APPLICATIONS OF AUTOMATIC SPEECH RECOGNITION TO SPEECH AND LANGUAGE DEVELOPMENT IN YOUNG CHILDREN

Martin Russell[1], Catherine Brown[1], Adrian Skilling[1], Rob Series[1], Julie Wallace[2], Bill Bonham[3] & Paul Barker[3]

[1]Speech Research Unit, DRA Malvern, Malvern, Worcs WR14 3PS, UK
[2]Education Department, Hereford and Worcester County Council, PO Box 73, Worcs WR5 2YA, UK
[3]Sherston Software Limited, Angel House, Sherston, Malmesbury, Wilts SN16 0LH, UK

ABSTRACT

Since 1990 the DRA Speech Research Unit has conducted research into applications of speech recognition technology to speech and language development for young children. This has been done in collaboration with Hereford and Worcester County Council Education Department (HWCC) and, more recently, with Sherston Software Limited, one of the UK's leading independent educational software publishers.

An initial project, known as STAR (Speech Training Aid Research), was prompted by HWCC's awareness of a requirement by teachers for a computerised 'Speech Training Aid' tool to aid young children in the development of a range of communications and language skills. The goal was to develop a computer-based system which was able to distinguish between 'good' and 'poor' pronunciations of a word, spoken by a child in response to a textual, pictorial or verbal prompt, from a 1,000 word children's vocabulary.

The same speech recognition technology has subsequently been integrated into Sherston Software's commercially successful range of animated 'Talking Books', which use stored digitised speech to enable the computer to read words out-loud to a child. This converts them into 'Talking & Listening Books' which, in addition to the existing functions, are able to 'listen' to a child reading and indicate words which have been read incorrectly.

1. THE CHALLENGE

The use of automatic speech recognition in computer based tools for speech and language development in children has enormous potential. While such tools are unlikely to be a substitute for the human interaction which occurs when a teacher or parent helps a child learn to read, they could vastly increase the individual assistance which a child receives, and allow valuable time with the teacher or parent to be used more effectively. Given these advantages, and the economic importance of literacy, it is not surprising that this problem is receiving attention from the speech technology research community (see, for example [1]).

From the perspective of speech technology, the question posed by HWCC was whether automatic speech recognition can be used to distinguish between 'good' and 'poor' pronunciations of a known word spoken by an unknown child. This raises the emotive question of what constitutes a 'good' or 'poor' pronunciation. Jones [2] defines 'poor' speech as a way of talking which it is difficult for most people to understand, caused by mumbling or the lack of definiteness of utterance. By contrast, 'good' pronunciation will enable a child to participate confidently in public, cultural and working life, and will aid accurate reading and spelling. 'Good' pronunciation occurs within the context of a variety of regional accents, and is clearly not the same as Received Pronunciation ('BBC English'). Factors such as a child's confidence in speaking are also relevant.

Assuming that 'good' and 'poor' pronunciation can be identified, there remains the question of whether current speech pattern processing techniques are sufficiently accurate to make the required distinction. This compliments the normal automatic speech recognition problem, where the goal is to correctly classify each utterance (including those which are 'poorly' spoken). Since the late 1980s the most successful automatic speech recognition systems have used phoneme-level hidden Markov models (HMMs) to model acoustic speech patterns. From this perspective the key issue (assuming that phonological variation due to legitimate accent differences can be accommodated by the use of appropriate network structures) is whether there is positive correlation between the probability of a particular acoustic pattern conditioned on a given HMM and a teacher's evaluation of the corresponding utterance as 'good' or 'poor'.

Next there is the issue of children's speech. Speech recognition research has concentrated on recognising adult speech, and in particular, many front-end analysis schemes are tailored to adult male speech. At the December 1995 IEEE ASR Workshop in Snowbird, children's speech appeared in the list of 'problems' at the end of several presentations. HMM-based systems have little explicit knowledge of human speech production and rely on automatic estimation of statistical model parameters from large speech corpora. Since the majority of available corpora consist predominantly of adult speech, it is not surprising that children's speech is found to be a problem. A further consequence of the nature of HMMs is that the question of a formal definition of 'good' or 'poor' speech may be largely academic. Since any such definition is likely to be couched in linguistic or phonetic terms, its usefulness in the context of an HMM based speech processing system is doubtful.

