Child Speech Recognition in Human-Robot Interaction: Evaluations and Recommendations

James Kennedy*
Plymouth University, U.K.

Séverin Lemaignan
Plymouth University, U.K.

Caroline Montassier
INSA Rouen, France

Pauline Lavalade
Université Pierre et Marie Curie, France

Bahar Irfan
Plymouth University, U.K.

Fotios Papadopoulos
Plymouth University, U.K.

Emmanuel Senft
Plymouth University, U.K.

Tony Belpaeme
Plymouth University, U.K.
Ghent University, Belgium

ABSTRACT

An increasing number of human-robot interaction (HRI) studies are now taking place in applied settings with children. These interactions often hinge on verbal interaction to effectively achieve their goals. Great advances have been made in adult speech recognition and it is often assumed that these advances will carry over to the HRI domain and to interactions with children. In this paper, we evaluate a number of automatic speech recognition (ASR) engines under a variety of conditions, inspired by real-world social HRI conditions. Using the data collected we demonstrate that there is still much work to be done in ASR for child speech, with interactions relying solely on this modality still out of reach. However, we also make recommendations for child-robot interaction design in order to maximise the capability that does currently exist.

Keywords

Child-Robot Interaction; Automatic Speech Recognition; Verbal Interaction; Interaction Design Recommendations

1. INTRODUCTION

Child-robot interaction is moving out of lab and into 'the wild', contributing to domains such as health-care [2], education [15, 25], and entertainment [20]. An increasing amount is being understood about how to design interactions from a nonverbal behaviour perspective [13, 14], but many of these domains hinge on effective verbal communication. This includes not only appropriate speech production by robots, but transcribing and understanding speech from young users as well. A prerequisite to this interpretation of speech is having a sufficiently accurate transcription of what is being said.

*Corresponding author: james.kennedy@plymouth.ac.uk

HRI '17, March 06 - 09, 2017, Vienna, Austria

ISBN 978-1-4503-4336-7/17/03…$15.00

DOI: http://dx.doi.org/10.1145/2909824.3020229

For this reason, high-quality Automatic Speech Recognition (ASR) is a vital component for producing autonomous human-robot interaction. ASR engines have undergone significant improvements in recent years, particularly following the introduction of new techniques such as deep learning [26]. However, these engines are commonly evaluated against standardised datasets of adult speech [23]. One might naively assume that these improvements will also translate to child speech, and will cope relatively well with noisy (i.e., real-world) environments, such as those experienced in applied HRI. However, this is often observed to not be the case, cf. [19].

In this paper we seek to evaluate the state-of-the-art in speech recognition for child speech, and to test ASR engines in settings inspired by real-world child-robot interactions. We record a variety of pre-determined phrases and spontaneous speech from a number of children speaking English using multiple microphones. We separate recordings by whether they are comparatively clean, or contain noise from the real-world environment. Through consideration of the results, we highlight the limitations of ASR for child speech, and also make a number of interaction design recommendations to maximise the efficacy of the technology currently available.

2. BACKGROUND

Speech recognition has undergone significant advances, building on or moving on from the use of Hidden Markov Models (HMM) towards using deep neural networks (DNN). DNNs have been shown to outperform older HMM based approaches by some margin against standard benchmarks [12]. For example, in a Google speech recognition task a deep neural network reduced the Word Error Rate (WER) to 12.3%, a 23% relative improvement on the previous state-of-the-art [12].

However, these benchmarks are based on adult speech corpora, such as the TIMIT corpus [17]. It has been noted by other researchers that there is a lack of corpora for children's speech, leading to a lack of training data and a lack of benchmarking for children's speech recognition models [5, 9, 11]. It is commonly assumed that the recent improvements observed in adult speech recognition mean that child speech recognition improved at the same pace, and recognising children's utterances can be achieved with a similar degree of success. However, anecdotal evidence suggests that this is not the

case; Lehman et al. [19] state that recognition of children's speech "remains an unsolved problem", calling for research to be undertaken to understand more about the limitations of ASR for children to ease interaction design.

Children's speech is fundamentally different from adult speech: the most marked difference being the higher pitched voice, due to children having a shorter, immature vocal tract. In addition, spontaneous child speech is marked by a higher number of disfluencies and, especially in younger children, language utterances are often ungrammatical (e.g., "The boy *putted* the frog in the box"). As such, typical ASR engines, which are trained on adult speech, struggle to correctly recognise children's speech [8, 24]. An added complexity is caused by the ongoing development of the vocal apparatus and language performance in children: an ASR engine trained for one age group is unlikely to perform well for another age group.

There have been various attempts to remedy this, from adapting adult-trained ASR engines to the spectral characteristics of children's speech [18, 22], to training ASR engines on child speech corpora [6, 8, 10], or combinations of both. For example, Liao et al. [21] have used spoken search instructions from YouTube Kids to train DNNs with some success, resulting in a WER between 10 and 20%. In [24] vocal-tract length normalisation (VTLN) and DNN are used in combination, and when trained on read speech of children aged between 7 and 13 years, result in a WER of approximately 10%. It should be noted that these results are achieved in limited domains, such as spoken search instructions, read speech, or number recognition [22]. Also, the circumstances in which the speech is recorded are typically more controlled than interactions encountered in HRI, where ambient noise, distance and orientation to the microphone, and language use are more variable.

Whilst children's speech recognition in general is a challenge, HRI brings further complexities due to factors such as robot motor noise, robot fan noise, placement and orientation of microphones, and so on. Many researchers adopt interaction approaches that do not rely on verbal interaction due to the unreliability of child ASR, particularly in 'wild' environments. Wizard of Oz (WoZ) approaches have proven popular to substitute for sub-optimal speech recognition and natural language interaction, but when autonomy is important, WoZ is impractical and the use of mediating interfaces to substitute for linguistic interaction has proven successful. Touchscreens, for example, can serve as interaction devices, they provide a focus for the interaction while constraining the unfolding interaction [1]. However, if we wish the field to continue to progress into real-world environments, then it is unrealistic to exclude verbal interaction due to the prevalence of this communication channel in natural interaction.

