The PF_STAR Children's Speech Corpus

Anton Batliner[1], *Mats Blomberg*[2], *Shona D'Arcy*[3], *Daniel Elenius*[2], *Diego Giuliani*[4], *Matteo Gerosa*[4], *Christian Hacker*[1], *Martin Russell*[3], *Stefan Steidl*[1], *Michael Wong*[3]

[1]University of Erlangen, Germany, [2]Kungl Tekniska Högskolan, Sweden
[3]The University of Birmingham, UK, [4]ITC-irst, Italy
m.j.russell@bham.ac.uk

Abstract

This paper describes the corpus of recordings of children's speech which was collected as part of the EU FP5 PF_STAR project. The corpus contains more than 60 hours of speech, including read and imitated native-language speech in British English, German and Swedish, read and imitated non-native-language English speech from German, Italian and Swedish children, and native-language spontaneous and emotional speech in English and German.

1. Introduction

This paper describes a multi-lingual corpus of native and non-native, read and spontaneous children's speech which was collected as part of the EU FP5 'PF_STAR' project[1].

Speech technology has huge potential for use by children. In addition to its conventional utility, there are applications such as interactive, computer-based pronunciation or reading tuition, or foreign language learning, in which speech is the key enabling technology (e.g [1]). It is surprising, therefore, that relatively little research has focussed on the development of speech technologies for children. It has been demonstrated that automatic speech recognition error rates are typically 100% greater for young children than for adults [2], that children exhibit significantly more intra- and inter-speaker variability [3], that bandwidth-restriction is a major contributor to error, and that teacher's assessment of pronunciation 'quality' correlate with recognition performance [4]. Some success has been reported using 'frequency warping' techniques to compensate for the effects of children's shorter vocal tracts [5]. However, most research has focussed on American English, little work has been done on European languages, and little is known about the effects of non-native language or spontaneous speech on automatic recognition of children's speech.

The goal of PF_STAR 'workpackage 5' was to establish baselines for automatic recognition of children's speech in British English, German, Italian and Swedish, to determine the effects of non-native language and spontaneity on recogniser performance, and to develop techniques to raise these performance baselines to a similar level to those for adult speech. To achieve this, there was a requirement for substantial corpora of transcribed children's speech data. This paper describes the corpora which were collected in the first year of the project. In total, approximately 63 hours of children's speech have been recorded, including approximately 16 hours of spontaneous and emotional speech, 47 hours of read (or imitated) speech, and 8 hours of non-native English children's speech.

[1]This work was conducted as part of EU FP5 PF_STAR (Preparing Future Multisensorial Interaction Research)

2. Corpus Design Considerations

The corpora were designed to enable a range of children's speech recognition 'baselines' to be established, including:

- Automatic recognition of native language, read speech from English, German, Italian and Swedish children.
- Automatic recognition of native language, spontaneous and emotional speech from English and German children.
- Automatic recognition of non-native language, read children's speech from German, Italian and Swedish children speaking English
- Human recognition of native-language English children's speech.

2.1. Recording Procedure

Where possible common recording standards were agreed. Common materials (the ITC-irst texts) and recording procedures were used for all non-native English recordings. In addition, the same 'AIBO' methodology was used at the two sites involved in the collection of spontaneous and emotional children's speech.

3. Common Recording Materials and Methodologies

Materials recorded for the native language corpora were chosen by the individual laboratories. However, in the case of the non-native language recordings in English and the recordings of spontaneous and emotional speech, the following common materials or procedures were agreed:

3.1. The ITC-irst English Words and Phrases

All of the ITC-irst English texts were chosen by ITC-irst in consultation with Italian teachers of English. They were judged to be appropriate for reading by 10 year old Italian children learning English.

3.1.1. The ITC-irst isolated words: ITC-W

This is a list of 220 English words, divided into 5 lists of 44 words so that each list covers the basic sounds of British English. These words were recorded by English, Italian and Swedish children.