In practice, it was accepted that a large corpus of children's speech would be required in order to make progress, and 'good' speech was defined implicitly by training an HMM set on material collected from children whose pronunciation was assessed as good by their teacher. The Speech Training Aid problem was treated as one of word spotting in continuous speech with a single key-word corresponding to a good pronunciation of the given word. The requirement to distinguish between different pronunciations of the same word, the real operational environment, and the need to cope with children's speech exacerbate the problem, but it is mitigated by the fact that only one key word is active at any one time.

2. A CORPUS OF CHILDREN'S SPEECH

A substantial corpus of speech of 5 to 7 year old children has been recorded. Recordings of five children, whose pronunciation was considered to be 'good' by their teacher, speaking each of the words from a 1,000 word Primary School Reading (PSR) vocabulary were compiled early in the project and used to develop the prototype Speech Training Aid system. The PSR vocabulary was specified with the advice of teachers, language experts and speech therapists, and was judged to be suitable for the age range involved. These recordings were supplemented, first by 'graded' recordings from eight additional children whose pronunciation was rated between very-good and very-poor, and subsequently by recordings of words which are minimally distinct from words in the PSR vocabulary. These recordings were used for early system assessment [4]. All of the children in this part of the corpus are from the Hereford and Worcester region of the UK

The corpus has been extended with recordings of children from Worcestershire, London, Edinburgh, Glasgow and Staffordshire using the prototype Talking and Listening Book system (see section 4). The scope of the corpus is shown in table 1.

	PSR	**PSR (graded)**	**PSR (min. distinct)**	**T&LB**
Number of children	5	8	8	181
Approximate size	5,000 words	1,000 words	80 words	15,000 words

Table 1: Speech data recorded and used during the project .

3. A SPEECH TRAINING AID USING AUTOMATIC SPEECH RECOGNITION

3.1. Previous Results

Previous 'off-line' speech recognition experiments with the 'graded' portion of the DRA Children's Speech Corpus have demonstrated that by comparing the probability of a speech pattern conditioned on a word-level HMM with the probability of the same pattern conditioned on a 'general speech' HMM, it is possible, on average, to distinguish between 'good' and 'poor' pronunciations of a given word spoken by a child [4]. The same experiments showed that there is a correlation between the percentage of words which are judged 'poor' by the automatic system and by teachers, though the automatic system rejects significantly more utterances than the teacher.

3.2. A Real-Time Speech Training Aid

The development of a prototype real-time Speech Training Aid began in 1994, using the Speech Research Unit's AURIX™ programmable, HMM-based continuous speech recogniser, configured as follows:

- Speech was analysed using a 27 channel critical-band filterbank (10kHz bandwidth) at 100 frames per second, followed by variable frame-rate (VFR) [3] analysis and a cosine transform. The representation comprised 8 cosine coefficients, mean filterbank channel amplitude, and 'VFR count', plus the corresponding 'delta' values. This is the normal parameterisation for adult, speaker-independent applications.
- HMMs were trained on the PSR subset of the children's speech corpus (section 2). Standard HMM initialisation and optimisation algorithms were used. The model set comprises 43 3-state monophone HMMs, three 1-state non-speech HMMs, a 'background adult speech' HMM, and a set of 3-state context-sensitive 'forward-looking' diphone HMMs (the latter determined from the training material using the Context Adaptive Phone (CAP) modelling approach [5]). The monophone HMMs, non-speech HMMs and 'background adult speech' HMM were combined to form a general speech HMM. The function of the 'background adult speech' model is to match verbal prompts from a teacher, which might otherwise match well with the word models.

The Speech Training Aid uses a Shure SM10 head-mounted microphone. The child is prompted to speak the required word (using either text, graphics, or verbally, or a combination) by host software running on a typical UK primary classroom computer. The speech signal is compared, in parallel, with a word-level HMM (obtained by concatenating the appropriate diphone HMMs), and the 'general speech' HMM. The child is judged to have produced a good pronunciation if the best match is obtained with the word model. A number of features are particularly important:

- Because AURIX is truly continuous, the child is not constrained to speak within a particular 'time window'. Everything that the child says is recognised, and a positive response is returned

only if AURIX's partial traceback algorithm returns a path through the word model.