3. RESEARCH QUESTIONS

The previous section highlights that the current performance of ASR for child speech remains unclear. We wish to address this by exploring different variables in the context of child speech, such as the type of microphone, the physical location of the speaker relative to a robot, and the ASR engine. These variables motivate a set of research questions presented below, all in the context of child speech. Their evaluation will be conducted with the aim of producing evidenced guidelines for designing verbal human-robot interactions with children.

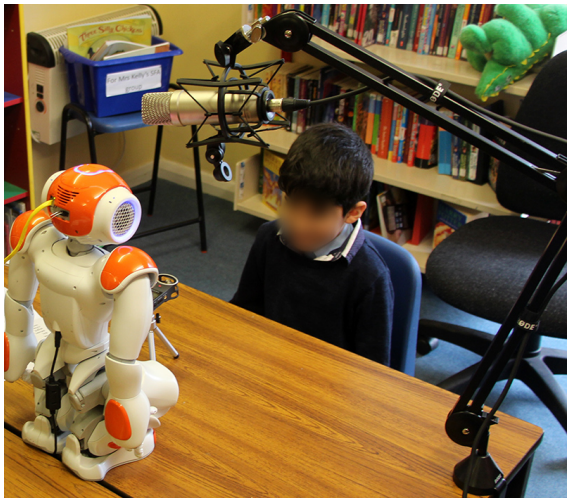

Figure 1: Equipment layout for recording children in a school. The Aldebaran NAO is turned on (but not moving) and records to a USB memory card. The studio microphone and portable microphone record simultaneously.

Q1 Do external microphones produce better results than robot-mounted microphones?

Q2 How can physical interaction setups be optimised for ASR?

Q3 Is there a benefit to using cloud-based or off-board ASR engines compared to a stock robot ASR engine?

Q4 What is the impact of 'real-world' noise on speech recognition in an HRI inspired scenario?

4. METHODOLOGY

In order to address the research questions posed in the previous section, a data collection and testing procedure was designed. At the time of writing, no corpus of child speech suitable for the intended analysis was publicly available. As such, there is a need for the collection of this data; the procedure for this will be outlined here.

4.1 Participants

A total of 11 children took part in our study, with an average age M=4.9, SD=0.3; 5F/6M. The age group is motivated by the many large-scale initiatives in the US, Europe and Japan exploring linguistic interactions in HRI [2, 3, 19, 20, 25], and the fact that this age group is preliterate, so cannot interact using text interfaces. All children had age-appropriate competency in speaking English at school. All participants gave consent to take part in the study, with the children's parents providing additional consent for participation, and recording and using the audio data. The children were rewarded after the study with a presentation of social robots.

4.2 Data Collection

In order to collect a variety of speech utterances, three different categories were devised: single word utterances, multi-word utterances, and spontaneous speech. The single word and multi-word utterances were collected by repeating

after an experimenter. This was done to prevent any issues with child reading ability. Spontaneous speech was collected through retelling a picture book, 'Frog, Where Are You?' by Mercer Mayer, which is a common stimulus for this activity in language development studies [4]. The single word utterances were numbers from 1 to 10, and the multi-word utterances were based on spatial relationships between two nouns, for example, 'the horse is in the stable'. Five sentences of this style were used; the full set can be downloaded from [16].

The English speech from children was collected at a primary school in the U.K. This served two purposes: firstly, to conduct the collection in an environment in which the children are comfortable, and secondly, to collect data with background noise from a real-world environment commonly used in HRI studies, e.g., [15]. An Aldebaran NAO (hardware version 5.0 running the NaoQi 2.1.4 software) was used as the robotic platform. This was selected as it is a commonly used platform for research with children, as well as for its microphone array and commercial-standard speech recognition engine (provided by Nuance). The robot would record directly from the microphones to a USB memory stick. Simultaneously, a studio grade microphone (Rode NT1-A) and a portable microphone (Zoom H1) were also recording. The studio microphone was placed above the robot and the portable microphone just in front of the robot (Fig. 1).

4.3 Data Processing

Encoding and Segmentation.

All audio files were recorded in lossless WAV format (minimum sampling rate of 44kHz). The audio files from each of the three microphones were synchronised in a single Audacity project. The audio files were then split to extract segments containing the speech under consideration. These segments were exported as lossless WAV files, resulting in 16 files per microphone (48 in total) per child. The spontaneous speech was transcribed and split into sentences. This produced a total of 222 spontaneous speech utterances of various lengths ($M = 7.8$ words per utterance, $SD = 2.6$). The full dataset (audio files and transcripts) is available online at [16].

Noisy vs. Clean Audio Recordings.

As the recordings of children in English were collected during the course of a school day, there is a range of background noise. To study the impact of noise on ASR performance, it is desirable to separate the recordings into those that have minimal background noise ('clean' recordings) and those that have marked background noise ('noisy' recordings). Some noise is unavoidable, or would be present in any HRI scenario, such as robot fan noise, so these were considered 'clean'. Other noise, such as birds outside, other children shouting from the adjacent room, doors closing, or coughing would be considered 'noisy'. This means that the clean recordings are not noise-free like those from a studio environment, but are a realistic representation of a minimal practical noise level in a 'wild' HRI scenario, thereby allowing us to evaluate recognition accuracy with greater veracity.

