# Comparison of Speaker Verification Performance for Adult and Child Speech

Saeid Safavi [1], Maryam Najafian [1], Abualsoud Hanani [2], Martin Russell [1], and *Peter Jančovič* [1]

[1] School of Electronic, Electrical and Computer Engineering, University of Birmingham, Birmingham, UK

[2] School of computer systems engineering, Birzeit University, Palestine

{sxs796, mxn978, m.j.russell, p.jancovic}@bham.ac.uk, ahanani@birzeit.edu

## Abstract

Although speaker verification is an established area of speech technology, previous studies have been restricted to adult speech. This paper investigates speaker verification for children's speech, using the PF-STAR children's speech corpus. A contemporary GMM-based speaker verification system, using MFCC features and maximum score normalization, is applied to adult and child speech at various bandwidths using comparable test and training material. The results show that the Equal Error Rate (EER) for child speech is almost four times greater than that for adults. A study of the effect of bandwidth on EER shows that for adult speaker verification, the spectrum can be conveniently partitioned into three frequency bands: up to 3.5-4kHz, which contains individual differences in the part of the spectrum due to primary vocal tract resonances, the region between 4kHz and 6kHz, which contains further speaker-specific information and gives a significant reduction in EER, and the region above 6kHz. These finding are consistent with previous research. For young children's speech a similar pattern emerges, but with each region shifted to higher frequency values.

**Index Terms**: speaker recognition, child speech, Gaussian mixture model, bandwidth, PF-STAR, ABI-1, ABI-2

## 1. Introduction

Several levels of information are contained in a speech signal over and above its linguistic content. Its primary function is communication, but it also conveys information about the speaker's identity, gender, social group, geographical origin, health and emotional state. Speech recognition is concerned with extracting the underlying linguistic message in an utterance, whereas speaker recognition is concerned with extracting the identity of the person speaking the utterance. As speech interaction with computers becomes more pervasive, and its applications become more private and sensitive, the value of automatic recognition of a speaker based on vocal characteristics increases.

Over recent years, there has been huge growth in social networking, offering new and varied ways of communicating via the internet. According to research by Ofcom [1], social networking sites are most popular with teenagers and young adults. For example, almost half of children aged from 8 to 17 who use the internet have set up their own profile on a social networking site. An automatic system that recognises the child based on his or her voice and confirms the identity of the individual with whom the child is communicating, could be a valuable safeguard for a child engaged in social networking. There are also applications in education. For example, an interactive tutor that could identify each child in a class could automatically continue a previous lesson, adapt its content to suit the child, and log the child's responses appropriately without the child needing to go through a formal login process.

The general area of speaker recognition is divided into Verification and Identification. The focus of this paper is verification, in which the goal is to determine from a voice sample if a person is who he or she claims. Furthermore, in either task the speech can be constrained to be a known phrase (text-dependent) or unconstrained (text-independent).

The most commonly used parameterisation for speaker verification is to represent an utterance as a sequence of Mel-Frequency Cepstral Coefficient (MFCC) vectors. The distribution of MFCC vectors, for a whole population or for individual speakers, is typically captured using a Gaussian Mixture Model (GMM) [2, 3]. The speaker GMM models are built by MAP adaptation of a Universal Background Model (UBM), a speaker-independent GMM constructed using data from a variety of speakers and background conditions. This approach has been very effective for various speaker identification and verification tasks [4]. An alternative is to use discrimination-based approaches, such as Support Vector Machines (SVM), which have been shown to obtain comparable, and in some cases better, performance than GMM based systems. The use of GMM supervectors, which consist of the stacked means of the mixture components, in the context of SVMs has also been successful [5]. A combination of the GMM and SVM approaches, where the GMM was used to calculate the likelihood values and the SVM classifier was then used to separate the likelihood values produced by a correct speaker and impostor, was used in [6] and showed slightly better results than the GMM system alone. Speaker verification systems usually employ a score normalisation procedure to cope with score variability and to make the decision threshold tuning easier.

