Correlating ASR Errors with Developmental Changes in Speech Production: A Study of 3-10-Year-Old European Portuguese Children's Speech

Annika Hämäläinen[1,2], Sara Candeias[1,3], Hyongsil Cho[1,2], Hugo Meinedo[1,5], Alberto Abad,[5,6], Thomas Pellegrini[4], Michael Tjalve[7], Isabel Trancoso[5,6], Miguel Sales Dias[1,2]

[1] Microsoft Language Development Center, Lisbon, Portugal
[2] ISCTE - University Institute of Lisbon (ISCTE-IUL), Lisbon, Portugal
[3] Instituto de Telecomunicações - pole of Coimbra, Coimbra, Portugal
[4] IRIT - Université Toulouse III - Paul Sabatier, Toulouse, France
[5] INESC-ID Lisboa, Lisbon, Portugal, [6] Instituto Superior Técnico, Lisbon, Portugal
[7] Microsoft & University of Washington, Seattle, WA, USA

{t-anhama, t-sacand, t-hych, t-humei, michael.tjalve, miguel.dias}@microsoft.com,
{alberto.abad, isabel.trancoso}@l2f.inesc-id.pt, pellegri@irit.fr

Abstract

Automatically recognising children's speech is a very difficult task. This difficulty can be attributed to the high variability in children's speech, both within and across speakers. The variability is due to developmental changes in children's anatomy, speech production skills et cetera, and manifests itself, for example, in fundamental and formant frequencies, the frequency of disfluencies, and pronunciation quality. In this paper, we report the results of acoustic and auditory analyses of 3-10-year-old European Portuguese children's speech. Furthermore, we are able to correlate some of the pronunciation error patterns revealed by our analyses – such as the truncation of consonant clusters – with the errors made by a children's speech recogniser trained on speech collected from the same age group. Other pronunciation error patterns seem to have little or no impact on speech recognition performance. In future work, we will attempt to use our findings to improve the performance of our recogniser.

Index Terms: automatic speech recognition, children's speech, acoustic analysis, auditory analysis, error analysis, European Portuguese, pronunciation quality

1. Introduction

Speech interfaces have tremendous potential in the education of children, with a wide variety of possible applications ranging from pronunciation training applications to educational games. However, automatically recognising children's speech is known to be very challenging. Recognisers trained on adult speech perform substantially worse on children's speech, and word error rates (WERs) are usually much higher than those on adult speech even when using a recogniser trained on children's speech [1-6]. As one might expect, the WERs gradually decrease as the children get older [1-6].

The difficulty of automatically recognising children's speech can be attributed to it being acoustically and linguistically very different from adult speech [1, 2]. For instance, due to their smaller vocal tracts, 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 adult 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, in the degree of spontaneity, in the frequency of disfluencies, in the values of fundamental and formant frequencies, as well as in pronunciation quality [1, 2, 7-11]. The highly variable values of acoustic parameters converge to adult levels at around 13-15 years of age [9]. Research on age-related pronunciation error patterns, so-called phonological processes or deviations, have also been carried out widely (e.g. [12-14]). Studying and understanding the acoustic and linguistic patterns of children's speech is important for designing and implementing well-functioning speech interfaces for children.

This study focuses on European Portuguese (EP) children's speech in the context of automatic speech recognition (ASR). The goal of the study was to identify which pronunciation patterns in EP children's speech are important from the point of view of ASR performance. Previous work on the pronunciation patterns in EP children's speech includes studies carried out to identify common age-related phonological processes [15-17]. Phonetically, EP has characteristics that make the study of children's speech very interesting. Examples of such characteristics include a high frequency of vowel reduction and consonantal clusters, both within words and across word boundaries [15]. These two characteristics make EP difficult for young speakers to produce; their articulatory muscles are not developed enough for skilfully articulating all the speech sounds and clusters of speech sounds of the language. In fact, when children attempt to imitate adult speech, they use certain processes to simplify the production of speech sounds. Such simplification processes might have a negative impact on ASR performance [2].

In this paper, we report findings from a detailed analysis of errors made by an automatic speech recogniser trained and tested with 3-10-year-old EP children's speech. We also analyse children's vowel formants and pronunciation quality with respect to adult speech, and couple our findings with the performance of the children's speech recogniser. We describe the methodology used in this study in Section 2, and detail our findings in Section 3. Section 4 discusses our conclusions and plans for future work.

