A Multimodal Educational Game for 3-10-Year-Old Children: Collecting and Automatically Recognising European Portuguese Children's Speech

Annika Hämäläinen[1,2], *Fernando Miguel Pinto*[1], *Silvia Rodrigues*[1], *Ana Júdice*[1],
Sandra Morgado Silva[3], *António Calado*[1], *Miguel Sales Dias*[1,2]

[1]Microsoft Language Development Center, Lisbon, Portugal
[2]ADETTI – ISCTE, IUL, Lisbon, Portugal
[3]Diferente Jogo, Caldas da Rainha, Portugal

{t-anhama, a-fpinto, v-antonc, Miguel.Dias}@microsoft.com, sandrasilva@diferencas.net

Abstract

Speech interfaces have tremendous potential in education. In this paper, we present our work in the Contents for Next Generation Networks project, an ongoing Portuguese industry-academia collaboration developing a multimodal educational game aimed at improving the physical coordination and the basic mathematical and musical skills of 3-10-year-old children. We focus on our work in the area of children's speech recognition: designing, collecting, transcribing and annotating a 21-hour corpus of prompted European Portuguese children's speech, as well as our first experiments with different acoustic modelling approaches. Our speech recognition results suggest that training children's speech models from scratch is a more promising approach than retraining adult speech models using children's speech when a sufficient amount of training data is available from the targeted age group. This finding also holds for adult female speech models retrained using children's speech. As compared with a baseline recogniser comprising gender-dependent adult speech models, the best-performing children's speech models that we have trained so far – gender-independent cross-word triphones trained with 17.5 hours of speech from 3-10-year-old children – resulted in a 45-percent (relative) decrease in word error rate in a task expecting isolated cardinal numbers, sequences of cardinal numbers or musical notes as speech input.

Index Terms: acoustic modelling, ASR, child-computer interaction, corpus, educational game, European Portuguese

1. Introduction

Speech interfaces have tremendous potential in the education of children. Speech provides a natural modality for child-computer interaction and can, at its best, contribute to a fun, motivating and engaging way of learning [1]. However, it is well known that automatically recognising children's speech is a very challenging task. Recognisers trained on adults' speech tend to suffer from a substantial deterioration in recognition performance when used by children [1-6]. Moreover, word error rates (WERs) on children's speech are usually much higher than those on adults' speech even when using a recogniser trained on age-specific speech – although they do show a gradual decrease as the children get older [1-7].

The difficulty of automatically recognising children's speech can be attributed to it being acoustically and linguistically very different from adults' speech [1, 2]. For instance, due to their vocal tracts being smaller, the fundamental and formant frequencies of children's speech are higher [1, 2, 7-9]. What is particularly characteristic of children's speech is its higher variability as compared with adults' speech, both within and across speakers [1, 2]. This variability is caused by rapid developmental changes in their anatomy, speech production et cetera, and manifests itself, for example, in speech rate, the degree of spontaneity, the frequency of disfluencies, fundamental and formant frequencies, and pronunciation quality [1, 2, 7-11].

In the context of education, speech interfaces have been developed, for instance, for interactive reading and pronunciation tutoring [1, 2, 12-14]. However, in the case of less-spoken languages, sufficiently large corpora of children's speech are often not available for developing speech-driven educational applications. This has, for instance, been the case for European Portuguese (pt-PT). In this paper, we describe our speech-related work in the Contents for Next Generation Networks (CNG) project, whose end product is a multimodal educational game for 3-10-year-old Portuguese children: developing of a 21-hour corpus of prompted children's speech, and modelling children's speech for automatic speech recognition (ASR) purposes.

The paper is further organised as follows. In Section 2, we introduce the CNG project and the educational game that the project partners are developing. We present the design, collection, transcription and annotation of the CNG Corpus of European Portuguese Children's Speech in Section 3 and, in Section 4, describe the children's speech recognition experiments that we have carried out so far and intend to perform in the future. Finally, in Section 5, we formulate our conclusions.