The use of high-level information, such as word usage, prosody or phone sequence analysis, for speaker verification has resurged in recent years [7].

It has been shown that acoustic and linguistic characteristics of children's speech are very different from those of adult's [8-10]. For example, children's speech is characterized by higher pitch and formant frequencies, and perceptually important features occur at higher frequencies [9]. Consequently, the impact of bandwidth reduction on speech recognition accuracy is greater for children's speech than for adults [11,12]. It has also been shown that children's speech exhibits greater levels of inter- and intra-speaker variability than adult speech [9]. Variability is highest for young children, converging to adult values when children reach the age of 13. Even for young children there is some evidence that the degree of variability varies significantly between individuals [11].

Although automatic recognition of children's speech has been the subject of considerable research effort, there is little published work on issues and algorithms related to automatic children's characteristics from his or her speech, [20, 21, 23].

For example, we do not know how the balance between increases in inter- and intra-speaker variability will affect speaker verification for child speech. In addition, we do not know the effect of bandwidth on speaker verification accuracy for children, although some studies of the effects of different frequency bands on adult speaker verification have been reported [22].

This paper describes an initial study of speaker verification performance for children's speech, using a text-independent GMM-based automatic speaker verification system. We present the results of speaker verification experiments for adult and child speech using comparable speech corpora, and study the effect of bandwidth on the verification performance. Our results demonstrate that the Equal Error Rate (EER) for children's speech is four times greater than for adults' speech. In addition, experimental results on the effect of bandwidth on speaker verification performance suggest that for both adult and child speech the spectrum can usefully be partitioned into three regions. For adult speech these are (i) the region up to 3.5-4kHz, which contains the primary vocal tract resonances and contributes to verification, (ii) the region between 3.5-4kHz and 6kHz, which contains further speaker-specific information, and (iii) the region above 6kHz. The importance of region (ii) has been reported elsewhere [20]. A similar partition of the spectrum is also valid for young children, but with the break points at higher frequencies, as one would predict. For older children these bands lie between those for young children and adults.

# 2. Corpora of child and adult speech

Three corpora of British English speech were used in this research: The PF-STAR corpus of children's speech [13], and the two "Accents of the British Isles" (ABI) corpora of regionally accented adults' speech: ABI-1 [14] and ABI-2. ABI corpus contains speech from 27 different locations representing distinct locations, as against with PF-STAR which has been recorded in two different locations. But we believe that accent differences do not have significant effect on the performance of speaker recognition system.

## 2.1. The PF-STAR children's speech corpus

The PF-STAR children's speech corpus [13] comprises 14 hours of recordings from 158 British children (52% male), from Birmingham and Malvern, aged between 4-15 years, but with 92% of the children aged 6-11. The speech was recorded at 22.05kHz sample rate using close talking and desk microphones in a relatively quiet environment (typically a room or space off the school library). The texts were presented to the children on a laptop using 'in-house' prompting and recording software. From the entire corpus, all data from 150 speakers were used; the remaining 8 speakers were the youngest children and did not record sufficient data to be included in the experiment.

## 2.2. The ABI speech corpora

The Accents of the British Isles (ABI) speech corpora were collected to support research into the implications of regional accents for speech and language technology. The two ABI corpora comprise recordings of speech representing twenty-six regional accents of British English plus 'Standard Southern English' (SSE). With the exception of SSE, all of the recordings were made on location in towns or cities that were judged to be representative of particular accents. The objective in each location was to record twenty subjects (ten men and ten women) who were born in the location and had lived there for all of their lives. The SSE speakers were selected by a phonetician. Each subject recorded approximately 15 minutes of read speech. The prompt texts were chosen for their relevance to applications or their phonetic content. The microphones, recording and prompting software, and sample rate are the same as for the PF-STAR corpus. The recordings were made in relatively quiet rooms in libraries or community centres.

