Challenges for computer recognition of children's speech

Martin Russell and Shona D'Arcy

Department of Electronic Electrical and Computer Engineering,
University of Birmingham, Birmingham B15 2TT, UK
m.j.russell@bham.ac.uk

Abstract

Some of the most compelling applications of spoken language technology in education involve children, but computer recognition of children's speech is particularly difficult. This paper reviews current approaches to children's speech recognition. It concludes that a fundamental challenge is to raise the performance of 'matched' children's speech recognition systems to the same levels that are achieved by state-of-the-art systems for adults, or, alternatively, to explain why this is not possible, for example by studying human recognition of children's speech. It is suggested that to achieve this it will be necessary to take account of the stages of acquisition and development of children's speech. These processes are well documented in the speech therapy literature, but their computational utility is still to be demonstrated.

1. Introduction

Some of the most compelling applications of spoken language technologies in education involve children. These include interactive tutors for reading, such as that developed in LISTEN [1], pronunciation coaching [2], and educational games. Unfortunately, Automatic Speech Recognition (ASR) is substantially more difficult for children's speech than for adults' speech. Wilpon and Jacobsen [3] report the results of digit recognition experiments using acoustic models trained and tested on a range of age groups. For example, using acoustic models trained on speech from subjects aged between 13 and 59 years, error rates are between 150% and 340% greater for children (aged between 8 and 12) than for adults (depending on the age group of the adults). Even if the acoustic models are matched (i.e. trained on speech from children), error rates for children's speech are between 60% and 176% greater than those for adults' speech using models trained on adult speech. Other studies ([4, 5]) report error rates for children's speech which are up to 100% worse for systems trained on adults' speech compared with children's speech.

This paper reviews current approaches to automatic recognition of children's speech. Techniques such as fequency warping, which are motivated by children's smaller vocal tracts, are considered alongside more generic adaptation techniques such as MAP [6] and MLLR [7]. It is noted that these techniques are able to raise the performance of 'adult' ASR systems on children's speech to that which one would expect from a matched children's ASR system. However, they do not address the fundamental problem noted above, that even with a matched system trained on children's speech, ASR performance is significantly poorer than that achieved for adults' speech with a matched ASR system.

It is suggested that a serious limitation of current approaches to ASR for children's speech is that they fail to take account of the chronological sequence of stages at which children acquire the ability to produce particular speech sounds accurately, and that the existence of these processes differentiates children's and adults' speech This type of phonetic development is well-documented in the speech therapy literature, however its computational utility appears to be untested.

Finally it is suggested that attempts to model this type of information and incorporate it into automatic systems for recognition of children's speech, might also lead to models of the development of speech perception in children, and to more 'human inspired' approaches to general ASR.

2. Vocal Tract Length Issues

Some of the mismatches between children's and adults' speech stem from obvious physiological differences, namely that children's vocal tracts are smaller than those of adults. Consequently, important structure in children's speech occurs at higher frequencies relative to adult speech. This is well-illustrated by studies of children's vowel formant frequencies. Narayanan and Potamianos [8] study the frequencies of the first two formants F1 and F2 for vowels spoken by 7, 10, 13 and 15 year old chidren and adults. Their results show an almost linear decrease in formant values along with a general 'compacting' of the vowel space as age increases. Figure 1 shows the results of a similar study for British English children's speech [9], based on speech from the PF-STAR children's speech corpus [10] and the ABI (Accents of the British Isles) corpus of adults' speech [11]. The decreases in the formant frequencies and compacting of the vowel space for adult data reported in [8] are evident in the figure. For comparison, figure 1 also shows vowel formant frequency data for a different population of British children [2] and the US English children's data from [8]. These new quadrilaterals occupy the same region of the vowel space as the original children's data, but their shapes are slightly different, reflecting accent differences and the different contexts in which the vowels occur.