2. The CNG project and educational game

The CNG project is an ongoing Portuguese industry-academia collaboration that studies speech and gesture as natural alternatives for child-computer interaction in the education of 3-10-year-old children. To this end, the project partners are developing a multimodal educational game that can be played in an immersive virtual environment (a CAVE, [15]) or with a desktop computer using Kinect for Windows, a motion sensing input device by Microsoft [16]. The game addresses two main areas of development: motor skill development (physical coordination skills) and cognitive development (attention, problem solving, mathematical and musical skills). In the game, the children will use their voice and gestures to complete different kinds of puzzles and tasks in 3D and 2D scenarios with educational themes (e.g. the Age of Discovery, dinosaurs). An intelligent virtual assistant with a synthesised voice will guide and help them throughout the game. In one of the themed scenarios, the Age of Discovery, the assistant might, for instance, say, "*Cheer up the sailors by counting from one to* *five and moving your body. With each number you say, you will need to move a part of your body.*" Speech input will be enabled for tasks related to mathematics (e.g. counting objects, simple mathematical operations) and music (e.g. completing musical note sequences). The expected speech input includes isolated cardinals, sequences of cardinals, and musical notes. The difficulty level of the game can be set up manually before the game starts but will also be adjusted automatically based on the children's performance.

The CNG game is developed for pt-PT. To the best of our knowledge, the only other speech-driven educational application for pt-PT is a speech therapy system that uses games to identify the phones that 5-6-year-old children have problem pronouncing, and to help them overcome their pronunciation problems [17]. For developing that system, a 158-minute corpus of isolated words was collected from 111 children belonging to the targeted age group. pt-PT children's speech is also available in the Portuguese Speecon Database [18]. However, it only contains speech from 52 children, of which only 15 children aged 8-10 belong to the age group targeted in the CNG project; the rest are 11-14-year olds.

3. The CNG Corpus

When it comes to speech material, the goal of the CNG project was to develop a corpus of about 20 hours of children's speech suitable for training and testing acoustic models (AMs) for the speech-driven parts of the CNG game. The resulting corpus is called the CNG Corpus of European Portuguese Children's Speech. The following subsections describe the design, collection, transcription and annotation of the corpus, as well as the details of the full corpus and the datasets used for the ASR experiments.

3.1. Corpus design

3.1.1. Speaker selection

As the CNG game is aimed at 3-10-year-old children, we only collected speech from speakers in that age range. Based on the children's capabilities (see Section 3.1.2), we split them into two age groups that are considered homogenous populations for the purposes of the CNG project: 3-6-year-old and 7-10-year-old children.

We collected speech from children attending nurseries and schools in and around the Portuguese cities of Lisbon, Leiria and Aveiro. The cities were chosen for practical reasons; the time and budget available for the data collection campaign were tight, so we had to concentrate on places where the project partners had existing contacts at nurseries and schools and were able to easily attend recording sessions. We tried to keep the ratio of girls and boys as even as possible but were not, for instance, able to aim at a specific ratio of speakers from the different areas.

3.1.2. Prompt design

Collecting speech from children poses some special challenges. First, children's attention span depends on their age [19]; they might get distracted from a prolonged recording task. Second, they may have difficulty reading or repeating long, complex words or sentences. Taking these challenges and the requirements of the CNG game into account, we designed four types of prompts to record: 292 phonetically rich sentences, musical notes (e.g. *dó*), isolated cardinals (e.g. *44*), and sequences of cardinals (e.g. *28, 29, 30, 31*). The phonetically rich sentences originated from the CETEMPúblico corpus of Portuguese newspaper language [20]. They were short (~4 words/sentence) and did not include any difficult words. In the case of 3-6-year-olds, the cardinals ranged from 0 to 30, and the sequences of cardinals consisted of 2-3 numbers. In case of 7-10-year-olds, the cardinals ranged from 0 to 999, and the sequences of cardinals consisted of 4 numbers.

The younger children produced a set of 30 prompts selected across the different types of prompts in a balanced way. This resulted in a bit more than one minute of speech per speaker. The older children read out a set of 50 prompts resulting in about 3 minutes of speech per speaker.

The differences in the contents and targeted number of prompts between the two age groups were designed based on our experiences from pilot recording sessions with children of different ages. They take into account the differences in the attention span and linguistic capabilities between the two age groups.

3.2. Data collection

We used the *Your Speech* online speech data collection platform [21] for collecting the speech data and some biographical information (age group, gender and region of origin) about the speakers. The platform was operated by recording supervisors trained for managing the recording sessions. The platform's web interface presented the speakers with each of the prompts to record; the recording of an utterance started when the recording supervisor clicked the *Record* button and ended when (s)he clicked the *Stop* button, or when no more speech was detected by the system. The recorded utterance was then uploaded to the web backend of the system and automatically checked for the presence of speech and clipping; if speech was indeed detected and if the utterance did not contain any samples of clipping, the recording supervisor could proceed to the next prompt using the *Next Phrase* button or, if unhappy with the utterance, have the speaker rerecord the utterance using the *Rerecord* button. If the automatic quality control was not passed, the speaker was requested to rerecord the utterance.