3.1.2. ITC-irst 'phonetically rich' sentences: ITC-S

This is a set of 50 phonetically rich English sentences, which were divided into 5 lists of 10 sentences and cover the basic

sounds of British English. These sentences were recorded by English, Italian and Swedish children.

3.1.3. ITC-irst 'generic' sentences: ITC-G

This is a list of 400 sentences, divided into 40 lists of 10 sentences. These were recorded by English, Italian and Swedish children.

3.2. The 'AIBO' Methodology for Spontaneous and Emotional Speech

Recordings of spontaneous, emotional, un-scripted children's speech were be made using the 'AIBO' methodology developed at the University of Erlangen [6]. A child gives spoken instructions to a Sony AIBO robot and believes that AIBO is responding to his or her commands. In reality AIBO is being controlled by a human 'Wizard-of-Oz'. Three versions of the experiment were conducted:

3.2.1. AIBO Experiment 1: "Parcours", 'obedient AIBO'

The child is asked to guide AIBO around a map, starting and ending at squares labelled 'start' and 'finish'. The map is printed on a floor carpet measuring approximately 2m $\times$ 3m. A number of cups are placed on the map, and the child was instructed to make AIBO look into each cup. Special squares are labelled with instructions (e.g. 'dance'), and the child is asked to make AIBO obey the instruction when it reaches the square. The 'Wizard' listens and watches through one or more video cameras. In Experiment 1 the Wizard tries to make AIBO follow the child's instructions, as if AIBO is controlled by a very high performance spoken language understanding system.

3.2.2. AIBO Experiment 2: "Parcours", 'disobedient AIBO'

The child is given the same task and instructions as in experiment 1. However, the Wizard causes AIBO to perform a fixed, pre-determined sequence of actions, which takes no account of what the child says. The motivation was to elicit emotional speech from the child. Experiment 2 simulates a very poor spoken language processing system.

3.2.3. AIBO Experiment 3: Object Localisation

In each of 5 'object localisation' tasks, children were instructed to direct AIBO towards one of several cups on the carpet. One of these cup was "poisoned". The children applied different strategies to direct AIBO. In the first task AIBO was 'obedient' to persuade the children that it could understand their commands. In the other tasks AIBO was 'disobedient'. In some tasks AIBO went directly toward the "poisoned" cup in order to evoke emotional speech from the children.

4. Native language, prompted speech recordings

Recordings of native language children's speech were made in British English, German and Swedish. The German recordings supplement the existing *Fluency* corpus. Experiments on native Italian children's speech used the existing *ChildIt* corpus.

4.1. Read German speech (extended 'Fluency' Corpus)

Read native-language speech was collected from 53 German children (age 10 - 13, 23 male, 30 female). Fifty-one of the children are the same as for the German 'AIBO' spontaneous speech recordings (6.1). The children read unfamiliar texts. The vocabulary size for the recorded data is 9,371 'words', including 1,306 fragments and non-words. The speech was recorded with a head-mounted microphone with preamplifier (dnt Call4U Comfort) and sampled at 44.1 kHz, with 16 bit quantisation, via the sound card of the data collection computer.

4.2. Prompted Swedish speech

The objective was to produce a generic corpus of Swedish children's speech, suitable for task-independent phone model training or adaptation. The recordings were made in the Stockholm area with low variation in the recording environment. Recordings were made in separate rooms at day-care and after-school centres. The children were prompted by an adult who read the text from a screen.

The microphones used were two of those used in the EU project 'Speecon', namely a headset (Sennheiser ME 104) and an omni-directional table microphone (MBF Haun). The latter unit was placed on a table approximately 50 to 100 cm from the subject. The recordings were made directly onto computer disk, at 32 kHz sampling rate and 24 bit resolution, using an external A/D converter connected to the USB port (to avoid any variation due to different computer sound cards). For each child a reference, 1kHz constant-amplitude sinewave was recorded via the microphone connector for calibration.

