Identification of Gender from Children’s Speech by Computers and Humans

Saeid Safavi, Peter Jančovič, Martin Russell, Michael Carey

School of Electronic, Electrical & Computer Engineering, University of Birmingham, UK
{sxs796, p.jancovic, m.j.russell, m.carey}@bham.ac.uk

Abstract
This paper presents results on gender identification (GI) for children’s speech, using the OGI Kids corpus and GMM-UBM and GMM-SVM systems. Regions of the spectrum containing important gender information for children are identified by conducting GI experiments over 21 frequency sub-bands. Results show that the frequencies below 1.8 kHz and above 3.8 kHz are most useful for GI for older children, while the frequencies above 1.4 kHz are most useful for the youngest children. The effect of using age-independent and age-dependent gender modelling (including the effects of puberty on boys voices) is explored. The application of intersession variability compensation is explored but experiments showed only little improvement. Experiments on human GI were also conducted and the results show that the humans do not achieve the performance of the machine.

Index Terms: paralinguistic speech processing, gender identification, child speech, gaussian mixture model, support vector machine, intersession variability compensation, frequency band

1. Introduction
In addition to its linguistic content, the acoustic speech signal also contains paralinguistic information, such as the speaker’s identity, accent, gender, age, or emotional state. Automatic recognition of such paralinguistic information can be useful in many application areas. For instance, it can be used to adapt speech models to the speaker, to guide a human computer interaction system to automatically adapt to different user needs, to enhance personal security and protection or in education.

Research effort into paralinguistic speech processing has been growing considerably over the last two decades. It has initially focused mainly on speaker recognition from adults’ speech, e.g., [1], but more recently also spread to speaker recognition for children’s speech, e.g., [2], recognition of accent, e.g. [3, 4, 5], emotions, age and gender. A recent review of paralinguistic speech processing is presented in [6]. Some earlier research on adult speech gender recognition demonstrated high performance [7, 8]. The task of age and gender recognition also became popular in last few years [9, 10, 11]. Research focused on exploring the use of features capturing different types of information from the speech signal and the use of different classification systems. Most studies employed mel-frequency cepstral coefficients (MFCCs). The use of TRAPS features to capture longer temporal context was explored in [10]. Several studies also considered the use of glottal and prosodic features. These were typically calculated on the whole utterance and included features such as the fundamental frequency, jitter/shimmer, articulation rate, and harmonic-to-noise ratio [12, 9, 10]. The latter could also be provided by estimating the spectral voicing information using the method presented in [13, 14]. Overall, the use of MFCC features, capturing vocal-tract information, was shown to provide best performance, which could be further improved by incorporating other features or combining multiple classifiers. The use of various classification approaches for age and gender identification has been explored. Early studies employed distance measures [7], Gaussian mixture model (GMM) and hidden Markov model (HMM) based recognisers [8]. More recently, the success of GMM - Universal Background Model (GMM-UBM) and GMM-Support Vector Machine (GMM-SVM) approaches to adult speaker recognition motivated its application to the age and gender recognition task [15]. The distribution of acoustic feature vectors for a population of speakers, is typically captured using a UBM, a gender/age-independent GMM constructed using data from a variety of speakers and background conditions. Age/gender dependent GMMs are then built either by using the UBM as an initial model in a standard EM training or by MAP adaptation of the UBM. In the GMM-SVM approach, the combination of GMM supervectors (comprising the stacked parameters of the GMM components) with SVMs combines the strengths of the generative GMM-UBM and the discriminative power the SVM. The GMM-UBM and GMM-SVM approaches have also been compared to the use of GMM and parallel phoneme recognition system [9, 15, 16]. The use of dynamic Bayesian networks employing prosodic features was explored in [9, 16]. Furthermore, techniques such as cepstral mean subtraction and variance normalisation have been applied to speaker age and gender identification tasks to enhance the performance of acoustic level modeling [16]. In all these studies, age and gender were recognised jointly, with a broad set of age classes corresponding to children, young adults, adults and seniors. The gender was not considered for the children class.

It has been shown that acoustic and linguistic characteristics of childrens speech are different from those of adults [17] [18] [19]. For example, childrens speech is characterised by higher pitch, and perceptually important features such as formants occur at higher frequencies [18]. Consequently, the impact of bandwidth reduction on speech recognition accuracy is greater for childrens speech than for adults [20] [21]. However, we do not know the significance of different frequency bands for gender identification for children.