The recording sessions took place in a quiet room. In the case of 3-6-year-olds, as well as the 7-10-year-olds that had problems reading the prompts, the recording supervisors read the prompts out first and the children then repeated them. The speech data were recorded using a noise-cancelling Life Chat LX 3000 USB headset and digitised at 16 bits and 22 kHz.

3.3. Transcriptions, annotations and quality control

Using an in-house transcription tool, a (single) native speaker trained for the task transcribed the corpus orthographically. In addition, she annotated the corpus using the tags listed in Table 1. The annotation scheme was designed to be compatible with the requirements of our in-house AM training tool.

The transcriber discarded sessions that contained recordings with consistently poor audio quality, or speech from non-native speakers or speakers with consistent problems repeating or reading the prompts out. After the transcription and annotation work, we identified typographical errors in the transcriptions by checking them against a large Portuguese lexicon. In addition, we used forced alignment to identify potentially problematic utterances; utterances that cannot be aligned are more likely to have problems in the quality of the speech, the transcriptions and/or the audio. We cast these utterances aside. As the proportion of utterances annotated with the <NPS/> tag was very low, we did not include them in the final version of the corpus, either. We have not carried out formal, systematic assessment of the reliability of the transcriptions and annotations included in the corpus.

3.4. Overview of the corpus and the datasets for ASR experiments

We collected a total of 21 hours of speech from 510 children – 30% of them aged 3-6, and 70% aged 7-10. This makes the CNG Corpus the largest currently available corpus of pt-PT children's speech, also covering some ages that other resources of pt-PT children's speech [17, 18] do not cover. The imbalance between the amount of speech collected from 3-6-year-olds and 7-10-year-olds is due to the difficulty of recording speech from very young speakers, as well as the limited amount of speech collected from each of them. 56% of the speakers in the corpus are girls and 44% boys. The vast majority (84%) are from the Leiria area.

For the purpose of ASR experiments, the corpus was randomly divided into three speaker-independent datasets that respect the proportions of speaker age and gender in the full corpus: a training set used for training AMs (85% of the data), a development test set for optimisation purposes (5% of the data), and an evaluation test set for the final testing of the AMs (10% of the data). Due to the limited number and type of prompts recorded, it was not possible to rule out the same prompts appearing in the three different datasets; with hindsight, this could have been avoided by using a corpus design that splits the speakers and prompts between the three datasets before the recordings. The main statistics of the corpus and the three datasets are presented in Table 2.

Table 1. *The tags used for annotating the recordings.*

Tag	Meaning
<FILL/>	Filled pauses (e.g. "umm", "er", "ah")
<NON/>	Non-human noises (e.g. mouse clicks, music)
<SPN/>	Human noises (e.g. coughs, audible breath)
<UNKNOWN/>	False starts; mispronounced, unintelligible or truncated words; words with considerable background noise
<NPS/>	Speech from non-primary speakers

Table 2. *The main statistics of the speech material.*

	Train	Devel.	Eval.	Total
#Speakers	432	26	52	510
#Word types	605	482	521	614
Ages 3-6	*557*	*218*	*319*	*560*
Ages 7-10	*585*	*458*	*494*	*591*
#Word tokens	102,537	6229	12,029	121,046
Ages 3-6	*9553*	*676*	*1148*	*11,424*
Ages 7-10	*92,984*	*5553*	*10,881*	*109,622*
hh:mm:ss	17:42:22	01:06:26	02:05:34	20:54:22
Ages 3-6	*02:30:24*	*00:10:22*	*00:18:31*	*02:59:17*
Ages 7-10	*15:11:58*	*00:56:04*	*01:47:03*	*17:55:05*

4. ASR experiments

The ASR functionality of the CNG game is implemented using the Microsoft Speech Platform Runtime (Version 11) [22], which contains a Hidden Markov Model (HMM) -based speech recogniser, and a language pack, which incorporates the language-specific components necessary for ASR: grammars, a pronunciation lexicon, and AMs (cf. [23]). The goals of the ASR-related work in the project are to create a pt-PT language pack that is specifically adapted to the CNG game, and to obtain the best possible recognition performance using existing techniques and tools compatible with the requirements of the Microsoft Speech Platform Runtime. The following subsections provide information about the grammars that we have authored for the CNG game, the pronunciation lexicon that we are using, as well as our first experiments with different acoustic modelling approaches.