To appropriately categorise the recordings as clean or noisy, each one was independently listened to by 3 human coders with the guidance from above as to what is considered clean vs. noisy. Overall agreement levels between coders was good, with Fleiss $\kappa = .74$ (95% CI [.65,.84]) for the fixed utterances and $\kappa = .68$ (95% CI [.60,.75]) for the spontaneous

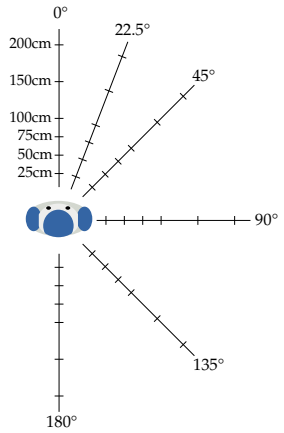

Figure 2: Locations at which speech it played to the NAO to explore how the physical layout of interactions may influence speech recognition rates.

utterances. A recording was categorised as noisy or clean if all 3 coders agreed it was respectively noisy or clean. Where there was any disagreement between coders, the recordings were omitted from analysis of noise impact (59 fixed and 54 spontaneous utterances were excluded). This resulted in 80 noisy recordings, and 37 clean recordings being analysed from the fixed utterances set and 83 clean/85 noisy recordings from the spontaneous utterances set. For some children, the NAO recording failed due to technical difficulties. Therefore, when comparing across microphones, the fixed utterance selection is reduced to 29 clean recordings and 60 noisy recordings.

Manipulation of the Sound Location.

To evaluate the impact of distance and angle on speech recognition, it was necessary to vary the distance between the robot and child, while at the same time keeping the speech utterances constant. As children struggle to exactly reproduce speech acts and over 500 utterances are needed to be recognised, we used pre-recorded speech played through an audio reference speaker (the PreSonus Eris E5) placed at different locations around the robot. In order to match the original volume levels, a calibration process was used where a recording would be played and re-recorded at the original distance between the child and the robot. The audio signal amplitudes between the original and recorded file were then compared. The speaker volume was iteratively revised until the amplitudes matched. This volume was then maintained as the speaker was moved to different distances and angles from the robot, while always facing the robot (to address, at least in part, Q2 from Sec. 3); see Fig. 2 for a diagram of these positions.

4.4 Measures

For recognition cases where a *multiple choice* grammar is used (i.e., the list of possible utterances is entirely predefined, and the recognition engine's task is to pick the correct one), the recognition percentage is used as the metric. Each word or sentence correctly recognised adds 1; the final sum is divided by the number of tested words or sentences. All Confidence Intervals calculated for the recognition percentage include continuity correction using the Wilson procedure. We

use the same metric when using template-based grammars (Sec. 5.2.1).

For the cases in which an open grammar is used, we use the Levenshtein distance as a metric at the *letter* level. This decision was made as it reduces punishment for small errors in recognition, which would typically not be of concern for HRI scenarios. For example, when using the Levenshtein distance at the word level (as with Word Error Rate), if the word 'robots' is returned for an input utterance of 'robot', this would be scored as completely unrecognised. At the letter level, this would score a Levenshtein distance of 1, as only a single letter needs to be inserted, deleted or substituted (in this example, the letter 's') to get the correct result. To compare between utterances, normalisation by the number of letters in the utterance is then required to compensate for longer inputs incurring greater possibility of higher Levenshtein distances.

5. RESULTS

This section will break down the results and analysis such that the research questions are addressed. The results are split into two main subsections concerning: 1) technical implementation details, and 2) general ASR performance. The intention is to then provide a practical guide for getting the best performance from ASR in HRI scenarios, as well as an indication of the performance level that can be expected more generally for child speech under different circumstances.

5.1 Technical Best Practices

Throughout this subsection, the ASR engine will remain constant so that other variables can be explored. In this case, the ASR engine used is the one that comes as default on the Aldebaran NAO, provided by Nuance (VoCon 4.7). A grammar is provided to this engine, consisting of numbers (as described in Sec. 4.2) and single word utterances. Longer utterances, along with open grammar and spontaneous speech will be explored in the subsequent subsection.

5.1.1 Type of microphone

Upon observation of the results it became clear that the robot-mounted microphone was vastly outperforming the portable and studio microphones. When visually comparing the waveforms, there was a noticeable difference in recorded amplitude between the NAO signal and the other two microphones. This was despite the standalone microphone input gains being adjusted to maximise the signal (whilst preventing peak clipping). To increase the signal amplitude whilst maintaining the signal-to-noise ratio, the files were normalised. This normalisation step made a significant difference to the results of the speech recognition. For the portable microphone, the recognition percentage after normalisation (70%, 95% CI [59%,79%]) was significantly improved compared to before normalisation (2%, 95% CI [0%,9%]); Wilcoxon signed-rank test[1] $Z = -7.483, p < .001, r = 0.67$. A similar improvement was observed for the studio microphone when comparing before (5%, 95% CI [2%,12%]) and after (81%, 95% CI [70%,88%]) normalisation; $Z = -7.937, p < .001, r = 0.71$ (Fig. 3). This suggests that the NAO microphones are tuned to maximise the speech level, and if external microphones are to be used, then normalisation of the recordings should be considered a vital step in processing prior to sending to an ASR engine. Therefore, for the remainder of the analysis here, only normalised files are used for the studio and portable microphones.

In exploring Q1, it is observed that the differences between microphones is smaller than may have been expected. The NAO microphones are mounted in the head of the robot near a cooling fan which produces a large amount of background noise. It could therefore be hypothesised that the ASR performance would greatly increase by using an off-board microphone, and that using a higher-quality microphone would improve this further. Using Friedman's test, a significant difference at the $p < .05$ level is found between the NAO (61%, 95% CI [48%,73%]), portable (65%, 95% CI [51%,76%]), and studio (84%, 95% CI [72%,92%]) micro-

[1]Due to the recognition being binary on single word inputs, the resulting distributions are non-normal, so non-parametric tests are used for significance testing.