### 2.2.1. The ABI-1 speech corpus

ABI-1 [14] comprises recordings of 280 subjects: twenty from each of 13 locations representing distinct accents of British English plus twenty subjects who were judged to speak Standard Southern English. ABI-1 consists of approximately 70 hours of recordings, with speakers' ages ranging from 16 to 79 years.

### 2.2.2. The ABI-2 speech corpus

ABI-2 was recorded using exactly the same methodology as ABI-1. It comprises approximately 70 hours of recordings of 286 speakers representing 13 regional accents of British English that are not covered in the original ABI-1 corpus. The material recorded is the same as in ABI-1, except that each subject recorded an additional set of 22 SCRIBE sentences.

# 3. Speaker verification system

## 3.1. Front-end processing

The front-end processing applied to all corpora is as follows. The speech is pre-emphasised and periods of silence were discarded using an energy-based speech activity detector (SAD). The speech was then segmented into 32-ms frames with a shift of 16-ms between frames, and a Hamming window was applied to each frame. The short-time magnitude spectrum, obtained by applying the FFT, is passed to a bank of 32 Mel-spaced triangular band-pass filters, spanning the frequency region from 64Hz to 11050Hz. Each speech frame is then represented as a 38 dimensional feature vector, consisting of 19 static MFCCs and 19 delta MFCCs. Finally, Feature Warping [15], with 3-seconds window, is applied on the MFCC feature vectors to reduce the effect of channel mismatch and additive noise.

## 3.2. Modelling

The speaker verification system is based on the Gaussian Mixture Model - Universal Background Model (GMM-UBM) approach. This is a likelihood ratio detector, in which the ratio is computed, for an unknown test utterance, of the probability of the utterance given the speaker model and the probability of the same utterance given the UBM. This score is normalized using "max" log likelihood ratio normalisation and then compared with a threshold to determine whether the utterance is accepted as being from the 'true speaker' or rejected as an 'impostor'.

One gender independent UBM was trained using approximately 4hrs of speech data from the ABI-1 corpus, with 10 iterations of the EM algorithm.

### 3.2.1. Speaker dependent training data

The test speakers were taken from the ABI-2 corpus for adults and the PF-STAR corpus for children. In total, 152 adult speaker-dependent GMMs and 150 children speaker-dependent GMMs are trained.

The speaker-dependent GMMs for adults and children were obtained by applying MAP-adaptation to the means of the GMM-UBM using the relevance factor 10 [3]. In all cases the adaptation was performed using approximately 48 seconds of speech data from each subject (after silence removal).

#### 3.2.2. Test data

The evaluation strategy follows the methodology used in the NIST 2003 Speaker Recognition Evaluation Plan [16].

The test data for the adults comprises 902 segments from the 152 target speakers. Each test segment is evaluated against 10 randomly chosen 'impostors' and the true speaker.

The test data for the children comprises 875 segments from the 150 target speakers. Each test segment is evaluated against 10 randomly chosen 'impostor' speakers from the same age group as the true speaker.

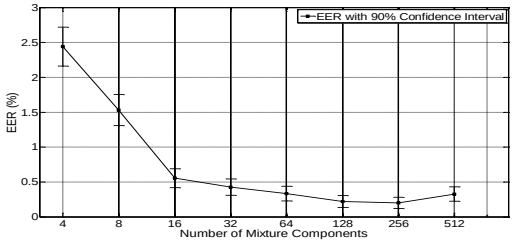
The speech duration of each test segment, for both the adult and child speech experiments, is fixed at 4.8 seconds (after the silence removal).