4.2.1. Material recorded

The children were divided into two groups: 4-5 year olds and 6-8 year olds, and spoke utterances from four different categories. The numbers of utterances in each category for 4-5 (respectively, 6-8) year olds are: 6 (6) answers to questions, 30 (60) sentences, 10 (10) 3-digit sequences, and 10 (10) name lists. There are 993 pairs of names and 1,255 distinct sentences for the younger group, and 1,000 name triples and and 2,513 sentences for the older group. The younger children spoke fewer sentences, which were shorter to facilitate correct pronunciation. The number of recorded children in the Swedish corpus is 198.

The recordings are annotated with phonetic transcriptions plus information on external noise and extra-linguistic sounds from the speaker. Pronunciation errors, truncated utterances and unintelligible stretches of speech are also marked. These events were labelled according to conventions from the 'SpeechDat' and 'SpeeCon' corpora.

4.3. Read (native) English speech

Recordings were made of 159 English children aged 4 to 14 in two Primary Schools; one in Birmingham (54 children, 24 male and 34 female) and one in Malvern (75 children, 40 male, 35 female) and at the University of Birmingham (30 children, 18 male, 12 female). The 'Malvern' recordings were made in an open library area in the school. On occassions (e.g. between lessons or at break times) this was busy and noisy, so that some of the recordings include varying degrees of background noise. The recordings at the Birmingham school were made in a closed room, and the noise levels are lower. The recordings at the university were made in an audiometric booth. The 30 children who made recordings at the university also took part in the Birmingham 'AIBO' recordings (6.2).

The recordings made at the schools used an Emkay head mounted microphone. Initial recordings also used a Telex desk

mounted microphone, but these are likely to be of limited value due to the level of background noise. At the university, parallel recordings were made using the Emkay and Shure SM10A microphones, attached to the same head mount. Recordings were sampled at 16kHz, using an Edirol UA-5 external sound card with USB interface. Prompts were displayed on a laptop.

4.3.1. Material Recorded

It was intended that each native English subject should read the following material: 20 'SCRIBE' sentences, 44 ITC-irst isolated words (3.1.1), 10 ITC-irst 'phonetically rich' sentences (3.1.2), 20 ITC-irst 'generic' phrases (3.1.3), an 'accent diagnostic' passage, and 20 digit triples. Some very young children could not manage all of this and only a subset was recorded.

- **The SCRIBE Sentences** are an anglicised version of the phonetically balanced US TIMIT 460 sentence set. Each SCRIBE sentence was graded according to its difficulty of pronunciation, and subsets were defined for reading by children aged between 5 and 15. For each age group the SCRIBE sentences were partitioned into lists of 10 sentences. SCRIBE sentences were included to enable comparison with data in existing corpora of adult speech.
- **The British English 'Accent Diagnostic' passage** is a standard text which can be used for accent classification in English. The passage was too difficult for some younger children.

The statistics of the read and imitated, native language component of the PF_STAR children's speech corpus are summarised in table 1.

Table 1: *Native language children's speech recordings*

language	style	num chldren	ages	hours
English	read	129	6-11	10
Swedish	imitated	99	4-5	7
Swedish	imitated	99	6-8	10
German	read	53	10-13	9.3

5. Non-native language, English read speech recordings

5.1. Non-native English recordings by Italian children - the 'ChildEn' corpus

The ChildEn corpus comprises English sentences read by Italian children. It was designed and collected by ITC-irst. The corpus includes speech data from 78 children (44 male and 34 female) aged around 10 from two primary schools in the Trento area (NE Italy). The children were selected from the fifth grade and had been studying English for 3 or 4 years.

Recording took place in a school computer room or library. Each child was taken out of class for about 25 minutes. During a recording session each child was assisted by an adult operator. The prompt text was presented on a computer screen and the child was asked to read it aloud. If necessary the text was read more than one time and just the last repetition was stored. In the first set of recordings, the child listened to a reference pronunciation, by a native adult speaker, one or more times before speaking. For unknown words, the child tended to imitate these reference pronunciations. The speech acquired in this way is referred to as 'imitated' speech. A second set of recordings was made in another school without these reference pronunciations. However, prompt texts were introduced to the children by the teacher some days earlier. For isolated words, a picture representing the meaning of the word was displayed on the screen in addition to the text. These recordings are referred to as 'read' speech. The recordings used a Shure SM10A head-worn microphone attached to a pre-amplifier. Speech was sampled at 16kHz with a resolution of 16 bits per sample, using the computer's internal A/D converter.