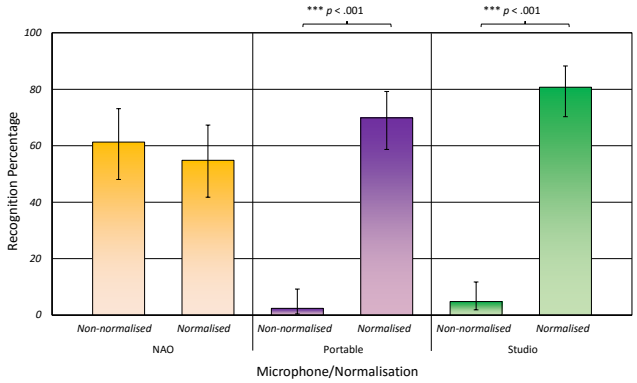

Figure 3: A comparison of recognition percentage of English words and short sentences spoken by children, split by microphone before and after normalisation. * indicates significance at the $p<.001$ level. The recognition is much improved for the portable and studio microphones following normalisation.**

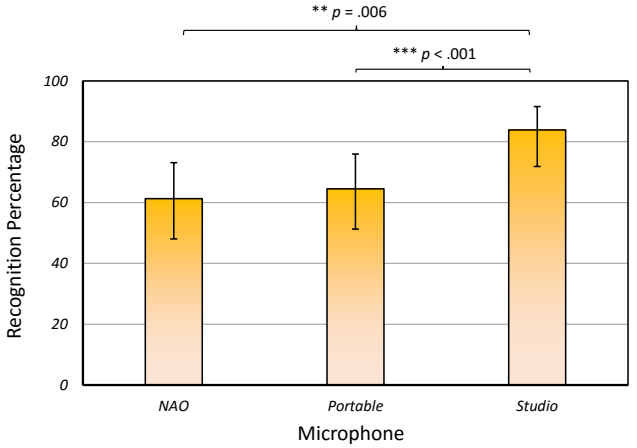

Figure 4: Recognition percentage of numbers spoken by children, split by microphone type (62 utterances). * indicates significance at the $p<.001$ level, ** indicates significance at the $p<.01$ level. The studio microphone provides the best ASR performance, but the difference between on- and lower quality off-board microphones is relatively small.**

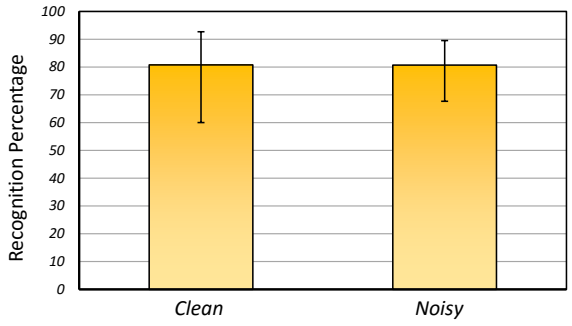

Figure 5: Recognition percentage of single word utterances spoken by children, split by background noise level (83 total utterances). Noise level does not have a significant effect on the recognition rate.

phones; $\chi^2(2) = 9.829, p = .007$. Post-hoc Wilcoxon signed-rank pairwise comparisons with Bonferroni correction reveal a statistically significant difference between the portable and studio microphones ($Z = -3.207, p < .001, r = 0.29$; Fig. 4), and between the NAO and studio microphones ($Z = -2.746, p = .006, r = 0.24$). Differences between the portable and NAO microphones ($Z = -0.365, p = .715, r = 0.03$) were not significant. This suggests that there is no intrinsic value to using an off-board microphone, but that a high quality off-board microphone can improve the ASR results. The difference between the robot microphone and the external studio grade microphone is fairly substantial, with a recognition percentage improvement of around 20%point ($r = 0.28$). It would be scenario specific as to whether the additional technical complexity of using a high-quality external microphone would be worth this gain, and indeed, in scenarios where the robot is mobile, use of a studio grade microphone may not be a practicable option.

5.1.2 *Clean vs. Noisy Recording Environment*

Splitting the files by whether they were judged to be clean or noisy (as described in Sec. 4.3), it was observed that the noise did not appear to have a significant impact on the results of the ASR. Using the studio microphone (i.e., the best performing microphone) for the number utterances, a Mann-Whitney U test reveals no significant difference between clean (81%, 95% CI [60%,93%]) and noisy (81%, 95% CI [68%,90%]) speech; $U = 740.5, p = .994, r = 0.00$ (Fig. 5). The apparent robustness of the ASR engine to noise is of particular benefit to HRI researchers given the increasingly 'real-world' application of robots, where background noise is often near impossible (nor desirable) to prevent.

However, this does not mean that noise does not play a role in recognition rates. In this instance, the ASR engine is restricted in its grammar; the effect of noise in open grammar situations is explored in the next subsection. Additionally, when the distance of the sound source to the microphone is varied, background noise becomes a greater factor.

5.1.3 *Sound Source Location*

Measurements were made as in Fig. 2 using the built in NAO microphone, with the replayed audio from the studio microphone (as described in Sec. 4.3). Due to the number of data points this generates (540 per child), the findings in full will not be produced here, but to get a high-level picture

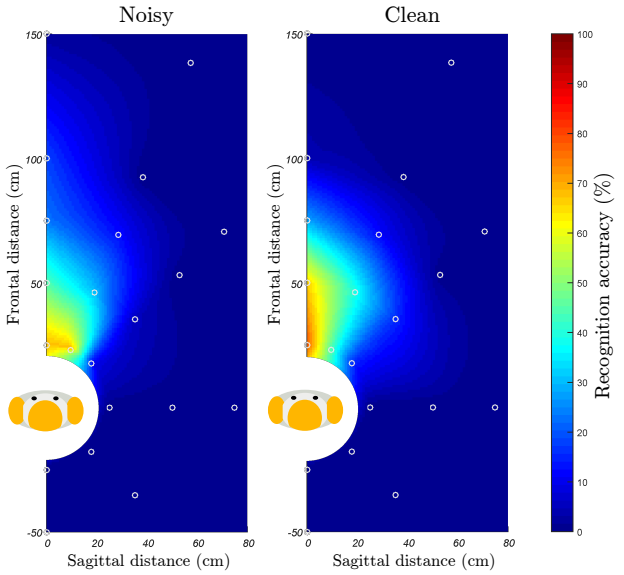

Figure 6: Interpolated heatmap of recognition percentage as a function of distance and orientation to the robot. Interpolation has been performed based on the measurements made at the small white circles. On the *left* is the heatmap for the noisy audio, whereas the *right* is for clean audio. The clean audio is better recognised at further distances from the robot, however, in both cases, recognition accuracy is 0% to the side and behind the robot.

of how the distance and orientation influences recognition rates, a heatmap can be seen in Fig. 6.