2. Methodology

To reach our goal, we analysed EP children's speech with specific reference to a speech recogniser built for a multimodal educational game aimed at 3-10-year-old Portuguese children

[18]. We trained and tested the recogniser with speech extracted from a corpus of EP children's speech that was specifically collected with the educational game in mind. When carrying out the analysis, we focused on utterances that had ASR errors, as well as on utterances that had been recognised correctly but with a low confidence score. In addition, we automatically computed the acoustic distance between phones produced by children and phones produced by adults. This section describes the speech material, the automatic speech recogniser, and the methodology used in our study. The results of our analysis are reported in Section 3.

2.1. Speech Material

We used speech extracted from the CNG Corpus of European Portuguese Children's Speech [18]. The corpus contains four types of utterances recorded from children aged 3-10: phonetically rich sentences, musical notes (e.g. *dó*), isolated cardinal numbers (e.g. *44*), and sequences of cardinal numbers (e.g. *28, 29, 30, 31*). The children were divided into two groups when developing the corpus: 3-6-year-olds and 7-10-year-olds. The prompts for both the cardinal numbers and the sequences of cardinal numbers were designed to be easier in the case of the 3-6-year-olds, who were also asked to produce fewer prompts. Depending on their age and reading skills, the children either read the prompts, or repeated them after a recording supervisor. The corpus comes with manually verified transcriptions, as well as annotations for filled pauses, noises, and incomplete, mispronounced and unintelligible words.

Table 1. The main statistics of the speech material.

	Training	Test
#Speakers	432	52
#Word types	605	521
Ages 3-6	*557*	*319*
Ages 7-10	*585*	*494*
#Word tokens	102,537	12,029
Ages 3-6	*9553*	*1148*
Ages 7-10	*92,984*	*10,881*
hh:mm:ss	17:42:22	02:05:34
Ages 3-6	*02:30:24*	*00:18:31*
Ages 7-10	*15:11:58*	*01:47:03*

2.2. Automatic Speech Recognition

For the automatic speech recognition experiments reported in [18], we trained and tested several different Hidden Markov Model (HMM) -based speech recognisers with EP children's speech. Table 1 summarises the datasets used for training and testing the recognisers. The best-performing recogniser, which we also used in this study, was a cross-word triphone recogniser trained using a standard acoustic model training procedure with decision tree state tying (see e.g. [19]). Thirty-eight phone labels were used for training the triphones, which have 14 Gaussian mixtures per state. The recogniser also comprises a silence model, a hesitation model and a noise model; the last two were trained utilising the annotations for filled pauses and noises that are available in the corpus. The recogniser was specifically trained for a multimodal educational game that expects isolated cardinal numbers, sequences of cardinal numbers and musical notes as speech input [18]. Therefore, we used constrained grammars for language modelling purposes: a list grammar for the musical notes, and structure grammars for the isolated cardinal numbers and the sequences of cardinal numbers. The grammar for the isolated cardinal numbers allowed cardinal numbers from 0 to 999, whereas the grammar for the sequences of cardinal numbers allowed sequences of 2-4 cardinal numbers ranging from 0 to 999; the grammars corresponded both to the recorded data and to the expected speech input. During the experimentation phase, we recognised the phonetically rich sentences using a list grammar consisting of the phonetically rich sentences recorded for the corpus; the educational game itself does not use this type of speech input.

For establishing a baseline, we used the female acoustic models from the EP language pack that comes with the Microsoft Speech Platform Runtime (Version 11) [20]. The models in the EP language pack comprise a mix of gender-dependent whole-word models and cross-word triphones trained using several hundred hours of read and spontaneous speech collected from adult speakers of EP. We used the female acoustic models because the acoustic characteristics of children's speech are more similar to adult female speech than to adult male speech [8, 9, 18].

Table 2. 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.

	Full Test Set	Ages 3-6	Ages 7-10
Baseline	18.1 ± 0.7	49.2 ± 3.0	14.9 ± 0.7
Children's ASR	10.0 ± 0.5	27.1 ± 2.6	8.2 ± 0.5

Table 3. The WERs (%) of the children's speech recogniser per utterance type.

	Full Test Set	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

Table 4. The number of word substitution, insertion and deletion errors made by the children's speech recogniser, excluding the phonetically rich sentences.