4.1. Grammars and pronunciation lexicon

We have authored several grammars for language modelling purposes: a list grammar for the musical notes and structure grammars for the isolated cardinals and the cardinal sequences. The grammar for the isolated cardinals allows cardinals from 0 to 999, whereas the grammar for the cardinal sequences allows sequences of 2-4 cardinals ranging from 0 to 999. In the experimentation phase, the phonetically rich sentences are simply recognised using a list grammar consisting of the 292 prompts (see Section 3.1.2); the CNG game itself will not include this type of speech input. Our pronunciation lexicon contains an average of 1.04 pronunciations for the words in the task, represented using a set of 38 phone labels.

4.2. Feature extraction

We carried out feature extraction of the children's speech data at a frame rate of 10 ms using a 25-ms Hamming window and a pre-emphasis factor of 0.98. We calculated 12 Mel Frequency Cepstral Coefficients (MFCCs) and log-energy with corresponding first, second and third order time derivatives, and reduced the total number of features to 36 using Heteroscedastic Linear Discriminant Analysis (HLDA).

4.3. Acoustic modelling

The following subsections describe our baseline recogniser, as well as the different recognisers built so far to test different acoustic modelling approaches. We omit detailed information about the baseline recogniser and our training techniques as commercially sensitive information.

4.3.1. Baseline recogniser

For establishing a baseline, we used the acoustic models from the pt-PT language pack [23]. They comprise a mix of gender-dependent (GD) whole-word models and cross-word triphones trained using several hundred hours of read and spontaneous speech collected from adult speakers. The baseline recogniser also includes a silence model, a hesitation model for modelling filled pauses, and a noise model for modelling human and non-human noises. In addition to recognising children's speech using the female and male AMs of the baseline recogniser in parallel (BL), we also recognised it using the female AMs (BLF) and the male AMs (BLM) separately.

4.3.2. Experimental recognisers

We trained and tested several different kinds of AMs to investigate the effect of different variables on speech recognition performance: the type of training technique, and the gender and age group of the children. The two main types of AMs were 1) the BL AMs retrained with the children's speech in the training set and 2) cross-word triphones trained using the children's speech only. The experimental AMs included GD AMs, gender-independent (GI)

AMs, and age group -dependent AMs. Due to the limited amount of training data from 3-6-year-olds, we were only able to train age group -dependent AMs for 7-10-year-olds with the types of training techniques used in our experiments so far.

We retrained the BL AMs using several different set-ups:

- BL$_{F}$ and BL$_{M}$ retrained with the training data from girls and boys, respectively (BL$_{GD}$)
- BL$_{F}$ retrained with the training data from both girls and boys (BL$_{GI}$)
- BL$_{F}$ and BL$_{M}$ retrained with the training data from 7-10-year-old girls and boys, respectively (BL$_{GD7-10}$)
- BL$_{F}$ retrained with the training data from 7-10-year-old girls and boys (BL$_{GI7-10}$)

The hesitation and noise models of the baseline recogniser were retrained utilising the <FILL/>, <NON/> and <SPN/> tags available in the CNG Corpus. The idea of only retraining the adult female AMs (BL$_{F}$) with children's speech stems from the fact that the acoustic characteristics of children's speech are more similar to adult female speech than to adult male speech [7-9, 24]. In fact, in the ASR experiments reported in [7], the WER obtained for children's speech using adult male AMs is more than twice higher than the WER achieved using adult female AMs. Furthermore, [7] reports better recognition performance with adult female models adapted with children's speech using maximum likelihood linear regression (MLLR) adaptation than with adult male and GI models adapted the same way.

We used a standard procedure with decision tree state tying (see e.g. [25]) to train several different kinds of cross-word triphone recognisers using children's speech only:

- GD triphones trained with the full training set (CNG$_{GD}$)
- GI triphones trained with the full training set (CNG$_{GI}$)
- GD triphones trained with the training data from 7-10-year-olds (CNG$_{GD7-10}$)
- GI triphones trained with the training data from 7-10-year-olds (CNG$_{GI7-10}$)

Similarly to the baseline recogniser and its derivatives, the cross-word triphone recognisers also included a silence model, a hesitation model and a noise model – the last two trained utilising the <FILL/>, <NON/> and <SPN/> tags available in the CNG Corpus. To find the optimal number of Gaussian mixtures per state, we trained and tested triphone recognisers with 4-16 Gaussian mixtures per state.