Two observations can be made from this data that have particular relevance for HRI researchers. The first is the platform-specific observation that with the NAO robot (currently one of the most widely used research platforms for social HRI) the utterance recognition rate drops dramatically once the sound source reaches a 45 degree angle to the robot head, and becomes 0 once it reaches 90 degrees. The implication of this is that when using the NAO, it is vital to rotate the head to look at the sound source in order to have the possibility of recognising the speech. This is of course dependent on the current default software implementation; four channels of audio exist, but for ASR only the front two are used, and so a workaround could be created for this. The second, broader observation, is that the background noise and distance seem to influence recognition rates when combined. Fig. 5 shows how little impact noise has when the files

Distance (cm)	Clean % [95% CI]	Noisy % [95% CI]
25	73 [52,88]	77 [64,87]
50	65 [44,82]	44 [31,58]
75	27 [12,48]	23 [13,36]
100	4 [0,22]	18 [9,30]

Table 1: ASR recognition rates for children counting from one to ten. Recordings were played frontally at different distances from the robot. Note how recognition falls sharply with distance when the speech contains noise.

are fed directly into the robot ASR, but when combined with distance, there is a marked difference beyond 50cm. Table 1 shows the measurements for the first metre directly in front of the robot; at 25cm the difference between clean and noisy files are minimal, however at 50cm, the difference is more pronounced, with recognition rates dropping fast.

5.2 ASR Performance with Children

The previous subsection addressed variables in achieving a maximal possible speech recognition percentage through modifying the technical implementation, such as different microphones, distances to a robot, orientation to a robot, and background noise levels. This subsection will provide a complementary focus on exploring the current expected performance of ASR with children under different speech and ASR engine conditions. This will include a comparison of differing length utterances, spontaneous utterances, and different ASR engines with varying grammar specifications. For all analyses in this section, the studio microphone signal is used to provide the best quality sound input to the speech engines (and provide a theoretical maximal performance).

5.2.1 Impact of Providing a Grammar

Tests on child speech in the previous subsection were performed with single word utterances, with a grammar consisting of only those utterances. This kind of multiple choice is relatively straightforward, and this carries over to slightly longer utterances too. We compare the recognition rate of the fixed multi-word utterances (34 spatial relation sentences as described in Sec. 4.2) under 3 conditions using the built-in NAO ASR: 1) with a fixed grammar containing the complete utterances, e.g., "one" or "the dog is on the shed" (i.e., multiple choice), 2) with a template grammar for the sentences (as seen in Fig. 8), and 3) with an open grammar. This progressively reduces the prior knowledge the ASR engine has about what utterances to expect. The full mix of noisy and clean utterances were used as there was no observed significant correlation in any of the three conditions between ASR confidence level and noise condition, nor between noise condition and resulting recognition rates. The grammar condition has a significant impact on the recognition percentage; Friedman's test $\chi^2(2) = 39.92, p < .001$. Post-hoc Wilcoxon signed-rank pairwise comparisons with Bonferroni correction reveal a statistically significant difference between the multiple choice (74%, 95% CI [55%,86%]) and template grammars (53%, 95% CI [35%,70%]); $Z = -2.646, p = .008, r = 0.32$. The template grammar in turn offers a significant improvement over the open grammar (0%, 95% [0%,13%]); $Z = -4.243, p < .001, r = 0.51$ (Fig. 7).

5.2.2 Comparison of ASR Engines

Finally, we look at how different ASR engines perform, under identical recording conditions. We compare the Google Speech API (as found in the Chrome web browser for instance), the Microsoft Speech API (as found in the Bing search engine), CMU PocketSphinx, and the NAO-embedded Nuance VoCon 4.7 engine; studies were run in August 2016. The audio samples are those recorded with the studio microphone; they include native and non-native speakers as well as noisy and clean samples; they include both the fixed sentences and the spontaneous speech; no grammar is provided to the engine (i.e., open grammar).

As performing recognition with an open grammar is a much harder challenge for recognition engines, the recognition percentage alone is no longer a sufficient measurement to compare between performance of ASR engines due to the very low number of exact utterance recognitions across all engines. Instead we use the Levenshtein distance (LD) at the letter level. As the utterance length for the spontaneous speech is also variable, the Levenshtein distance is normalised by utterance length (as per Sec. 4.4). This provides a value between 0 and 1, where 0 means the returned transcription matches the actual utterance, and 1 means not a single letter was correct. Values in between indicate the proportion of letters that would have to be changed to get the correct response, therefore lower scores are better. Table 2 provides one recognition example with the corresponding Levenshtein distances.

While the LD provides a good indication of how close the result is from the input utterance, the examples in Table 2

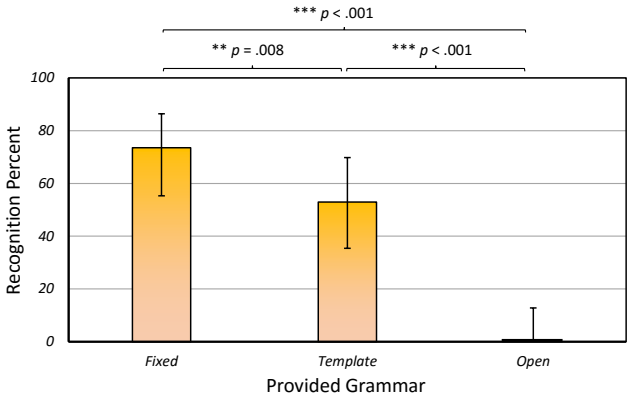

Figure 7: Recognition percentage when providing a fixed grammar, a template grammar, and an open grammar on short utterances. The fixed 'multiple choice' grammar produces the best recognition, followed by a template. The open grammar, on average, recognises almost no sentences correctly.