	Full Test Set	Ages 3-6	Ages 7-10
Substitutions	345	60	285
Insertions	198	15	183
Deletions	303	60	243

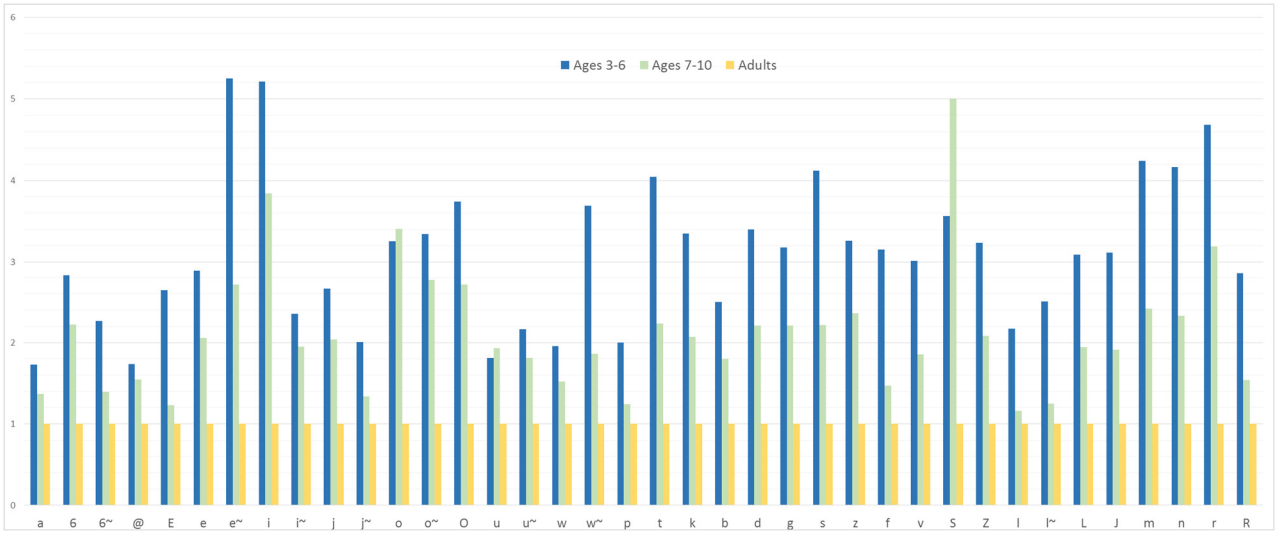

Fig. 1. Average GOP scores for 3-6-year-old and 7-10-year-old speakers, relative to the GOP scores of adult speakers.

Table 2 summarises the speech recognition results obtained with the baseline recogniser and the children's speech recogniser. The children's speech recogniser significantly outperformed the baseline recogniser. It improved by 45% relative over the performance of the baseline recogniser. The improvement was also 45% when calculated separately for both 3-6-year-olds and 7-10-year-olds. Similar to other studies [3-5, 7], the WERs were considerably higher in the case of the younger children.

Table 3 lists the WERs of the children's speech recogniser for each of the recorded utterance types. It also includes the overall WERs without phonetically rich sentences, which represent a prompt type that is not applicable to the educational game. Table 4 presents the corresponding number of substitution, insertion and deletion errors made by the children's speech recogniser; the higher number of errors in the case of the 7-10-year-olds reflects the larger amount of test data in their case. The results in Table 3 make it 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 leaves room for improvement in the case of 7-10-year-olds, it may already be acceptable for the educational game – in particular in the case of musical notes and isolated cardinal numbers.

2.3. Automatic Evaluation of Pronunciation Quality

The Goodness of Pronunciation (GOP) algorithm was originally introduced to assess the quality of non-native speakers' phoneme-level pronunciation in the context of Computer-Assisted Language Learning (CALL) [21]. In this study, we used it to automatically evaluate the quality of the pronunciations produced by the children in our corpus. We investigated if the phone segments with high GOP scores corresponded to mispronunciations or articulatory phenomena typical of EP children's speech. For this study, we computed the GOP scores by first carrying out a forced alignment of the speech utterances using the canonical pronunciations of the words in the orthographic transcriptions of those utterances. The GOP score for each phone segment p was then computed by calculating the likelihood ratio that the phone realisation corresponds to the phoneme that should have been spoken according to the canonical transcription, using the following formula:

$$GOP(p) = \frac{\left|\log \frac{\prod_t P(O_t|p)}{\prod_t max_q(O_t|q)}\right|}{N_p} \qquad (1)$$

where N_p is the number of frames in phone segment p. To obtain the posterior probabilities for the GOP analysis, we used the hybrid HMM/MLP speech recognition system described in [22]. In this case, we used an MLP-based context-independent acoustic model with 39 softmax outputs (corresponding to the 38 European Portuguese phonemes + silence). Notice that each one of the softmax outputs of the MLP acoustic model is interpreted as the posterior probability of the corresponding phoneme in connectionist speech recognition systems. The numerator in Eq. 1 was obtained from the forced alignment using the MLP outputs corresponding to the canonical transcriptions of words. The denominator was simply obtained through free phone recognition, i.e., based on the maximum value of MLP outputs. Essentially, the higher a GOP score is, the more likely it is that the phone in question was mispronounced.