## 4. Experimental results and discussion

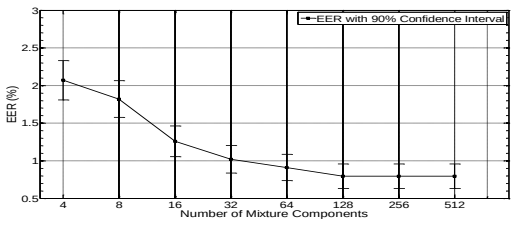
The standard NIST software is used to measure verification performance [17]. A 90% confidence interval is calculated for the EER and indicated by error bars in the results. To calculate the confidence interval, the parametric method from [18] is used to calculate the error margins on the False Accept Rate (FAR) and False Reject Rate (FRR) at a given threshold.

### 4.1. Effect of number of mixture components

First we analyse speaker verification performance for adults and children for various numbers of mixture components, using the maximum 11.025kHz bandwidth. The results are presented in figure 1. It can be seen that the best performance for adults is 0.22% EER with 128 or 256 mixture components. This confirms that speaker verification for adults using clean, wide-band speech is a relatively easy task [19]. For children the best performance is 0.8% EER, also obtained with 128 mixture components. This indicates that the speaker verification EER for children is nearly four times worse than for adults.




(a)



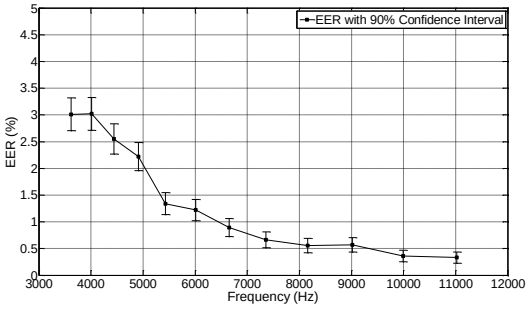

(b)

Figure 1: Speaker verification performance in terms of equal error rate (EER) for adult speech (a) and child speech (b) when using the full bandwidth and various numbers of mixture components.

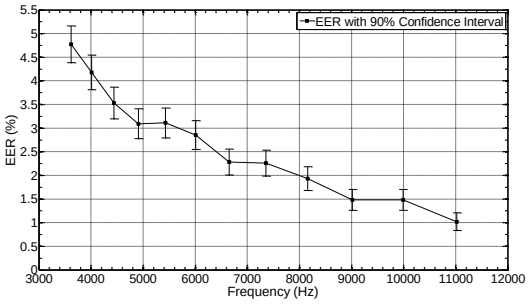
### 4.2. Effect of bandwidth

In this section, we study the effect of bandwidth on verification performance for adults and children. From Section 4.1, it is clear that the EER for the adult data in our study is low when 128 mixture components are used. Therefore, in order to obtain results that are statistically more reliable, the experiments in this section are performed for both adults and children using GMMs with just 32 mixture components. This is consistent with [19, 20], where 32 component GMMs were also used and gave good performance on TIMIT.

To achieve bandwidth reduction the same 32 band-pass filter-bank analysis from the previous experiments was performed, but the vector passed to the DCT for calculation of the cepstral features consisted of different numbers of logarithm filter-bank energies, varying from 21, corresponding to the bandwidth of 3.6kHz, to 32, corresponding to the maximum bandwidth of 11.025kHz.




(a)




(b)

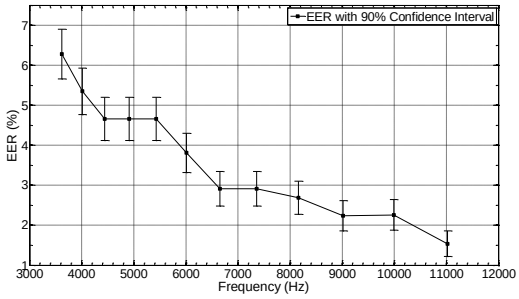
Figure 2: The speaker verification performance in terms of equal error rate (EER) for adult speech (a) and child speech (b) as a function of the bandwidth of speech signal.