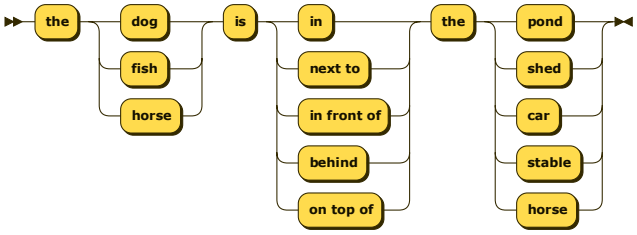

Figure 8: Template for the grammar provided to the ASR for the fixed utterances. 75 different sentences can be generated from this grammar.

Google API	*then the wraps looks at the dog*	[LD=0.17]
Microsoft API	*rat look at dogs*	[LD=0.48]
PocketSphinx	*look i personally*	[LD=0.83]

Table 2: Recognition results and Levenshtein distance for three ASR engines on the input utterance "then the rat looked at the dog". The NAO-embedded Nuance engine did not return any result.

evidence that this metric does not necessarily reflect *semantic* closeness. In this particular case, the Bing result "rat look at dogs" is semantically closer to the original utterance than the other answers. For this reason, we assess recognition performance in open grammar using a combination of three metrics: 1) the Levenshtein distance; 2) raw accuracy (i.e., the number of exact matches between the original utterance and the ASR result); 3) a manually-assessed 'relaxed' accuracy. The utterance would be considered accurate in the 'relaxed' category if small grammatical errors are present, but not semantic errors. Grammatical errors can include pluralisation, removal of repetitions, or small article changes ('the' instead of 'a'). For example, if an input utterance of "and then he found the dog" returned the result "and then he found a dog", this would be considered accurate, however "and then he found the frog" would produce a similar LD, but the semantics have changed, so this would not be included in the relaxed accuracy category.

Table 3 shows that when the input utterance set is changed to use spontaneous speech, the average normalised LD does not change much for any of the ASR engines. Nor do the LD rates change much when only clean spontaneous speech is used, providing further evidence for the minimal impact of noise as established in Sec. 5.1.2. However, there is a marked difference between Google and the other recognition engines. The average LD from Google is around half that of the other engines, and the number of recognised sentences in both the strict and relaxed categories is substantially higher. The recognition performance remains however generally low: using relaxed rules, the currently best performing ASR engine (Google Speech API) for our data recognises only about 18% of a corpus of 222 child utterances (utterances have a mean length of $M = 7.8$ words, $SD = 2.6$).

To help decide whether or not the results returned from Google would actually be usable in autonomous HRI scenarios, it is necessary to determine when the utterance is correctly recognised. This is typically indicated through the *confidence value* returned by the recognition engine. To further explore this, we assess the number recognition percentage at different thresholds within the confidence level (Fig. 9). A total of 101 results from the 222 passed to the recogniser returned a confidence level (a confidence value is not returned when the uncertainty of the ASR engine is too high). To achieve just below 50% semantically correct recognition accuracy, the confidence threshold could be set to 0.8, which would only include 36 utterances. While a clear improvement over the 18% previously achieved when not taking into consideration the confidence value, a 50% recognition rate is arguably not sufficient for a smooth child-robot verbal interaction, and would still require the system to reject nearly 2/3 of the child utterances.

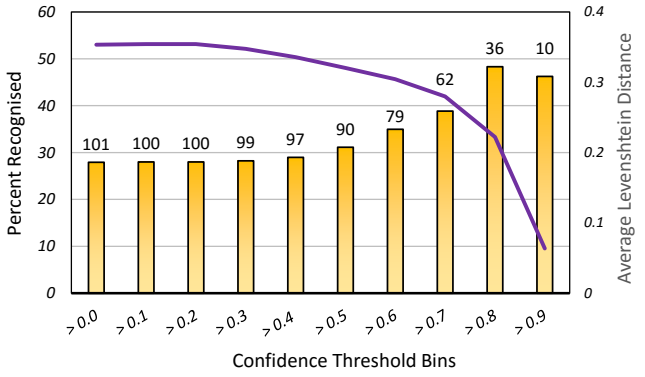

Figure 9: Histogram of recognition percentage (using the relaxed, manually coded criteria) for spontaneous speech grouped by confidence levels (indicated by the number above each bar) returned by Google ASR. The average Levenshtein Distance is also shown on the secondary axis. Recognition increases with higher confidence ranges, but few utterances have a high confidence.

6. DISCUSSION

Our results show that, at the time of writing, automatic speech recognition still does not work reliably with children, and should not be relied upon for autonomous child-robot interaction.

Speech segmentation is one aspect that we did not investigate. The segmentation of speech units and rejecting non-speech parts is an important factor in speech recognition. For example, noise can be mistakenly recognised by ASR engines as speech, or a pause in the middle of a sentence might interrupt the segmentation. Existing solutions (like a beep sound indicating when to talk) are not ideal for children of this age. Our manual segmentation likely leads to better results than would be expected with automatic segmentation.

We did not analyse if gender had an effect on ASR due to the age of the children used in the study. It has been shown that there are no significant differences in the vocal tract between genders at the age under consideration (5-6 years old) [7], so we do not expect differing performance based on gender.

Mitigation strategies for poor ASR performance depend on the ASR engine. We have specifically investigated the use of constrained grammar with the NAO's Nuance engine; and the use of the recognition confidence with the Google ASR. While severely constraining the interaction scope, none of these techniques were found to provide satisfactory results. In our most favourable test case (children speaking numbers from one to ten in front of the robot, at about 25cm; the robot having an explicit 'multiple choice' grammar), the ASR would return an incorrect result in one of four cases, and could not provide any meaningful confidence value. This result is disappointing, particularly when considering that interactions based on 'multiple choice' are difficult to rely on with children, as they tend not to remember and/or comply to the given set of recognisable utterances.