The acoustic model used for the GOP analysis was trained using a corpus of adult speech collected from broadcast news, which is dominated by speech from news anchors and other trained/experienced speakers. It can be assumed that this acoustic model, which has mainly been trained with carefully pronounced speech, together with the canonical transcriptions of the words in question, provides us with a good reference to compare the children's speech with.

We computed the GOP scores for all the utterances in our training and test sets. To serve as a reference, we also calculated the GOP scores for a small in-house corpus of European Portuguese young to middle-aged adults' speech containing phonetically rich sentences read out by speakers aged 25-59. Figure 1 presents the average GOP scores for all the phones in the training and test sets with children's speech, relative to the GOP scores for the young to middle-aged adults' speech (for each phone, the average GOP score was normalised by dividing it by the corresponding average GOP score for adult speech). It illustrates how the quality of the phones produced by the 3-6-year-olds is generally speaking poorer than the quality of the phones produced by the 7-10-year-olds. Furthermore, it

highlights some the phonemes that the children have the most problems producing (e.g. [i], [S], [r]). To analyse which phonemes the children in our *test set* had the most problems with, we picked the 500 phone realisations that had the highest GOP scores (relative to adult speakers' GOP scores) in each age group and calculated the percentage of each phoneme in those subsets. Table 5 shows the phonemes with the highest proportion of potential problems (≥5% of potential problems in the dataset) in the case of the 3-6-year-olds, and Table 6 presents the same information for the 7-10-year-olds.

Table 5. Phones with the highest proportion of potential problems in the case of 3-6-year-old test set speakers.

Phone	Percentage of Top-500 GOP scores
r	14.2
s	10.8
t	10.4
i	9.2
k	6.8
6	6.2
S	5.2

Table 6. Phones with the highest proportion of potential problems in the case of 7-10-year-old test set speakers.

Phone	Percentage of Top-500 GOP scores
S	45.4
i	21.8
s	9.8
k	6.4

2.4. Auditory Analysis

We analysed the word substitution, insertion and deletion errors made by the children's speech recogniser on the set of test utterances excluding the phonetically rich sentences (see Section 2.2 and Table 4). In total, we analysed 87 errors made in the case of the 3-6-year-olds and 39 errors made in the case of the 7-10-year-olds. In some cases, the recogniser did not output any words for the whole utterance. A preliminary analysis of the utterances with recognition errors suggested that the word substitution errors would be the most interesting errors for a thorough auditory phonetic analysis, so we focused on those types of errors in particular. To get a better overall picture of the pronunciation patterns that might be important from the point of view of ASR performance, we also analysed utterances that had been recognised correctly but with a low confidence score (51 utterances from the 3-6-year-olds and 51 utterances from the 7-10-year-olds). Apart from listening to utterances with speech recognition issues, we have also listened to a number of utterances containing phones with large GOP values.

Two qualified phoneticians, one an expert in Portuguese phonetics and another an expert in general auditory phonetics, carefully listened to all the test utterances that had been misrecognised by the children's speech recogniser. They transcribed the children's phonetic realisations of the misrecognised words using SAMPA (Speech Assessment Methods Phonetic Alphabet; [23]), compared their transcriptions with the standard transcriptions of the words in question, and categorised the differences between the two. The results of the auditory analysis are reported in Section 3.