Table 3. *WERs (%) with a 95% confidence interval for all, for 3-6-year-old, and for 7-10-year-old speakers in the evaluation test set.*

	All Eval.	Ages 3-6	Ages 7-10
BL	18.1 ± 0.7	49.2 ± 3.0	14.9 ± 0.7
BL$_{F}$	18.1 ± 0.7	49.2 ± 3.0	14.9 ± 0.7
BL$_{M}$	32.8 ± 0.9	69.9 ± 2.7	28.9 ± 0.9
BL$_{GD}$	11.9 ± 0.6	29.4 ± 2.7	10.1 ± 0.6
BL$_{GI}$	11.5 ± 0.6	27.8 ± 2.6	9.8 ± 0.6
BL$_{GD7-10}$	13.1 ± 0.6	37.8 ± 2.9	10.5 ± 0.6
BL$_{GI7-10}$	13.0 ± 0.6	36.8 ± 2.8	10.5 ± 0.6
CNG$_{GD}$	10.5 ± 0.6	30.9 ± 2.7	8.3 ± 0.5
CNG$_{GI}$	**10.0 ± 0.5**	**27.1 ± 2.6**	**8.2 ± 0.5**
CNG$_{GD7-10}$	12.1 ± 0.6	36.5 ± 2.8	9.4 ± 0.6
CNG$_{GI7-10}$	11.5 ± 0.6	35.0 ± 2.8	9.1 ± 0.6

Table 4. *The WERs (%) of the best-performing recogniser (CNG$_{GI}$) per prompt type.*

	All Eval.	Ages 3-6	Ages 7-10
Phonetically rich	10.4	25.6	6.6
Musical notes	4.2	13.3	2.2
Isolated cardinals	6.3	27.4	3.9
Sequences of cardinals	10.6	33.3	9.7
Overall (excl. phon. rich)	9.8	29.3	8.7

4.4. Speech recognition results

Table 3 reports the WERs for the most relevant/best-performing recognisers: the baseline recognisers (BL, BL$_{F}$ and BL$_{M}$), the baseline recognisers retrained with children's speech (BL$_{GD}$, BL$_{GI}$, BL$_{GD7-10}$ and BL$_{GI7-10}$), 14-Gaussian triphone recognisers trained using the training data from both 3-6-year-olds and 7-10-year-olds (CNG$_{GD}$ and CNG$_{GI}$), and 12-Gaussian triphone recognisers trained using the training data from 7-10-year-olds only (CNG$_{GD7-10}$ and CNG$_{GI7-10}$).

All models that had specifically been adapted for children's speech significantly outperformed the baseline recogniser comprising GD AMs trained using adult speech (BL). The error rates obtained with the adult female AMs (BL$_{F}$) were identical to those obtained with the combination of the adult female and adult male AMs (BL). In other words, the adult female AMs were always chosen to recognise children's speech. This further illustrates that children's speech is more similar to adult female than to adult male speech. Recognition performance with adult male AMs (BL$_{M}$) was significantly worse. Similar to other studies (e.g. [3-5, 7]), the WERs were considerably higher in the case of the younger children. The best-performing recogniser for both age groups was the GI cross-word triphone recogniser trained using all children's speech in the training set (CNG$_{GI}$). Its performance did not significantly differ from that of the corresponding GD recogniser (CNG$_{GD}$), however. Overall, its performance was 45% (relative) better than that of the GD baseline recogniser (BL). The improvement was 45% also when calculated separately for both 3-6-year-olds and 7-10-year-olds.

In the case of 7-10-year-olds, training cross-word triphones using children's speech led to significantly better recognition performance than retraining the otherwise similar baseline AMs with children's speech (CNG$_{GD}$ vs. BL$_{GD}$, CNG$_{GI}$ vs. BL$_{GI}$ etc.). This was the case also when retraining the adult female AMs with children's speech (CNG$_{GI}$ vs. BL$_{GI}$). However, in the case of 3-6-year-olds, there were no significant differences between the two types of training techniques. For instance, in the case of 7-10-year-olds, the performance of the best-performing cross-word triphone recogniser (CNG$_{GI}$) was 16% (relative) better than that of the best-performing recogniser comprising adult female AMs that had been updated with speech collected from both girls and boys (BL$_{GI}$). The improvement was only 3% (relative; not significant) in the case of 3-6-year-olds. These findings might be related to the fact that we had much less training data from 3-6-year-olds than from 7-10-year-olds, and suggest that training children's speech models from scratch is a more promising approach than retraining adult speech models using children's speech especially when a sufficient amount of training data is available from the targeted age group.