- Changing the probability of the 'word' and 'general speech' components of the syntax, biases the system towards, or away from, the general speech model, and hence makes it more or less discerning. This 'general speech model bias' parameter can be set by the teacher on an 'accuracy' scale between 'high' and 'low'.
- The combination of a general speech HMM, AURIX's standard adaptive noise tracking algorithm, and the use of a close talking microphone help ensure that the system is able to function in a noisy classroom environment.

3.3. Trials of the Speech Training Aid

A pilot study to evaluate the Speech Training Aid has been taking place at a Worcester primary school since January 1995. Forty children aged between 5 and 7 have had regular (once per week), supervised use of the system. At first the trials took place in a quiet library area, but they have subsequently been moved into a classroom. Several important factors became clear:

- Children were less intimidated than adults by the act of talking to a machine. They regard it as exciting, but not unnatural.
- Using a computer appears to be a significant motivating factor for children. Apart from the superficial attraction of the technology, there was also evidence that a child considers a computer 'non-judgmental' and is more likely to attempt an unfamiliar word.
- A good child-machine interface is essential.

Initially teachers were skeptical of the technology's capabilities and robustness to the classroom environment, and of its ability to maintain children's attention. However, as trials have progressed the children's interest has not wavered and the teachers' enthusiasm has grown such that they are now putting additional children forward to use the system. Although no quantitative measures of the systems educational benefits have been made, those involved in the trial now believe that there is evidence that the system can aid clarity of speech and improve confidence.

4. TALKING AND LISTENING BOOKS

4.1. 'Talking' Books

'Talking Books' are animated, computer-based books which include the ability to play-back stored recordings of speech. Sherston Software is one of the UK's leading Primary School software publishers, and has experience of speech technology through adding a 'talking' facility to existing reading schemes and proprietary titles. By employing a speech compression technique, the speech data storage requirements have been reduced such that the software will run on platforms which are typically available at the younger end of UK Primary Schools (in particular, a CD-ROM is not required).

A recent pilot study of the Talking Books involved 32 infant children from four classes. Preliminary results are encouraging. Children exposed to the Talking Books have shown significant increases in word accuracy compared with those who were not. Teachers have also been favourable [5].

4.2. 'Talking and Listening' Books

The addition of speech recognition technology to 'Talking Books', to convert them into 'Talking and Listening Books', is compelling. Ideally, this would create an interactive, computer based reading scheme which would not only read out-loud to the child, but also enable the child to read out-loud to the computer. However, reading tuition presents additional challenges to the speech technology. The Speech Training Aid application concentrates on accuracy of pronunciation of individual words, whereas fluency and comprehension are emphasised in the teaching of reading. To cope with fluency it is necessary to move towards continuous speech recognition, which is within the capabilities of AURIX™ but presents additional problems in terms of robustness and the complexity of the child-machine interface. Also, approaches to reading which attach primary importance to comprehension may require that the system is able to cope with the omission of non-essential words and even the use of synonyms.

A prototype Talking and Listening Book system has been developed by integrating AURIX™ and Sherston Software's 'Talking Books'. The initial version used the Speech Training Aid 'vocabulary independent' phoneme-level HMMs trained on the PSR corpus (section 2). Prototype 'Talking and Listening Books' have been used by nearly 200 children, and audio recordings of the majority of these interactions have been made and used to adapt the HMM set incrementally towards this vocabulary.

The prototype Talking and Listening Books have been demonstrated 'live' at key educational exhibitions, including the World Conference on Computers in Education in July 1995 and the British Educational Training and Technology exhibition in January. The response from educational professionals who have seen these demonstrations is overwhelmingly positive. School trials are scheduled to begin in the near future.

5. THE CHILD-MACHINE INTERFACE

Whenever the systems described in this paper are used by children it is clear that the design of the child-machine interface is of paramount importance. The initial Talking and Listening Book system prompted a child to speak the required word, then moved on to the next word either after an acceptable pronunciation had been produced, or after a fixed time period. This gave too little control to the child and was unacceptable. Either the child would feel pressure as the end of the time period

approached, or frustration at having to wait for the next word if their utterances were not accepted (either because of poor pronunciation or recognition error). In the most recent system, control is given to the child, who decides when to move on to the next word, or to play-back the recorded speech.