2.5. Acoustic Analysis of Vowel Formants

We analysed EP children's vowels acoustically by computing the average formant values for the phonetically rich sentences in the training and test sets – a total of 1848 and 7077 phonetically rich sentences recorded from the 3-6-year-olds and the 7-10-year-olds, respectively. To be able to compute the average formant values, we obtained phoneme-level segmentations by carrying out a forced alignment of the phonetically rich sentences using the hybrid HMM/MLP speech recognition system discussed in Section 2.3 and [22]. We used context-independent acoustic models for the forced alignment, as they are considered more suitable for linguistically motivated research than context-dependent models (e.g. [24]). We extracted the formant values (by calculating the average of three values taken at 1/3, 1/2 and 2/3 point of each vowel realisation), filtered out aberrant values, and drew the vowel charts using the Praat software [25]. To define the threshold values for filtering, we used the average F2 values for EP adult females [26] as a reference (cf. Section 2.2). Formant values that were 400 Hz below or above the reference values were considered as artefacts and were discarded. After filtering, we were left with a set of 5100 and a set of 24100 vowels for computing the average F1/F2 values for the 3-6-year-olds and the 7-10-year-olds, respectively. Figure 2 illustrates the F1/F2 values for the nine oral vowels of EP, showing the expected shift in formant frequencies. In addition to the children's formant values, the figure presents the average F1/F2 values for 20-30-year-old females in our corpus of young to middle-aged adults' speech. As one might expect, the children's formant values are higher than those of the adult females, with the younger group of children having the highest formant values. The F1/F2 chart is discussed in more detail in Section 3.2.

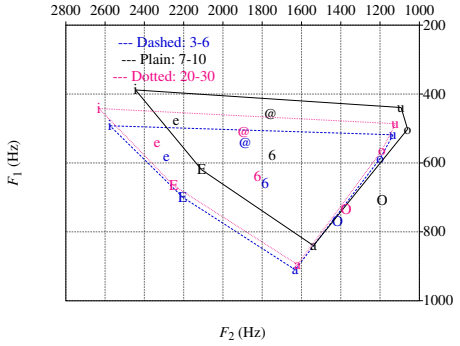

Fig. 2. F1/F2 chart for 3-6-year-olds (dashed line), 7-10-year-olds (plain line) and young female adults (dotted line).

3. Results

This section describes the findings from the auditory analysis (see Section 2.4) and the analysis of vowel formants (see Section 2.5). Before describing any pronunciations in this section, we must clarify that we have chosen to use a phonetic – rather than a phonological – representation of sound patterns because it is closer to the physical reality of language.

3.1. Consonants

Previous studies [16, 17] have shown a high occurrence of consonant cluster reductions in EP children's speech, and we observed the same phenomenon in our data. In fact, ASR errors were often related to the reduction of consonant clusters, especially in the case of liquids. For example, the word *três* ('three') was often pronounced as [t"eS] instead of the standard pronunciation [tr"eS]. This mispronunciation accounted for 10% of the misrecognitions and 10% of the correct recognition results with a low confidence score that we analysed. Considering the fact that children acquire the ability to accurately produce liquid consonants, such as [l] and [r], at the latter stage of their language acquisition process (at around 4 or 5 years old of age), this finding is not surprising. The high frequency of potential problems with [r], which can be seen in Figure 1 and Table 5, seems to support this finding.

The word *um* (["u~]; 'one') was sometimes incorrectly recognised as the word *onze* (["o~z@]; 'eleven'). We hypothesise that these ASR errors were related to background noise in the recordings or to the audible breathing of the speakers right after the production of the word *um*, which might have led the recogniser to confuse *um* with *onze*, whose pronunciation includes the alveolar fricative [z].

As for fricative consonants, the substitution of the phones [s] and [z] with their palatal equivalents [S] and [Z] was common in the case of the 3-6-year-olds. Examples of such substitutions include:

- *sete* ('seven'): [s"Et@] → [S"Et@]
- *cinco* ('five'): [s"i~ku] → [S"i~ku]
- *dezasseis* ('sixteen'): [d@z6s"6jS] → [dZ6S"6jS]
- *dezassete* ('seventeen'): [d@z6s"Et@] → [dZ6S"Et@]
- *dezoito* ('eighteen'): [d@z"Ojtu] → [dZ"Ojtu]

Interestingly, Table 5 suggests that the GOP algorithm is also able to identify the children's problems producing [s] and [S]. The auditory analysis we have carried out so far suggests that the [s]/[S] and [z]/[Z] substitutions might be correlated with ASR errors. However, before drawing any conclusions, we intend to investigate the matter further by analysing the [s] and [S] realisations with very high GOP scores (see Table 5 but also Table 6 for the older children).

When analysing the pronunciation of plosives, we observed the velar consonant [k] often being substituted with an alveolar stop in words like *quinze* ('fifteen'; [k"i~z@] → [t"i~z@]) and *catorze* ('fourteen'; [k6t"orz@] → [t6t"orz@]). This fronting process has also been reported in the literature [16] as one of the most common pronunciation patterns in EP children's speech. Interestingly, the GOP algorithm was also able to pick up the children's problems producing [k] (see Tables 5 and 6). Another interesting observation is that this phone substitution, which crosses phonological categories, did not seem to have any major impact on ASR performance. This is probably due to the nature of our ASR task: the restricted grammars (see Section 2.2) together with the relatively small vocabulary might have allowed us to recover from some mispronunciations.