The WERs of the experimental recognisers also illustrate that GD AMs do not lead to improved recognition performance in the case of 3-10-year-old children; the performance of otherwise similar GD and GI recognisers (BL$_{GD}$ vs. BL$_{GI}$, BL$_{GD7-10}$ vs. BL$_{GI7-10}$

etc.) did not differ from each other significantly. Although the effect was not significant, the GI recognisers did seem to have a tendency for better recognition performance than the GD recognisers. This finding could be expected based on the fact that the differences in the fundamental and formant frequencies of girls' and boys' speech only become more pronounced at around 11 years of age [7, 9].

Recognition performance deteriorated in the case of 7-10-year-olds when we only used training data from their own age group to retrain or train the AMs, although the effect was only significant in the case of the cross-word triphone recognisers (CNG$_{\text{GD}}$ vs. CNG$_{\text{GD7-10}}$ and CNG$_{\text{GI}}$ vs. CNG$_{\text{GI7-10}}$). Even though training AMs with data from the targeted age or age group might be expected to lead to improved recognition performance [1, 2], this result supports the view that a broad diversity of the training data aids recognition performance [4]. Unsurprisingly, the WERs of 3-6-year-olds increased significantly with AMs that had been retrained or trained from scratch using the training data from 7-10-year-olds only (BL$_{\text{GD}}$ vs. BL$_{\text{GD7-10}}$, BL$_{\text{GI}}$ vs. BL$_{\text{GI7-10}}$ etc.).

Table 4 lists the WERs of the best-performing children's speech recogniser (CNG$_{\text{GI}}$) for each of the recorded prompt types. It also includes the overall WERs without phonetically rich sentences, which represent a prompt type that is not relevant for the CNG game. It is clear that the recognition performance of 3-6-year-olds leaves much to be desired. While the recognition performance of the different types of prompts also leave space for improvement in the case of 7-10-year-olds, it is probably already acceptable for the CNG game – in particular in the case of musical notes and isolated cardinals.

4.5. Discussion and future work

For now, the best-performing AMs (CNG$_{\text{GI}}$) have been delivered for use in the first versions of the CNG game, together with the grammars and pronunciation lexicon discussed in Section 4.1. However, we will continue to explore ways to optimise recognition performance.

We are particularly interested in improving on the poor recognition performance of 3-6-year-olds. There are probably at least three reasons for this poor performance. First, as already mentioned earlier, recognition performance correlates with children's age; regardless of the optimisation methods that we will use, the WERs on older children's speech are likely to remain lower than those on younger children's speech. Second, we had much less training data from 3-6-year-olds than from 7-10-year-olds; the AMs we trained using speech from speakers belonging to both age groups were effectively optimised for 7-10-year-olds. Third, many of the 3-6-year-olds recorded for the corpus had difficulty repeating the prompts correctly, especially in the case of prompts containing several words to memorise, and this resulted in a lot of disfluencies and hesitations in the data. In [10], recognition performance did not suffer significantly from disfluencies and hesitations in children's speech. However, we must investigate if this is also the case with our speech data, which also contains speech from speakers younger than those tested by [10]. In addition to the difficulty repeating the prompts correctly, the younger children were very challenging to record speech from because they often reacted to the recording situation with shyness (see also [17]). For all of these reasons, it might be interesting to collect more speech by recording children's verbal interaction with the CNG game itself, and to use those data to adapt the current children's speech AMs. Firstly, the recording situation would be much less intimidating. Secondly, the produced speech would probably be more suitable for training AMs for the CNG game than the speech collected during the data collection campaign: in the case of children, vowel durations are known to be significantly higher and speaking rate lower for read speech than for spontaneous speech, the effect being more pronounced in younger children [1]. Another benefit of collecting more speech is that we could offset the current bias towards regional accents from the Leiria area where most of the speech data was collected (see Section 3.4).

Before embarking on a speech data collection using a preliminary version of the CNG game, however, we will experiment with existing optimisation methods to try to improve on the recognition performance of both 3-6-year-olds and 7-10-year-olds. Some of the factors making children's speech recognition particularly challenging are the higher frequencies of their fundamental and formant frequencies and the high level of variability in these frequencies across children of different ages. Vocal Tract Length Normalisation (VTLN) has been shown to lead to improved recognition performance on children's speech both in the case of AMs trained using speech from another age group [6, 7, 26] and in the case of AMs trained using speech from the same age group [6, 7, 26]. We are currently looking into applying VTLN at both the training and the recognition stages of our experiments (cf. [26]).