A further factor is higher-level control of the software by the child. Talking and Listening Books involve animation, and talking-and-listening by both computer and child. A simple and unambiguous interface is needed to enable the child to operate these functions easily. Good design of this interface is essential to make these systems useable by an unsupervised child

6. SYSTEM PERFORMANCE

Figure 1 shows the results of an off-line laboratory evaluation of the January 1996 prototype Talking and Listening Book system. Technical details are as defined in section 3.2, except that the HMM parameters were trained on a combination of the PSR corpus and a subset of the Talking and Listening Books corpus. The system was tested on the 'minimally distinct PSR' subset of the corpus. The figure shows percentage false acceptance (one member of a minimally distinct pair is accepted as a valid pronunciation of the other) and false rejection (a 'good' pronunciation of a given word is rejected) as a function of the bias towards the general speech model, which was discussed in section 3.2. The 'equal error' rate is approximately 30%.

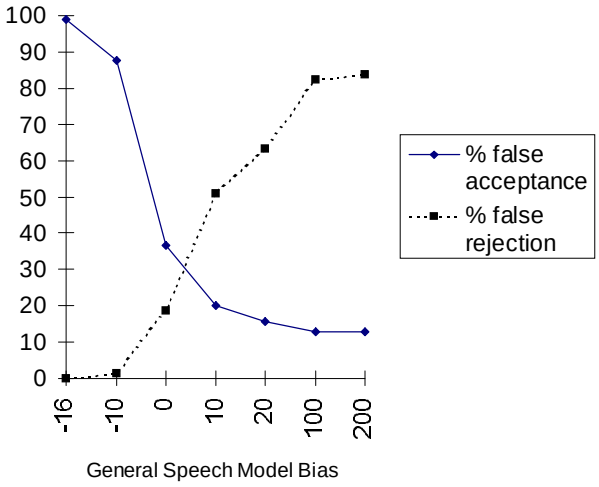

Figure 1: Performance of the January 1996 Talking and Listening Book system as a function of the general speech model bias parameter

Although this figure provides a 'benchmark' for calibration, its relevance to actual performance is debatable. The ability to distinguish between minimal pairs is important, particularly for the Speech Training Aid application, but many of the distinctions which the system is required to make in the reading application are less demanding. Furthermore, the bias which is used may not correspond to the equal-error-rate. For example, when a child uses the system for the first time a low value would be chosen to bias the system towards acceptance.

7. CONCLUSIONS

The results of the trials of the Speech Training Aid system and experiences with the prototype Talking and Listening Books suggest that there are benefits to be gained from deploying current speech recognition technology in tools for speech and language development in young children. Children are motivated by the opportunity to use computers in this aspect of their education and do not appear to be intimidated by talking to a machine. Teachers involved in the Speech Training Aid trials, although initially skeptical about the technology, are now enthusiastic and believe that it is of educational benefit.

Suitable design of the child-machine interface is critical in determining acceptability and educational usefulness. Indeed it is arguable that these considerations are currently more important than the performance of the speech recognition technology.

In order to realize the benefits of the technology, it must be made widely available to primary schools at an affordable cost. This can only be achieved by making use of the limited computing power which is already available in the primary school classroom. To this end, the speech recognition technology which underpins the AURIX™ system is currently being ported to run on this type of computer.

8. REFERENCES

1. Jack Mostow, Steven F Roth, Alexander G Hauptmann and Matthew Kane, "A prototype reading coach that listens", *Proc. Twelfth National Conference on Artificial Intelligence (AAAI'94) Seattle, WA, August, 1994.*

2. D Jones, 'The pronunciation of English', *Cambridge University Press, 1967.*

3. S M Peeling and K M Ponting, "Variable frame rate analysis in the ARM continuous speech recognition system", *Speech Communication 10, pp 155-162, 1991.*

4. R W Series, "A Speech Training Aid", *Proc. ESCA Workshop on Speech and Language Technology for Disabled Persons, Stockholm, Sweden, May 31 - June 2, 1993.*

5. J Medwell, "Talking books and the teaching of reading", *forthcoming.*

6. R K Moore, M J Russell, P Nowell, S N Downey and S R Browning, "A comparison of Phoneme Decision Tree (PDT) and Context Adaptive Phone (CAP) based approaches to vocabulary-independent speech recognition", *Proc. IEEE International Conf. on Acoustics, Speech and Signal Processing, ICASSP'94, Adelaide, Australia, April 1994.*