We also found a devoicing deviation for the alveolar fricative [z] in words like *zero* and *doze*:

- *zero* ('zero'): [z"Eru] → [s"Eru] or [z"Eru] → [S"Eru]
- *doze* ('twelve'): [d"oz@] → [d"os@]

To further analyse devoicing deviations in EP children's speech, we will carry out an acoustic analysis of VOT (Voice Onset Time) in future research.

3.2. Vowels

The automatic analysis of pronunciation quality (see Figure 1 and Tables 5 and 6) suggested that some of the vowels pronounced by the children in our corpus – most notably [i] and [e~]) – deviate from the same vowels pronounced by adults. However, based on our auditory analysis, vowels are usually pronounced correctly by the children, and any deviations in their pronunciation do not seem to result in ASR errors. In fact, we could not identify any word substitution errors caused by deviations in the pronunciation of vowels. Again, this might be because of the restricted grammars used in the ASR experiments.

As for word deletion errors, one specific word caught our attention: the word *e* ('and') was often deleted by the recogniser in the case of cardinal numbers between 22 and 99. Although monosyllabic function words are known to be a common source of ASR errors, these errors also seemed to correlate with a pronunciation pattern that we could observe in the children's speech. In Portuguese, the orthographic form of these cardinal numbers includes *e* between the tens and the units (e.g. *vinte e cinco* ('twenty-five')). However, there are two alternative ways of pronouncing these cardinal numbers: one with the *e* (pronounced as an unstressed [i]) and another without. The speakers in the corpus often merged the pronunciation of *e* into the final vowel of the previous word. This phenomenon, which is typical of EP continuous speech also in the case of adult speakers, gives rise to a change in the syllable structure of the syntagm, which seemed to cause the children's speech recogniser to make a number of word deletion errors. The high proportions of [i] in Tables 5 and 6 reflects this particular phenomenon, examples of which include, for instance:

- *vinte e cinco* ('twenty-five'):
  [v"i~t@ i s"i~ku] → [v"i~t i s"i~ku]
- *cinquenta e quatro* ('fifty-four'):
  [si~k"we~t6 i k"watru] → [s"i~kwe~t i k"watru]

The phenomenon of merging two adjacent segments is common in EP adult speech in the case of weak vowels, such as the near-open central vowel [6], mainly when they appear in unstressed syllables and in the context of a syntagm [15, 27]. We discovered this phenomenon in the children's speech when analysing phone realisations with high average GOP scores for [6] (see Table 5). However, we could not identify any connection with ASR errors in this case. Examples of this merging, or assimilation, phenomenon include:

- *a coisa agora* ('the stuff now'):
  [6 k"ojz6 6g"Or6] → [6 k"ojz ag"Or6]
- *era assim* ('it was like that'): ["Er6 6s"i~] → ["Er as"i~]

The vowel formants F1 and F2 (see Figure 2) showed age-related tendencies that did not seem to correlate with ASR errors. Although the vowel triangles of the 3-6-years-olds are very similar to those of the 7-10-year-olds, the triangle of the 3-6-year-olds has higher F1 values, mainly for close and mid-close vowels. This slight increase in F1 values could be expected as the "closer" articulation of the 3-6-year-olds is related to their vocal tracts being smaller than those of the 7-10-year-olds. The centralization of the front vowels [i], [e] and [E] is reinforced by the total absence of lip rounding, showing that

children become more skilled in their ability to control the articulators with age. This is a view shared by many experts in child language acquisition [12, 14].

3.3. Other Characteristics of EP Children's Speech

We also observed other linguistic events, such as truncated words and repetitions (e.g. [k"wa k"watru] for *qua- quatro* ('fo-four')), especially in the case of the 3-6-years-olds. We expected to observe these events, well-known as hesitations or disfluencies, as they are a characteristic of read speech [27]. However, similarly to [10], they did not have an impact on ASR performance.

