, Sept. 2005.

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

[11] S. Oviatt. Talking to thimble jellies: Children's conversational speech with animated characters. In *Proc. of ICSLP*, Beijing, China, Oct. 2000.

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