In the first set of recordings each child read aloud 44 words (3.1.1), 10 phonetically rich sentences (3.1.2), and 10 generic sentences (3.1.3). The same lists of words and phonetically rich sentences were used for every 5th child, so that some prompt texts were read more than one time. The second recording sessions used different prompts. Two lists of 50 words were prepared, with each list selected so as to ensure coverage of the basic sounds of British English. Most of these words came from the lists in (3.1.1). Two lists of 25 phonetically rich sentences were also prepared, each obtained by merging 2 lists from (3.1.1) and adding 5 more sentences. During recording sessions, children were asked to read alternatively one of two sets of prompt texts, each consisting of 50 words and 25 phonetically rich sentences.

In summary, the 'imitated' part of the corpus contains speech material from 53 speakers (26 males and 27 females). A total of 672 different prompt texts were read one or more times, giving a total of 3,393 utterances. The overall vocabulary size is 618. The "read" part of the corpus contains speech from 25 speakers (18 males and 7 females). 150 different prompt texts were read many times giving a total of 1,875 utterances. The vocabulary size is 220. The complete ChildEn corpus consists of 5,268 utterances, with an overall duration of 3h:28m:26s.

5.2. Non-native English recordings by German children

Recordings were made of 57 German children speaking English. Of these, 51 are the same as for the 'AIBO' and read German speech recordings (6.1), and two additional children are the same as for the read German speech recordings. The age range of the children is 10 - 13. In addition, 4 children aged 13-15 years were recorded. The recording settings are the same as for the German read speech. The vocabulary size is 920 words and 3.4 hours of speech were recorded. Originally it was intended that all partners should record the ITC-irst sentences and word lists (3.1). However, these children had only been learning English for half a year and had relatively poor English skills. Consequently only those ITC-irst word lists with words already known to the children were used, and these were supplemented with known texts from a textbook.

5.3. Non-native English recordings by Swedish children

Forty Swedish children were recorded reading the ITC-irst words and sentences (3.1). Each child read 44 words, 10 phonetically rich sentences and 10 'generic' sentences. In addition to text prompts, the children could listen to the same reference pronunciations used by ITC-irst. This option was used in 15% of the recordings. The recordings were made with a Shure SM10A head-mounted microphone, 16 kHz sampling frequency, and 16 bit resolution. The same external A/D converter was used as in the native Swedish recordings.

Table 2: *Non-native-language children's speech recordings in English (L1 denotes native language)*

L1	style	num chldren	ages	hours
Italian	imitated	53	9-11	2.1
Italian	read	25	9-11	1.3
Swedish	read	40	10-11	1.3
German	read	57	10-15	3.4

6. Native language spontaneous and emotional recordings - the 'AIBO' corpora

6.1. The German AIBO Corpus

Spontaneous and emotional German speech was collected from 51 children (age 10 - 13, 21 male, 30 female) from two schools using the methodology described as experiments 2 and 3 in (3.2). The goal was to create a new type of corpus of children's speech which is both 'natural' and emotional. The speech is intended to be 'natural', because children do not disguise their emotions to the same extent as adults, and 'spontaneous', because the children were not told to use specific instructions but were encouraged to talk to the AIBO like they would talk to a fiend.

The recordings took place in two class-rooms, one in each school. The only people in the room were the child, the supervisor, who gave instructions, the 'wizard' and a third assistant. Speech was recorded using a wireless head set (UT 14/20 TP SHURE UHF-series with microphone WH20TQG) and a DAT-recorder, with 48 kHz sampling rate, and 16bit quantisation. Each recording session lasted approximately 30 minutes. Because of the experimental design, the 25.5 hours of recordings contain a huge amount of silence (reaction time of the AIBO), which caused a substantial reduction after raw segmentation. Finally we obtained about 9.2 hours of speech.