Compared with adult speech corpora, some children in this study uttered words with a reduced duration and/or a quiet voice. We believe that there is a psychological explanation for this: especially the younger children often reacted to the recording situation with shyness [18]. The words with a short duration and/or a low volume - in particular monosyllabic words with a simple syllable structure, accounted for a large part of the word deletion errors made by the recogniser. Examples of words that were frequently deleted include, for instance, *e* ('and'; ["i]), *um* ('one'; ["u~]), and *sim* ('yes'; [s"i~]).

4. Conclusions and Discussion

The goal of this study was to identify pronunciation patterns in European Portuguese children's speech that might be important from the point of view of ASR performance. We carefully analysed the errors made by an automatic speech recogniser trained and tested on 3-10-year-old children's speech. Furthermore, we analysed children's vowel formants and pronunciation quality with respect to adult speech, and coupled our findings with the performance of our children's speech recogniser. Our analyses confirmed the general tendencies in European Portuguese children's pronunciation that have been described by others but they also provided us with valuable information on the pronunciation patterns that actually have an impact on ASR performance. Most notably, the simplification of consonant clusters clearly had a negative impact on ASR performance. Using the findings from our analyses, we intend to derive pronunciation rules for adding relevant pronunciation variants into a pronunciation lexicon used by our children's speech recogniser. Such an approach has previously led to significant decreases in word error rates when automatically recognising preschool children's speech [28].

One of the techniques that we used to spot pronunciation error patterns in children's speech was the Goodness of Pronunciation (GOP) algorithm. The results of this analysis nicely correlated with the pronunciation error patterns found by two phoneticians listening to utterances that the children's speech recogniser had recognised incorrectly or correctly but with a low confidence score. In addition, the results of the GOP analysis suggested that 3-6-year-old children might also have frequent problems pronouncing some other phonemes, such as [6] and [t] (see Table 5). The high GOP scores for [6] were related to a common phone merging phenomenon in European Portuguese and did not result in ASR errors in the case of our children's speech recogniser. We have not yet carried out a thorough auditory analysis of the [t] realisations corresponding to the high GOP scores (see Table 5). However, our first impression is that a large number of the high scores result from [t] having been substituted with [k] – a phenomenon that is commonly referred to as backing and is typical of phonological disorders in children's speech [29]. We intend to investigate this matter further in the near future.

Due to the nature of the corpus and the restricted grammars used in the ASR experiments, the analyses reported in this paper clearly have their limitations. The types of utterances in the corpus are not fully representative of everyday language, and the restricted grammars are likely to have helped us recover from pronunciation errors that might have led to ASR errors had a larger vocabulary and a language model been used instead. For these two reasons, the findings of the study are hard to generalise to European Portuguese children's speech recognition tasks other than our own. Moreover, the data in the children's speech corpus is read or repeated speech and, as such, not fully representative of the speech input expected in the multimodal educational game that the children's speech recogniser was built for. Therefore, future studies will have to focus on collecting speech data with a wider variety of utterance types to ensure the diversity of the data from the phonetic and phonological point of view. In addition to that, the setting of future recordings will need to be revised to make sure that the recorded data is more representative of the type of speech that is of interest to us (spontaneous speech instead of read or repeated speech). The best option would be to collect more speech data by recording children's verbal interaction with the multimodal educational game itself.

5. Acknowledgements

Microsoft Language Development Center carried this work out in the scope of the following projects: (1) QREN 5329 Fala Global, co-funded by Microsoft and the European Structural Funds for Portugal (FEDER) through POR Lisboa (Regional Operational Programme of Lisbon), as part of the National Strategic Reference Framework (QREN), the national program of incentives for Portuguese businesses and industry; (2) QREN 7943 CNG – Contents for Next Generation Networks, co-funded by Microsoft and FEDER through COMPETE (Operational Program for Competitiveness Factors), as part of QREN. Sara Candeias is involved in the LetsRead: Automatic Assessment of Reading Ability of Children project of Instituto de Telecomunicações, co-funded by FCT. This work has also been partially been supported by FCT grant PEst-OE/EEI/LA0021/2013.