This paper presents the results of experiments on gender identification (GI) from children’s speech and is organized as follows. Section 2 describes the OGI Kid’s corpus of children’s speech, which is used in all experiments. Our GI systems are described in section 3, and our experiments and results are presented in section 4. First, we describe a study of the utility of the information in different frequency bands for childrens’ gender identification. Results show that the frequencies below 1.8 kHz and above 3.8 kHz are most useful for GI for older children, while the frequencies above 1.4 kHz are most useful for the youngest children. In the case of older children, the use of only a narrow frequency sub-band can provide performance similar
to the use of full bandwidth. Next, we explore the effect of using age-independent and age-dependent gender modelling and GMM-UBM and GMM-SVM modelling. Results show that the GMM-SVM system outperforms considerably the GMM-UBM model in most cases. The use of age-dependent modelling can provide further substantial GI performance improvements, especially for the GMM-UBM system. We also analyse the effect on GI of the voice breaking for children in the oldest group and show that having a separate models for boys with broken and unbroken voice can provide substantial improvements. Further, we present the effect of employing intersession variability compensation. Results indicate only relatively small improvement for GI task. Finally, we present a comparison with human GI performance. Results show that considerably better performance is obtained by the automatic system than by humans.

2. The OGI kids speech corpus and data description

The OGI Kids Speech corpus [22] is a collection of spontaneous and read speech recorded at the Northwest Regional School District near Portland, Oregon. The CSLU Toolkit is used for data collection. It comprises recordings of words and sentences from approximately 1100 children. A gender-balanced group of approximately 100 children per grade from Kindergarten (5-6 year olds) through to grade 10 (15-16 year olds) participated in the collection. For each utterance, the text of the prompt was displayed on a screen, and a human recording of the prompt was played, in synchrony with facial animation using the animated 3D character “Baldi”. The subject then repeated the prompt, which was recorded via a head-mounted microphone and digitized at 16 bits and 16 kHz.

Three different test sets from the OGI data are used in the experiments presented in this paper.

TS1: To investigate the effect of different frequency bands on GI performance for general childrens speech, 687 speakers were chosen randomly (kindergarten to 10th grade).

TS2: To investigate the effect of different frequency bands and using age dependent model on GI performance for speech from children of different ages, 3 different age groups were selected, each containing 76 speakers. These are AG1: kindergarten to 3rd grade (5-9 year olds), AG2: 4th to 7th grade (9-13 year olds), and AG3: 8th to 10th grade (13-16 year olds).

TS3: To investigate the effect on gender identification of the voice breaking for male speakers going through puberty, the data in AG3 group were split into three sub-sets, denoted as ‘boys broken’, ‘boys unbroken’ and girls. Each sub-set contained data from 18 speakers for training. For testing, all 191 speakers from AG3 (both male and female) were used.

3. Gender identification systems

3.1. Signal analysis

Feature extraction was performed as follows. Periods of silence were discarded using an energy-based Speech Activity Detector (SAD). The speech was then segmented into 20-ms frames (10-ms overlap) and a Hamming window was applied. The short-time magnitude spectrum, obtained by applying an FFT, is passed to a bank of 24 Mel-spaced triangular bandpass filters, spanning the frequency region from 0 Hz to 8000 Hz. Table 1 shows the center frequency of each filter (the cut-off frequencies of a filter are the centre frequencies of the adjacent filters). To investigate the effect of different frequency regions on GI performance, experiments were conducted using frequency band limited speech data comprising the outputs of groups of 4 adjacent filters. We considered 21 overlapping sub-bands, where the Nth sub-band comprises the outputs of filters N to N + 3 (N=1 to 21). Each set of 4 filter outputs was transformed to 4 Mel Frequency cepstral coefficients (MFCCs) plus 4 delta and 4 delta-deltas, and feature warping [23] was applied. For the full bandwidth experiments the outputs of all 24 filters were transformed into 19 MFCCs plus 19 deltas and 19 delta-deltas.

| Table 1: The Center Frequencies for 24 Mel-spaced Band-Pass Filters |
|--------------------------|----------|--------------------------|----------|
| FILTER NUMBER | CENTER FRE. (Hz) | FILTER CENTER FRE. (Hz) |
| 1 | 156 | 13 | 3843 |
| 2 | 281 | 14 | 2062 |
| 3 | 406 | 15 | 2343 |
| 4 | 500 | 16 | 2656 |
| 5 | 625 | 17 | 3000 |
| 6 | 750 | 18 | 3375 |
| 7 | 875 | 19 | 3812 |
| 8 | 1000 | 20 | 4312 |
| 9 | 1125 | 21 | 4906 |
| 10 | 1281 | 22 | 5531 |
| 11 | 1437 | 23 | 6281 |
| 12 | 1625 | 24 | 7093 |

3.2. Modelling

Our GI systems are based on the GMM-UBM [1, 24] and GMM-SVM [24] methods.

In the GMM-UBM approach, a UBM is built using all utterances from 418 speakers. The age-independent gender models are obtained by MAP adaptation (adapting means only) of the UBM, using the gender-specific enrollment data. The result is one UBM and 2 gender GMMs. To investigate the effect of using age-dependent gender models on GI performance, age-dependent models are obtained by MAP adaption of the UBM, using the age and gender-specific enrollment data.