Template-based grammars (or 'slot-filling' grammars) where the general structure of the sentence is known beforehand, and only a limited set of options are available to fill the 'gaps' are a potentially interesting middle-ground between 'multiple choice' grammars and open speech. However, we show that in our test case (grammar depicted in Fig. 8), the correct utterance was recognised in only 50% of the cases, again without any useful confidence value.

In the realm of open grammars, the Google Speech API returned the most accurate results by a large margin. When run on grammatically correct, regular sentences (the ones generated from the grammar depicted in Fig. 8), it reaches 38% accuracy in recognition when minor grammatical differences are allowed. This result, while likely not yet usable in today's applications, is promising. However, when looking

	Google		Bing		Sphinx		Nuance	
	M LD [95%CI]	*% rec.*	*M* LD [95%CI]	*% rec.*	*M* LD [95%CI]	*% rec.*	*M* LD [95%CI]	*% rec.*
fixed (n=34)	**0.34** [0.24,0.44]	*11.8 [38]*	0.64 [0.56,0.71]	*0 [0]*	0.68 [0.64,0.73]	*0 [0]*	0.76 [0.73,0.80]	*0 [0]*
spontaneous (n=222)	**0.39** [0.36,0.43]	*6.8 [17.6]*	0.64 [0.61,0.67]	*0.5 [2.4]*	0.80 [0.77,0.84]	*0 [0]*	0.80 [0.78,0.82]	*0 [0]*
spontaneous clean only (n=83)	**0.40** [0.35,0.45]	*6.0 [16.9]*	0.63 [0.58,0.68]	*1.2 [1.2]*	0.78 [0.72,0.85]	*0 [0]*	0.78 [0.75,0.81]	*0 [0]*

Table 3: Comparison between four ASR engines using fixed, all spontaneous, and clean spontaneous speech utterances as input. Mean average normalised Levenshtein Distance (*M* LD) indicates how good the transcription is. *% rec* indicates the percentage of results that are an exact match for the original utterance, with the values in square brackets *[ ]* indicating matches with 'relaxed' accuracy.

at children's spontaneous speech, the recognition rate drops sharply (to around 18% of successful recognition). This difference can be explained by the numerous disfluencies and grammatical errors found in natural child speech. To provide an example, a relatively typical utterance from our data was "and... and the frog didn't went to sleep". The utterance has a repetition and disfluency at the start, and is followed by grammatically incorrect content. This is, in our opinion, the real challenge that automatic child speech recognition faces: the need to account for the child-specific language issues, beyond the mere differences between the acoustic models of adults vs. children. This is a challenge not only for speech-to-text, but as well for later stages of the verbal interaction, like speech understanding and dialogue management.

Our results allow us to make a number of recommendations for designing child-robot interaction scenarios that include verbal interaction. Most of these are also applicable to adult settings and would be expected to contribute to a smoother interaction.

- Constrain the interaction by leading the child to a limited set of responses. This typically works well for older children, but carries the risk of making the interaction stale.
- Use additional input/output devices. A touchscreen has been found to be a particularly effective substitute for linguistic input [1, 14], but also other devices –such as haptic devices– should be considered.
- Place the young user in the optimal location for ASR. The location and orientation relative to the microphone (and robot) has a profound impact on ASR performance (Sec. 5.1.3). A cushion, stool or chair can help children sit in the optimal location.
- Constrain the grammar of the ASR. While not all ASR engines allow for this (cf. Bing), some will allow constraints or "hints" on what is recognised. This proves to be valuable in constrained interaction settings, for example, when listening only for numbers between 1 and 10 (Sec. 5.2.1).
- Background noise appears to be less of an issue than initially anticipated. It appears that the current ASR engines have effective noise cancelling mechanisms in place. Nevertheless, "the less noise, the better" remains true, particularly when interacting at a distance from the robot (Sec's 5.1.2 & 5.1.3).
- A lack of ASR performance does not mean that the robot should not produce speech, as speech has been found to be particularly effective to engage children.

We opted to evaluate the ASR capabilities of the Aldebaran NAO platform, as it is the most commonly used robot in commercial and academic HRI. While the NAO system under performs for child speech, some performance could be gained through using a high-quality external microphone and cloud-based ASR, with Google as clear favourite.

7. CONCLUSION

Language is perhaps the most important modality in human-to-human interaction and as such, functional natural language interaction forms a formidable prize in human-machine interaction. Speech recognition is the entry point to this and while there has been steady progress in speaker-independent adult speech recognition, the same progress is currently lacking from children's speech recognition. For various reasons –pitch characteristics of children's voices, speech disfluencies, and unsteady developmental changes– child speech recognition is expected to require a multi-pronged approach and recognition performance in unconstrained domains is currently too low to be practical.

This has a profound impact on the interaction between children and technology, especially where pre-literacy children are concerned, typically ages 6 and younger. As they have no means of entering input other than by speaking to the device, the interaction with pre-literacy children stands or falls with good speech recognition.

Our results show that natural language interactions with children are not yet practicable. Today, building rich and natural interactions between robots and children still requires a complex alchemy: a careful design of the interaction that leads the responses of the young user in such a way that restrictive ASR grammars are acceptable, the understanding and production of rich non-verbal communication cues like gaze, and a judicious use of supporting technology such as touchscreens.

8. ACKNOWLEDGEMENTS

This work has been supported by the EU H2020 L2TOR project (grant 688014), the EU FP7 DREAM project (grant 611391), and the EU H2020 Marie Sklodowska-Curie Actions project DoRoThy (grant 657227).

9. REFERENCES

[1] P. Baxter, R. Wood, and T. Belpaeme. A touchscreen-based 'sandtray' to facilitate, mediate and contextualise human-robot social interaction. In *Proceedings of the 7th annual ACM/IEEE international conference on Human-Robot Interaction*, pages 105–106. ACM, 2012.