Young children may have problems accurately producing particular speech sounds or clusters of speech sounds; they might, for instance, systematically substitute one consonant cluster for another [2]. This is likely to have a negative effect on speech recognition performance [2]. When there is not enough training data for acoustic models to learn such pronunciation patterns, a pronunciation lexicon customised for the pronunciation patterns of the targeted age group might improve recognition performance [1]. For instance, [27] obtained significant decreases in WER in the case of preschool children by studying how they pronounce words with respect to their canonical pronunciations, and by deriving pronunciation rules to add relevant pronunciation variants into their pronunciation lexicon. We are currently studying the pronunciation patterns of the 3-6-year-olds in the CNG corpus, with the goal of incorporating this information into our pronunciation lexicon.

In the case of adults, using restricted grammars may be a good strategy for modelling the kind of structured input (e.g. cardinal numbers) also expected by the CNG game. However, children might not restrict their responses to the utterances vital to the task only; they might, for instance, use utterances expressing excitement or disappointment, or try to interact with the characters on the screen in a way that is not related to the task at hand [10, 28]. The failure to model this kind of extraneous speech is likely to lead to deteriorated recognition performance. Therefore, we intend to analyse children's interaction with a preliminary version of the CNG game to determine the best language modelling approach for the final application.

5. Conclusions

In this paper, we presented the design, collection, transcription and annotation of the largest currently available European Portuguese children's speech corpus, which contains 21 hours of prompted speech recorded from 510 children aged 3-10. The corpus comes with manual orthographic transcriptions and annotations indicating filled pauses, noises and damaged words (e.g. mispronunciations). It was specifically designed for training and testing acoustic models for an educational game that teaches 3-10-year-old children basic mathematical and musical skills. However, it could also prove useful for developing other speech-driven applications for children and for use in children's speech research. The corpus is available at request for R&D activities. Please contact Miguel Sales Dias (Miguel.Dias@microsoft.com) for further information.

In addition, we presented our first ASR experiments, aimed at finding the best acoustic modelling approach for the aforementioned educational game. Our speech recognition results suggest that, when a sufficient amount of training data is available from the targeted age group, training children's speech models from scratch is a more promising approach than retraining adult speech models using children's speech. This finding also holds for adult female models retrained using children's speech. Our recognition results also show that gender-dependent models do not lead to increased recognition performance in the case of 3-10-year-old children. As compared with a baseline recogniser comprising gender-dependent adult speech models, the best-performing children's speech models that we have trained so far – gender-independent cross-word triphones trained with 17.5 hours of speech from 3-10-year-old children – result in a 45-percent (relative) decrease in word error rate in a task expecting isolated cardinal numbers, sequences of cardinal numbers or musical notes as speech input.

6. Acknowledgements

The QREN 7943 CNG – Contents for Next Generation Networks project is co-funded by Microsoft, the Portuguese Government, and the European Structural Funds for Portugal (FEDER), through COMPETE and QREN. The authors are indebted to the children, recording supervisors, nurseries and schools that took part in the data collection.

7. References

[1] Gerosa, M., Giuliani, D., Narayanan, S. and Potamianos, A., "A Review of ASR Technologies for Children's Speech", in *Proc. WOCCI*, Cambridge, MA, USA, 2009.

[2] Russell, M. and D'Arcy, S., "Challenges for Computer Recognition of Children's Speech", in *Proc. SLaTE*, Farmington, PA, USA, 2007.

[3] Potamianos, A. and Narayanan, S., "Robust Recognition of Children's Speech," *IEEE Speech Audio Process.*, 11(6):603-615, 2003.

[4] Wilpon, J. G. and Jacobsen, C. N., "A Study of Speech Recognition for Children and the Elderly", in *Proc. ICASSP*, Atlanta, GA, USA, 1996.

[5] Elenius, D. and Blomberg, M., "Adaptation and Normalization Experiments in Speech Recognition for 4 to 8 Year Old Children", in *Proc. Interspeech*, Lisbon, Portugal, 2005.

[6] Gerosa, M., Giuliani, D. and Brugnara, F., "Speaker Adaptive Acoustic Modeling with Mixture of Adult and Children's Speech", in *Proc. Interspeech*, Lisbon, Portugal, 2005.