Verification results, in terms of EER, for adults and children as a function of the bandwidth are depicted in figure 2. For adults (figure 2(a)), it is evident that it is useful to partition the spectrum into three regions: (i) up to 3.5-4kHz, (ii) 3.6-4kHz to 5.5kHz, and (iii) above 5.5kHz. Region (i), corresponding to the primary resonances of the vocal tract, clearly contains speaker-specific information. However in these experiments there appears to be no benefit from including frequencies above 3.6kHz in this region. Region (ii) contributes a 58% reduction in EER. The importance of this region for speaker verification has been noted in [20], where it is reported that including the effect of the piriform fossa, which is speaker dependent and changes little during speech production, in the speech production model introduces spectral changes in the frequency region between 4kHz and 5kHz.

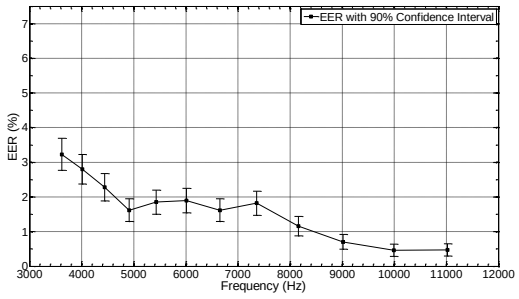
Region (iii) accounts for a further reduction in error rate of 76%, but over a much larger frequency range.

Figure 2(b) shows the corresponding results for children's speech. However, a clearer picture emerges from figures 3(a) and 3(b), where the results for younger children (aged 5 to 9 years) and older children (aged 10 to 15 years) are presented separately. For younger children, the boundary for region (i) appears to be between 4.5 and 5.5kHz. In contrast with the case for adults, there is useful information in the 3.6 to 4.5kHz region, presumably because the primary vocal tract resonances occur at higher frequencies for children, with smaller vocal tracts. Region (ii) lies between 4.5-5.5kHz and 6.5kHz, approximately 1kHz higher than for adult speech, and contributes a 37% reduction in EER. It would be interesting to know if this is consistent, in terms of physiology, with the corresponding result for adult speech. Region (iii), comprising frequencies above 6.5kHz, contributes a 64% reduction in EER over 4.5kHz.

The results for older children (figure 3(b)) are similar to those for adults.




(a)




(b)

Figure 3: Speaker verification performance in terms of equal error rate (EER) for younger children aged from 5 to 9 years (a) and older children aged from 10 to 14 years (b) as a function of the bandwidth of speech signal.

## 5. Conclusions

In this paper, we compared speaker verification performance for adults and children and, in both cases, investigated the effects of bandwidth on EER. We found that, as in the case of automatic speech recognition, verification performance is significantly poorer for children than for adults, with best EERs for children and adults of 0.8% and 0.22%, respectively. This suggests that any advantage stemming from increased inter-speaker variability in children is countered by the increase in intra-speaker variability.

Turning to bandwidth, we found, as reported elsewhere, that in terms of its contribution to speaker verification performance, the spectrum can be usefully partitioned into three frequency bands. For adult speech these are: (i) up to 3.5-4kHz, (ii) 3.5-4kHz to 5.5kHz, and (iii) above 5.5kHz. Similar bands occur for child speech, but with boundaries that are approximately 1kHz greater than for adults.

We conclude that speaker verification and speech recognition for child speech pose similar challenges.

## 6. References


[1] "Engaging with social networking sites," Online: http://stakeholders.ofcom.org.uk/market-data-research/media-literacy/medlitpub/medlitpubrss/socialnetworking/summary/, accessed on 28 March 2011.

[2] Rose, R. C. and Reynolds, D. A., "Text independent speaker identification using automatic acoustic segmentation," in Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, vol. 1, pp. 293-296, 1990.

[3] Reynolds, D. A., Quatieri, T. F. and Dunn, R. B., "Speaker verification using adapted Gaussian mixture models," Digital Signal Processing, vol. 10, pp. 19-41, 2000.

[4] Reynolds, D. A., "Speaker identification and verification using Gaussian mixture speaker models," Speech Commun., vol. 17, pp. 91-108, 1995.