[2] T. Belpaeme, P. Baxter, R. Read, R. Wood, H. Cuayáhuitl, B. Kiefer, S. Racioppa, I. Kruijff-Korbayová, G. Athanasopoulos, V. Enescu, R. Looije, M. Neerincx, Y. Demiris, R. Ros-Espinoza, A. Beck, L. Cañamero, A. Hiolle, M. Lewis, I. Baroni, M. Nalin, P. Cosi, G. Paci, F. Tesser, G. Sommavilla, and R. Humbert. Multimodal Child-Robot Interaction: Building Social Bonds. *Journal of Human-Robot Interaction*, 1(2):33–53, 2012.

[3] T. Belpaeme, J. Kennedy, P. Baxter, P. Vogt, E. E. Krahmer, S. Kopp, K. Bergmann, P. Leseman, A. C. Küntay, T. Göksun, A. K. Pandey, R. Gelin, P. Koudelkova, and T. Deblieck. L2TOR - second language tutoring using social robots. In *Proceedings of the 1st International Workshop on Educational Robots*, Paris, France, 2015.

[4] R. A. Berman and D. I. Slobin. *Relating events in narrative: A crosslinguistic developmental study*. Psychology Press, 2013.

[5] P. Cosi, M. Nicolao, G. Paci, G. Sommavilla, and F. Tesser. Comparing open source ASR toolkits on Italian children speech. In *Proceedings of the Workshop on Child Computer Interaction*, 2014.

[6] S. Fernando, R. K. Moore, D. Cameron, E. C. Collins, A. Millings, A. J. Sharkey, and T. J. Prescott. Automatic recognition of child speech for robotic applications in noisy environments. *arXiv preprint*, arXiv:1611.02695, 2016.

[7] W. T. Fitch and J. Giedd. Morphology and development of the human vocal tract: A study using magnetic resonance imaging. *The Journal of the Acoustical Society of America*, 106(3):1511–1522, 1999.

[8] M. Gerosa, D. Giuliani, S. Narayanan, and A. Potamianos. A review of ASR technologies for children's speech. In *Proceedings of the 2nd Workshop on Child, Computer and Interaction*, pages 7:1–7:8. ACM, 2009.

[9] P. Grill and J. Tučková. Speech databases of typical children and children with SLI. *PloS one*, 11(3):e0150365, 2016.

[10] A. Hagen, B. Pellom, and R. Cole. Children's speech recognition with application to interactive books and tutors. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding, 2003*, pages 186–191. IEEE, 2003.

[11] A. Hämäläinen, S. Candeias, H. Cho, H. Meinedo, A. Abad, T. Pellegrini, M. Tjalve, I. Trancoso, and M. Sales Dias. Correlating ASR errors with developmental changes in speech production: A study of 3-10-year-old European Portuguese children's speech. In *Proceedings of the Workshop on Child Computer Interaction*, 2014.

[12] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97, 2012.

[13] C.-M. Huang, S. Andrist, A. Sauppé, and B. Mutlu. Using gaze patterns to predict task intent in collaboration. *Frontiers in Psychology*, 6, 2015.

[14] J. Kennedy, P. Baxter, and T. Belpaeme. Nonverbal Immediacy as a Characterisation of Social Behaviour for Human-Robot Interaction. *International Journal of Social Robotics*, in press.

[15] J. Kennedy, P. Baxter, E. Senft, and T. Belpaeme. Social Robot Tutoring for Child Second Language Learning. In *Proceedings of the 11th ACM/IEEE International Conference on Human-Robot Interaction*, pages 67–74. ACM, 2016.

[16] J. Kennedy, S. Lemaignan, C. Montassier, P. Lavalade, B. Irfan, F. Papadopoulos, E. Senft, and T. Belpaeme. Children speech recording (English, spontaneous speech + pre-defined sentences). Data set, 2016. http://doi.org/10.5281/zenodo.200495.

[17] L. F. Lamel, R. H. Kassel, and S. Seneff. Speech database development: Design and analysis of the acoustic-phonetic corpus. In *Speech Input/Output Assessment and Speech Databases*, 1989.

[18] L. Lee and R. Rose. A frequency warping approach to speaker normalization. *IEEE Transactions on Speech and Audio Processing*, 6(1):49–60, Jan 1998.

[19] J. F. Lehman. Robo fashion world: a multimodal corpus of multi-child human-computer interaction. In *Proceedings of the 2014 Workshop on Understanding and Modeling Multiparty, Multimodal Interactions*, pages 15–20. ACM, 2014.

[20] I. Leite, H. Hajishirzi, S. Andrist, and J. Lehman. Managing chaos: models of turn-taking in character-multichild interactions. In *Proceedings of the 15th ACM International Conference on Multimodal Interaction*, pages 43–50. ACM, 2013.

[21] H. Liao, G. Pundak, O. Siohan, M. Carroll, N. Coccaro, Q.-M. Jiang, T. N. Sainath, A. Senior, F. Beaufays, and M. Bacchiani. Large vocabulary automatic speech recognition for children. In *Proceedings of Interspeech*, 2015.

[22] A. Potamianos and S. Narayanan. Robust recognition of children's speech. *IEEE Transactions on Speech and Audio Processing*, 11(6):603–616, 2003.

[23] M. L. Seltzer, D. Yu, and Y. Wang. An investigation of deep neural networks for noise robust speech recognition. In *Proccedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7398–7402. IEEE, 2013.

[24] R. Serizel and D. Giuliani. Deep-neural network approaches for speech recognition with heterogeneous groups of speakers including children. *Natural Language Engineering*, 1:1–26, 2016.

[25] F. Tanaka, K. Isshiki, F. Takahashi, M. Uekusa, R. Sei, and K. Hayashi. Pepper learns together with children: Development of an educational application. In *Proceedings of the IEEE-RAS 15th International Conference on Humanoid Robots*, HUMANOIDS 2015, pages 270–275. IEEE, 2015.

[26] D. Yu and L. Deng. *Automatic Speech Recognition: A Deep Learning Approach*. Springer, 2015.