Provided that there are no bandwidth restrictions, these differences are mitigated by frequency warping or Vocal Tract Length Normalisation (VTLN) (for example, [8, 12]), which is used to compensate both for major differences between children's and adults' speech and also between speech from children of different ages. In cases where bandwidth is restricted the utility of frequency warping is compromised. For example, if bandwidth is reduced from 8kHz to 4kHz the increase in computer speech recognition word error rate is more than 100% greater for children than for adults [9]. Moreover, it seems that the effects of bandwidth reduction on computer recognition of 'easy' children's speech and human recognition of more general children's speech are very similar [9].

Further reasons for the large error rates for recognition of

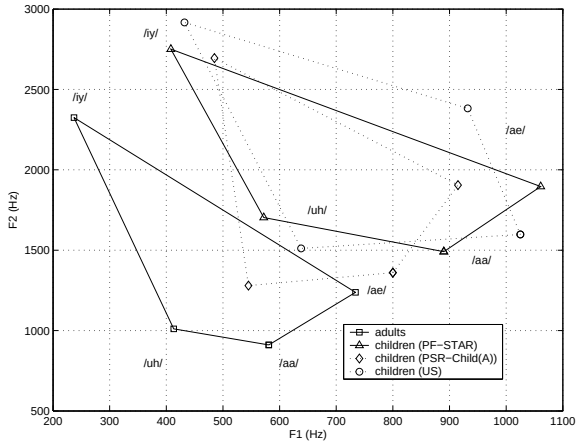

Figure 1: Comparison of first and second formant frequency values for the vowels /iy/, /uh/, /aa/ and /ae/ for British children, British adults and American children

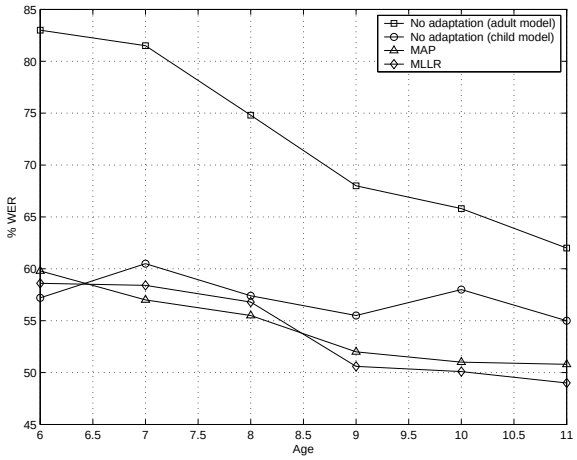

Figure 2: Recognition of children's speech using 'adult speech' models with no dapatation, MAP and MLLR adaptation, and with 'child speech' models with no adaptation

children's speech are the high levels of different types of variability found in children's speech compared with adults' speech. In [13], Lee, Potamianos and Narayanan present a comprehensive study of the magnitude and variation of a range of acoustic speech parameters as a function of speaker age. These include duration, fundamental frequency, formant frequencies and spectral envelope. They demonstrate magnitude differences and higher levels of variability for children's speech, converging to those for adults between the ages of 12 and 15 years.

There is some evidence [14] that this variability is not distributed uniformly across all children, and that ASR performance for a particular child is correlated with a teacher's judgement of pronunciation ability.

3. Generic Adaptation

Another approach to reducing error rates in automatic recognition of children's speech is to apply generic adaptation techniques, such as Maximum A Posteriori (MAP) adaptation [6] or Maximum Likelihhod Linear Regression (MLLR) adaptation [7], or combinations or variants of both, often in addition to frequency warping. Since the vast majority of effort in ASR research has been directed towards improved systems for adults, these techniques are particularly useful to improve the performance of these 'adult' systems on children's speech.

For example, figure 2 (based on results from [15]) shows percentage word error rates for children's speech from a 1800 word vocabulary subset of the British English part of the PF-STAR children's speech corpus [10], with no grammar. The figure shows the results obtained using 'adult' acoustic models trained on the WSJCAM0 British English speech corpus [19] with no adaptation (No adaptation (Adult model)), age-dependent MAP adaptation (MAP) and age-dependent MLLR adaptation (MLLR). The figure also shows performance without adaptation using 'age-independent' acoustic models trained on children's speech (No adaptation (Child model)). Both MAP and MLLR adaptation deliver the largest improvements for the youngest children, raising the performance with models trained on adults' speech to a similar level to that for models trained on children's speech.