6. References

[1] Gerosa, M., Giuliani, D., Narayanan, S., Potamianos, A.: A Review of ASR Technologies for Children's Speech. In: Workshop on Child, Computer and Interaction, Cambridge, MA (2009)

[2] Russell, M., D'Arcy, S.: Challenges for Computer Recognition of Children's Speech. In: Workshop on Speech and Language Technology in Education, Farmington, PA (2007)

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

[4] Wilpon, J.G., Jacobsen, C.N.: A Study of Speech Recognition for Children and Elderly. In: IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 349–352. Atlanta, GA (1996)

[5] Elenius, D., Blomberg, M.: Adaptation and Normalization Experiments in Speech Recognition for 4 to 8 Year Old Children. In: Interspeech, Lisbon (2005)

[6] Gerosa, M., Giuliani, D., Brugnara, F.: Speaker Adaptive Acoustic Modeling with Mixture of Adult and Children's Speech. In: Interspeech, Lisbon (2005)

[7] Gerosa, M., Giuliani, D., 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., 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., 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., Potamianos, A.: Creating Conversational Interfaces for Children. IEEE Speech Audio Process. 10(2), 65-78 (2002)

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

[12] Bowen, C.: Children's Speech Sound Disorders.Wiley-Blackwell, Oxford (2009)

[13] Grunwell, P.: Clinical Phonology (2nd ed.). Wiliams & Wilkins, Baltimore, MD (1987)

[14] Miccio, A.W., Scarpino, S.E.: Phonological Analysis, Phonological Processes. In: Ball, M.J., Perkins, M.R., Muller, N. Howard, S. (eds.) The Handbook of Clinical Linguistics. Wiley-Blackwell, Malden (2008)

[15] Candeias, S., Perdigão, F.: Syllable Structure in Dysfunctional Portuguese Children Speech. Clinical Linguistics & Phonetics 24(11), 883-889. Informa Healthcare (2010)

[16] Guerreiro, H., Frota, S.: Os processos fonológicos na fala da criança de cinco anos: tipologia e frequência (vol. 3). Instituto de Ciências da Saúde, UCP, Lisbon (2010)

[17] Almeida, L., Costa, T., Freitas, M.J.: Estas portas e janelas: O caso das sibilantes na aquisição do português europeu. In: Conferência XXV Encontro Nacional da Associação Portuguesa de Linguística, Porto (2010)

[18] Hämäläinen, A., Miguel Pinto, F., Rodrigues, S., Júdice, A., Morgado Silva, S., Calado, A., Sales Dias, M.: A Multimodal Educational Game for 3-10-year-old Children: Collecting and Automatically Recognising European Portuguese Children's Speech. In: Workshop on Speech and Language Technology in Education, Grenoble (2013)

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

[20] Microsoft Speech Platform Runtime (Version 11). Online: http://www.microsoft.com/en-us/download/details.aspx?id=27225, accessed 25 March 2013

[21] Witt, S.M.: Use of Speech Recognition in Computer-Assisted Language Learning. PhD thesis. Cambridge University, Cambridge (1999)

[22] Meinedo, H., Abad, A., Pellegrini, T., Neto, J., Trancoso, I.: The L2F Broadcast News Speech Recognition System. In: FALA, pp. 93-96. Vigo (2010)

[23] Wells, J.C.: Portuguese. http://www.phon.ucl.ac.uk/home/sampa/portug.htm (1997)

[24] Vieru, B., Boula de Mareüil, P., Adda-Decker, M.: Characterisation and Identification of Non-Native French Accents. Speech Commun. 53(3), 292-310 (2011)

[25] Boersma, P.: Praat, a System for Doing Phonetics by Computer. Glot International 5(9/10), 341-345 (2001)

[26] Pellegrini, T., Hämäläinen, A., Boula de Mareüil, P., Tjalve, M., Trancoso, I., Candeias, S., Sales Dias, M., Braga, D.: A Corpus-Based Study of Elderly and Young Speakers of European Portuguese: Acoustic Correlates and Their Impact on Speech Recognition Performance. In Interspeech, Lyon (2013)

[27] Veiga, A., Celorico, D., Proença, J., Candeias, S., Perdigão, F.: Prosodic and Phonetic Features for Speaking Styles Classification and Detection. Toledano, D.T., Ortega, A., Teixeira, A., Gonzalez-Rodriguez, J., Hernandez-Gomez, L., San-Segundo, R., Ramos, D. (eds.) Advances in Speech and Language Technologies for Iberian Languages, Communications in Computer and Information Science 328, 89-98. Springer (2012)

[28] Cincarek, T., Shindo, I., Toda, T., Saruwatari, H., Shikano, K.: Development of Preschool Children Subsystem for ASR and Q&A in a Real-Environment Speech-Oriented Guidance Task. In: Interspeech, Antwerp (2007)

[29] Crosbie, S., Holme, A., Dodd, B.: Intervention for Children with Severe Speech Disorder: A Comparison of Two Approaches. Int. J. Lang. Comm. Dis. 40(4), 467-491 (2005)