The complete German AIBO database has been transcribed orthographically and the emotions/user-states have been labelled. The vocabulary size is 1,200 words. In addition, prosodic peculiarities such as 'hyper-articulation' and 'lengthening' were annotated. Where the pronunciation of a word differs from the standard pronunciation, an alternative is given after the word. There are a total of 13,791 utterances in the corpus.

6.2. The British-English AIBO Corpus

Thirty English children, aged 4-14, took part in AIBO "parcours" experiments 1 (3.2.1) and 2 (3.2.2). Recordings were made in a special multimedia studio in CETADL (the Centre for Educational Technology and Distance Learning) at the University of Birmingham. The recordings used two head-mounted wireless microphones: a UT 14/20 TP SHURE UHF-series with microphone WH20TQG, as used at the University of Erlangen, and a Senheiser ew100 range lapel microphone (SK100 transmitter, EK100 receiver), which was clipped to the SHURE head-mount. The Senheiser microphone has a digital wireless link, and gave a much cleaner signal than the SHURE microphone. Each child's speech was also recorded using existing wall-mounted microphones in CETADL. The speech was sampled at 16kHz using the Edirol UA-5 external sound card with USB interface. During each session, the outputs of the wall mounted microphone and three video cameras were also recorded on VHS video tape. The cameras showed the child's face, an aerial view of the "parcours", and a view of both the child and the "parcours". All of the data has been transcribed orthographically. Emotions/user-states have been labelled at the University of Erlangen.

The statistics of the PF_STAR spontaneous and emotional 'AIBO' children's speech recordings are summarised in table 3.

Table 3: *Native language 'AIBO' spontaneous and emotional children's speech recordings*

Language	style	num chldren	ages	hours
German	spontaneous	51	10-13	9.2
English	spontaneous	30	4-14	10

7. Conclusions

This paper has described the new PF_STAR corpus of recordings of British, German, Italian and Swedish children. The corpus includes recordings of read or imitated native-language speech from British, German and Swedish children aged 4-12, read or imitated non-native language English speech from 10 and 11 year old German, Italian and Swedish children, and spontaneous and emotional speech from British and German children aged between 4 and 14. The majority of the non-native English recordings are based on a common set of English words and phrases designed by ITC-irst, and these were also recorded by the British children. Both sets of recordings of spontaneous emotional children's speech (German and English) used the same 'AIBO' methodology from the University of Erlangen. In total, the corpus includes nearly 65 hours of recordings of speech from 611 children. Experiments in which the PF_STAR corpus has been used to establish performance baselines for recognition of children's speech in British English, German, Italian and Swedish are reported elsewhere.

8. References

[1] Mostow, J., Roth, S.F., Hauptmann, A.G. & Kane, M. "A prototype reading coach that listens", Proc. 12th National Conference on Artificial Intelligence (AAAI'94), Seattle, WA, pp 785-792, 1994.

[2] Wilpon Jay G., Jacobsen Claus N., "A Study of Speech Recognition for Children and the Elderly", Proc. ICASSP'96, Vol. 1, pp. 349-352, 1996.

[3] Lee, S., Potamianos, A. & Narayanan, S., "Acoustics of children's speech: Developmental changes of temporal and spectral parameters," Journal of the Acoustical Society of America, pp. 1455-1468, Mar. 1999.

[4] Li, Q. & Russell, M., "An Analysis of the Causes of Increased Error Rates in Children's Speech Recognition", Proc. ICSLP 2002.

[5] Narayanan, S. & Potamianos, A., "Creating conversational interfaces for children", IEEE Trans. Speech and Audio Processing, vol. 10, pp. 65-78, Feb. 2002.

[6] Batliner, A., Hacker, C., Steidl, S., Nöth, E., D'Arcy, S., Russell, M. & Wong, M. "'You stupid tin box' - children interacting with the AIBO robot: A cross-linguistic emotional speech corpus", Proc. LREC 2004, pp 171-174, 2004