The use of these techniques with acoustic models trained on children's speech results in more modest improvements [15].

More recently, Cui and Alwan [18] have described an approach to adaptation of children's speech to 'adult' ASR systems based on alignment of formant-like spectral peaks. For limited adaptation data this technique outperforms conventional approaches to adaptation.

4. Challenges in ASR for Children's Speech

The application of frequency warping described in section 2 is motivated explicitly by physiological differences between the vocal tracts of children and adults. However, generic adaptation techniques such as MAP and MLLR simply attempt to remove mismatches between the expected values of speech parameters encoded in the acoustic model set and the corresponding values observed in the real speech data after alignment. In particular they take no account of the underlying causes of these mismatches and, once vocal tract length normalisation has been performed, they assume implicitly that the remaining differences are due to similar types of variability to that observed in adults' speech. The measure of success (as in figure 2) is the extent to which adaptation reduces the error rates obtained with models trained on adults' speech to levels which are similar to those obtained with models trained on 'matched' children's speech. However, it has already been noted in section 1 that ASR performance using 'matched' models is substantially poorer for children's speech.

This suggests that the fundamental challenge for computer recognition of children's speech is to raise the performance for 'matched' systems (i.e. those trained on children's speech) to similar levels to those which are currently achieved by state-of-the-art systems for adult's speech, or, if this is not possible, to understand properly why not.

An approach to assessing the underlying difficulty of recognising children's speech is to conduct human speech recognition experiments. Lippmann [16] suggests that computer speech recognition error rates for adults' speech are typically an order of magnitude greater that those for humans, and that the gap is wider for 'difficult' speech. D'Arcy and Russell [17] report a word error rate of 3.6% for experiments in which 40 adults listened to short phrases from the PF-STAR corpus, spoken by British English children aged between 6 and 11. These phrases

were judged suitable for 10 year old Italian children learning English, and hence are quite simple. The error rate is higher than expected, and suggests that the speech may be inherently difficult to recognise. However, further work is needed.

Of course, if the target application is automatic pronunciation verification or interactive reading tuition, then there are other challenges, such as defining 'good pronunciation' and developing models and metrics which are able to detect it, and understanding the 'trade-offs' between pronunciation and fluency in reading. However, these issues are beyond the scope of the current discussion.

5. Developmental Issues

It is well known that variability in an adult's speech stems from a very wide range of factors. These include natural, conscious and subconscious variations in the production of individual sounds, variations in the 'inventory' of sounds itself that may arise from social and educational factors or the presence of a regional accent, and possibly variations due to different levels of capability that an individual might have to produce particular sounds.

In the case of children this latter issue appears to be a major consideration. Speech acquisition by children, or the chronological sequence of stages at which children acquire the ability to produce particular speech sounds or combinations of speech sounds accurately, is well documented in the speech therapy and phonetics literature. McLeod and Bleile [20] present a detailed overview with an extensive bibliography. It is natural to ask whether this knowledge is computationally useful, and if so, how it can be used. For example, to what extend can speech development factors which can be characterised at the phonetic or phonological level explain the large variability in children's speech relative to adults' speech reported in [13]. Alternatively, is the variability due to these developmental factors swamped by variability due to, for example poor articulator control?

5.1. Acquisition of consonants

For example, figure 3 (based on data from [21, 22]and [20]) shows typical accurate use of consonants by children as a function of age. The graph labelled '(Bowen (1999))' shows the percentage of consonants used accurately by 75% of Australian children of different ages, while the graph labelled '(McLeod & Bleile (2003))' shows the percentage of consonants used correctly by children as a function of age. The figures agree closely for ages greater than 4 years, and suggest that at least some of the young children whose speech is being used to develop and test ASR systems will not have mastered a complete inventory of consonants.

5.2. Acquisition of consonant clusters

For more complex structures, such as consonant clusters, this problem will be even more acute. For example, according to [20] more than 15% of children aged 6 years or older will substitute $/tw/$ for $/tr/$, $/\theta kw/$ for $/skw/$, $/\theta pl/$ for $/spl/$, $/\theta pr/$ or $/spw/$ for $/spr/$, $/\theta tr/$ or $/stw/$ for $/str/$, or $/\theta kr/$ or $/skw/$ for $/skr/$.