[5] Campbell, W. M., Sturim, D. E., Reynolds, D. A. and Solomonoff, A. "SVM based speaker verification using a GMM supervector kernel and NAP variability compensation," ICASSP, pp. 312-322, 2006.

[6] Bimbot, F., Bonastre, J. F., Fredouille, C., Gravier, G., Magrin-Chagnolleau, I., Meignier, S., Merlin, T., Ortega-García, J., Petrovska-Delacrétaz, D. and Reynolds, D. A., "A tutorial on text-independent speaker verification," EURASIP Journal on Applied Signal Processing, vol. 2004, pp. 430-451, 2004.

[7] Doddington, G. "Speaker recognition based on idiolectal differences between speakers," in Proc. Of Eurospeech, pp. 2521-2524, 2001.

[8] Nittrouer, S. and Whalen, D., "The perceptual effects of child-adult differences in fricative-vowel coarticulation," J. Acoust. Soc. Am., vol. 86, pp. 1266-1276, 1989.

[9] Lee, S., Potamianos, A. and Narayanan, S., "Acoustics of children's speech: Developmental changes of temporal and spectral parameters," J. Acoust. Soc. Am., vol. 105, pp. 1455-1468, 1999.

[10] Gerosa, M., Lee, S., Giuliani, D. and Narayanan, S., "Analyzing children's speech: An acoustic study of consonants and consonant-vowel transition," in Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing, vol. I, pp. 393-396, 2006.

[11] Li, Q. and Russell, M. J., "An analysis of the causes of increased error rates in children's speech recognition," in Proc. of Int. Conf. on Spoken Language Processing, 2002.

[12] Yildirim, S., Narayanan, S., Boyd, D. and Khurana, S., "Acoustic analysis of preschool children's speech," in Proc. 15th ICPhS, pp. 949–952, 2003.

[13] Batliner, A., Blomberg, M., D'Arcy, S., Elenius, D., Giuliani, D., Gerosa, M., Hacker, C., Russell, M., Steidl, S. and Wong, M., "The PF STAR children's speech corpus," in Proc. of Interspeech, pp. 2761-2764, 2005.

[14] D'Arcy, S., Russell, M. J., Browning, S. R. and Tomlinson, M. J., "The Accents of the British Isles (ABI), corpus," pp. 115-119, 2004.

[15] Pelecanos, J. and Sridharan, S., "Feature warping for robust speaker verification," in Proc. of Speaker Odyssey, 2001.

[16] Martin, A., "2003 Speaker Recognition Evaluation," pp. 1, 2008.

[17] Martin, A., Doddington, G., Kamm, T., Ordowski, M. and Przybocki, M., "The DET curve in assessment of detection task performance," in Proc. of Eurospeech, pp. 1895-1898, 1997.

[18] Bolle, R. M., Pankanti, S. and Ratha, N. K., "Evaluation techniques for biometrics-based authentication systems (FRR)," in Proc. of Int. Conf. on Pattern Recognition, pp. 835-841, 2000.

[19] Reynolds, D. A., "Large population speaker identification using clean and telephone speech," IEEE Signal Processing Letters, vol. 2, pp. 46-48, 1995.

[20] Safavi, S., Jancovic, P., Russell, M,J. and Carey, M,J., " Speaker recognition for children's speech," Interspeech 2012., Portland, USA, pp. 1836-1839, 2012.

[21] Safavi, S., Russell, M,J. and Jancovic, P., "Identification of age-group from children's speech by computers and humans," Interspeech 2014., Singapour, 2014.
[22] Lu, X. and Dang, J., "An investigation of dependencies between frequency components and speaker characteristics for text-independent speaker identification," Speech Commun., vol. 50, pp. 312-322, 2008.
[23] Safavi, S., Jancovic, P., Russell, M,J. and Carey, M,J., "Identification of gender from children's speech by computers and humans," Interspeech 2013., Lyon, France, pp. 2440-2444, 2013.