[7] Gerosa, M., Giuliani, D. and Brugnara, F., "Acoustic Variability and Automatic Recognition of Children's Speech", *Speech Commun.*, 49(10-11):847-860, 2007.

[8] Huber, J.E., Stathopoulos, E. T., Curione, G. M., Ash, T. A. and Johnson, K., "Formants of Children, Women and Men: The Effects of Vocal Intensity Variation", *J. Acoust. Soc. Am.*, 106(3):1532-1542, 1999.

[9] Lee, S., Potamianos, A. and Narayanan, S., "Acoustics of Children's Speech: Developmental Changes of Temporal and Spectral Parameters", *J. Acoust. Soc. Am.*, 10:1455-1468, 1999.

[10] Narayanan, S. and Potamianos, A., "Creating Conversational Interfaces for Children", *IEEE Speech Audio Process.*, 10(2):65-78, 2002.

[11] Eguchi, S. and Hirsh, I. J., "Development of Speech Sounds in Children", *Acta Otolaryngol. Suppl.*, 257:1-51, 1969.

[12] Beck, J., Jia, P. and Mostow, J., "Automatically Assessing Oral Reading Fluency in a Computer Tutor that Listens", *TICL*, 1:61-81, 2004.

[13] Hagen, A., Pellom, B., Vuuren, S. V. and Cole, R., "Advances in Children's Speech Recognition within an Interactive Literacy Tutor", in *Proc. HLT/NAACL*, Boston, MA, USA, 2004.

[14] Russell, M. J., Series, R. W., Wallace, J. L., Brown, C. and Skilling, A., "The STAR System: An Interactive Pronunciation Tutor for Young Children", *Comput. Speech Lang.*, 14(2):161-175, 2000.

[15] Soares, L. P., Pires, F., Varela, R., Bastos, R., Carvalho, N., Gaspar, F. and Sales Dias, M., "Designing a Highly Immersive Interactive Environment: The Virtual Mine", *Comput. Graph. Forum*, 29(6):1756-1769, 2010.

[16] Kinect for Windows. Online: http://www.microsoft.com/en-us/kinectforwindows/, accessed on 4 Apr 2013.

[17] Lopes, C., Veiga, A. and Perdigão, F., "A European Portuguese Children Speech Database for Computer Aided Speech Therapy", in *Proc. PROPOR*, Coimbra, Portugal, 2012.

[18] The Portuguese Speecon Database. Online: http://catalog.elra.info/product_info.php?products_id=798, accessed 4 Apr 2013.

[19] Unger, H. G., *Encyclopedia of American Education*, Third Edition, Facts on File Inc., New York, USA, 2007.

[20] CETEMPúblico. Online: http://www.linguateca.pt/cetempublico/, accessed 4 Apr 2013.

[21] Freitas, J., Calado, A., Braga, D., Silva, P. and Sales Dias, M., "Crowd-Sourcing Platform for Large-Scale Speech Data Collection", in *Proc. FALA*, Vigo, Spain, 2010.

[22] Microsoft Speech Platform Runtime (Version 11). Online: http://www.microsoft.com/en-us/download/details.aspx?id=27225, accessed 4 Apr 2013.

[23] Microsoft Speech Platform Runtime Languages. Online: http://www.microsoft.com/en-us/download/details.aspx?id=27224, accessed 4 Apr 2013.

[24] Caldas de Oliveira, L., "eCIRCUS: Children Voices Against Bullying in Schools", keynote at *1st Microsoft Workshop on Speech Technology - Building Bridges between Industry and Academia,* Porto Salvo, Portugal, 2007. Online: http://www.microsoft.com/pt-pt/mldc/news/mldcworkshop.aspx, accessed 8 Apr 2013.

[25] Young, S., Evermann, G., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V. and Woodland, P., *The HTK Book (for HTK Version 3.2.1)*, Cambridge University, Cambridge, UK, 2002.

[26] Giuliani, D. and Gerosa, M., "Investigating Recognition of Children's Speech", in *Proc. ICASSP*, Hong Kong, 2003.

[27] Cincarek, T., Shindo, I., Toda, T., Saruwatari, H. and Shikano, K., "Development of Preschool Children Subsystem for ASR and Q&A in a Real-Environment Speech-Oriented Guidance Task", in *Proc. Interspeech*, Antwerp, Belgium, 2007.

[28] Strommen, E. F. and Frome F. S., "Talking Back to Big Bird: Preschool Users and a Simple Speech Recognition System", *Educ. Technol. Res. Dev.*, 41(1):5-16, 1993.