A detailed analysis of the problems which children encounter with word initial consonant clusters is presented in [23].

5.3. Acquisition of vowels

According to [20], accurate production of individual vowels is achieved by 3 years, but accurate production of vowels in more comples contexts such as polysyllabic words will not occur until a child is at least 6 years old. However, no explicit account is taken of such phenomena in modern ASR systems targetted at children's speech.

6. Implications for ASR

It is difficult to estimate the extent to which these phenomena affect the performance of an ASR system for children's speech. However, it has already been noted that there is some evidence that ASR performance correlates with human judgement of pronunciation quality, and it seems likely that a subjective assessment of these developmental issues will inform this human judgement.

It is interesting to note that, in general, speech acquisition in children appears to complete at a similar age to that at which the variability measured in [13] for children's speech converges to adult levels [24].

One would expect consonant cluster errors of the types listed in 5.2 to harm ASR performance, and this could be tested by a careful phone-level analysis of ASR errors for different age groups. Consonant substitution or deletion errors in children's speech would also be expected to compromise MAP and MLLR adaptation, because in either case the model parameters would be moved towards inappropriate data.

If these phenomena are significant for ASR, then it remains to be seen whether they can be accomodated by modifications to pronunciation dictionaries, or whether, for example, more complext models of children's production of consonant clusters are needed. Ultimately this is likely to depend on whether the effects observed in real speech data conform to the phone-level descriptions described above, or whether these descriptions are compromised by basic limitations in children's speech production.

7. Conclusions

This paper presents a review of computer recognition of children's speech. We conclude that, although adaptation of 'adult' systems for improved performance on children's speech is practically important, a more fundamental challenge is to raise the performance of 'matched' systems (systems trained on children's speech) to the levels achieved by state-of-the-art ASR systems for adult speech. Alternatively, the challenge is to show (for example by conducting sufficient human speech recognition experiments for children's speech) that there are fundamental obstacles to achieving this level of ASR performance.

It is suggested that the patterns of phonetic and phonological development in children's speech differentiate it from adults' speech and are potential sources of variation which are not accommodated explicitly in current children's ASR systems. These phenomena are well documented in the speech therapy literature, but their impact on ASR error rates and their computational usefulness, in terms of their potential for reducing error rates by accounting for them explicitly in ASR systems for children, is untested.

Finally, an interesting possibility is that efforts to develop computational frameworks to accommodate the evolution of speech production in children may lead to a better understanding of how to build formal models of the development of speech

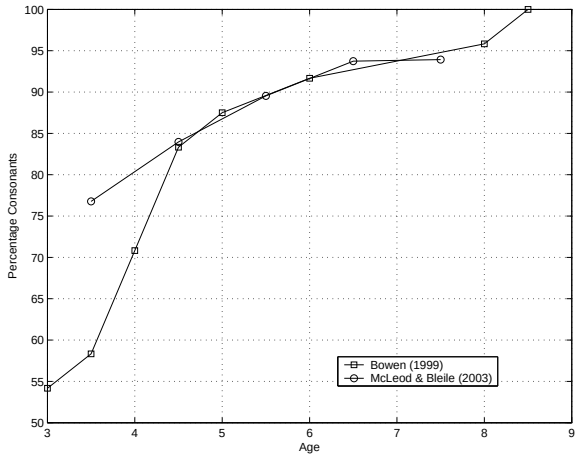

Figure 3: Percentage of consonants accurately used by children as a function of age. Percentage of consonants used accurately by 75% of children (according Bowen (1999) [21]), and percentage correct consonants (according to McLeod & Bleile (2003) [20])

perception in children. This, in turn, might lead to more 'human-like' approaches to geberal ASR.

8. References

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

[2] M J Russell, R W Series, J L Wallace, C Brown and A Skilling, *The STAR system: an interactive pronunciation tutor for young children*, Computer Speech and Language, Vol. 14, Number 2, 161-175, April 2000.

[3] J. Wilpon and C. Jacobsen, *A study of speech recognition for children and the elderly*, Proc. ICASSP'96, Atlanta, USA, 1996.

[4] M. Blomberg and D. Elenius, *Comparing speech recognition for adults and children*, Proc. XVIIth Swedish Phonetics Conf., 156-159, 2004.

[5] D. Giuliani and M. Gerosa, *Investigation of recognition of children's speech*, Proc. ICASSP'03, Hong Kong, China, vol. 2, 137-140, 2003.

[6] J.L. Gauvain and C.H. Lee, *Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains*, IEEE Trans. on Speech and Audio Proc., Vol. 2, 291-298, 1994.

[7] C.J. Leggetter and P.C. Woodland, *Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models*, Computer Speech and Language, vol. 9, pp.171-185, 1995.

[8] S. Narayanan and A. Potamianos, *Creating conversational interfaces for children*, IEEE Trans. Speech and Audio Proc., Vol. 10, No. 2., 65-78, 2002.

[9] M. Russell, S. D'Arcy and Li Qun, *The effects of bandwidth reduction on human and computer recognition of children's speech*, submitted to IEEE Sig. Proc. Letters.

[10] A. Batliner, M. Blomberg, S. D'Arcy, D. Elenius, D. Giuliani, M. Gerosa, C. Hacker, M. Russell, & M. Wong, *The PF_STAR Children's Speech Corpus*, Proc. Interspeech 2005, Lisbon, Portugal, 2761-2764, September 2005.

[11] S.M. D'Arcy, M.J. Russell, S.R. Browning and M.J. Tomlinson, *The Accents of the British Isles (ABI) Corpus*, Proc. Modélisations pour l'Identification des Langues, MIDL 2004, Paris, pp 115-119.

[12] S. Das, D. Nix and M. Picheny, *Improvements in children's speech recognition performance*, Proc. ICASSP'98, Seattle, WA, USA, 1998.

[13] S. Lee, A. Potamianos and S. Narayanan, *Acoustics of children's speech: Developmental changes of temporal and spectral parameters*, Journal of the Acoustical Society of America, Vol 10, 1455-1468, 1999.

[14] Q. Li and M. Russell, *An analysis of the causes of increased error rates in children's speech recognition*, Proc. ICSLP'02, Denver, CO, USA, 2002.

[15] S. D'Arcy, *The effect of age and accent on automatic speech recognition performance*, PhD thesis, submitted to the University of Birmingham, Birmingham, UK, 2007.

[16] R.P. Lippmann, *Speech recognition by machines and humans*, Speech Comm. 22, 1-15, 1997.

[17] S. D'Arcy and M. Russell, *A comparison of human and computer recognition accuracy for children's speech*, Proc. Interspeech 2005, Lisbon, Portugal, 2197-2200, September 2005.

[18] X. Cui and A. Alwan, *Adaptation of children's speech with limited data based on formant-like peak alignment*, Computer Speech and Language, Vol. 20, Issue 4, 400–419, 2006.

[19] T. Robinson, J. Fransen, D. Pye, J. Foote and S. Renals, *WSJCAM0: A British English speech corpus for large vocabulary continuous speech recognition* In Tech. report 0-7803-2431-5/95, Cambridge University Engineering Dept, 1995.

[20] S. McLeod and K. Bleile, *Neurological and developmental foundations of speech acquisition*, American Speech-Language-Hearing Assoc. Convention, Chicago, 2003 (see: http://www.speech-language-therapy.com/ASHA03McLeodBleile.pdf).

[21] C. Bowen, *Phonetic development*, see: http://members.tripod.com/Caroline_Bowen/home.html, 1999.

[22] M.G.E. Kilminster and E.M. Laird, *Articulation development in children aged three to nine years*, Australian Journal of Human Communication Disorders, 6, 1, pp 23–30, 1978.

[23] A.B. Smit, *Phonological error distributions in the Iowa-Nebraska articulations norms project: Word-initial consonant clusters*, Journal of Speech and Hearing Research, 36, pp 931–947, 1993.

[24] A. Fourcin, *Personal communication*, 2007.