In our GMM-SVM system, the speech data from each gender category (and also age category in the case of using the age-dependent models) were used to estimate the parameters of a GMM by MAP adaptation of the UBM. The adapted GMM mean vectors are then concatenated into a supervector [24], and the gender classes are assumed to be linearly separable in the supervector space. The supervectors are used to build one SVM for each gender/age class, by treating that gender/age class as the target class and the others as the background class.

4. Experimental results and discussion

4.1. Isolated frequency sub-bands GI for children’s speech

In this section, we study the effect of different sub-bands on GI performance for childrens speech. Experiments are conducted separately on 21 sub-bands, each consisting of four consecutive channels (see Section 3.1 for more details), and using the age-dependent GMM-UBM system. For each sub-band, three age-dependent models of each gender are trained. The models have 64 mixture components, which was found to be adequate for these 12 dimensional sub-band features.

Figure 1(a) presents the average results over all age groups.
It is evident that the performance even when using a narrow frequency region is in most cases well over chance. Further, it can be seen that the most useful sub-bands for GI are from 8 to 12 (frequency range 0.9 kHz–2.6 kHz). This corresponds to the location of the second formant for vowels, which was also found to provide the best GI performance for adult speakers in [25]. Figure 1(b) shows the performance for each age group. It can be seen that the trend is different for each age group. For AG3 (i.e., the oldest children), the frequency sub-bands up to 9 (frequencies up to 1.8 kHz) and above 19 (frequencies above 3.8 kHz) provide somewhat similar performance of around 75%, while the middle frequency sub-bands give lower performance. For AG2, the performance does not vary largely across the frequency bands. The peak performance is achieved at sub-bands 9 and 10 (frequency range 1.0 kHz–2.1 kHz). For AG1, the performance is close to chance for sub-bands up to 7 (frequencies up to 1.4 kHz) and then increases to around 65% for sub-band 11 and 12 and stays fluctuating around 60% for the remaining higher sub-bands. It may be that the insignificance of sub-bands up to 7 for young children, and their increasing utility as the age of the child increases is due to greater and more consistent vocal effort in older children.

4.2. Full-bandwidth GI for children's speech

This section demonstrates the effect of using different modelling approaches. Experiments were performed using the full bandwidth speech.

4.2.1. Age-independent modelling

First, we demonstrate the effects of employing the generative GMM-UBM and discriminative GMM-SVM systems when using age-independent modelling. For each of the systems, we performed experiments using different numbers of mixture components. The best results were obtained when using 1024 and 512 mixture components for the GMM-UBM and GMM-SVM system, respectively, and these are presented in Table 2. It can be seen that the GMM-SVM system outperforms considerably the GMM-UBM.

<table>
<thead>
<tr>
<th>System</th>
<th>GI rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GMM-UBM (age independent)</td>
<td>67.39</td>
</tr>
<tr>
<td>GMM-SVM (age independent)</td>
<td>77.44</td>
</tr>
</tbody>
</table>

4.2.2. Age-dependent modelling

This section demonstrates the effect of using age-dependent modelling, in which all training data is split into three age groups (as described in Section 2) and a model is created for each age group. During recognition, models corresponding to the age of the speaker of the testing utterance were used. The results of the experiments are presented in Table 3. It can be seen that the performance of the GMM-UBM and the GMM-SVM system improved by 4.23% and 1.74%, respectively, in comparison to corresponding age-independent systems.

<table>
<thead>
<tr>
<th>System</th>
<th>GI rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GMM-UBM (age dependent)</td>
<td>71.76</td>
</tr>
<tr>
<td>GMM-SVM (age dependent)</td>
<td>79.18</td>
</tr>
</tbody>
</table>

We now analyse the results obtained by the age-dependent GMM-UBM and GMM-SVM systems for each age group. These are presented in Table 4, in which ‘B’, ‘G’, and ‘Av’ denotes boys, girls and average. One would expect the GI performance to be lowest for youngest children, i.e. AG1, and to improve as the age increases. Indeed, we can see that the identification rate achieved by each system for AG2 is considerably higher than for AG1 – the performance increase is 12.81% for the GMM-UBM and 7.34% for the GMM-SVM system. However, the performance is unexpectedly low for AG3 (i.e., the oldest children) – in comparison to AG2, the performance improves only by 2.52% for the GMM-UBM and decreases by 8.23% for the GMM-SVM system. This may be related to the fact that the boys in AG3 fall into two subsets, according to whether or not their voices have broken as a consequence of puberty. It is possible that the GMM-UBM system is better able to accommodate this than the GMM-SVM system.

We further analyse the effect of voice breaking in AG3. This is performed using the GMM-UBM system. Table 5 (a)
shows the GI confusion matrix for AG3 when using a single model for boys and a single model for girls. It can be seen that there is a high confusion for the gender of boys being recognised as girls. We speculate that this is because the model for ‘boys’ covers broken and unbroken voices, and consequently some boys whose voices have not broken may achieve a better match with the ‘girls’ speech model. These results, and the fact that changes in the voice coinciding with puberty is prominent mainly in boys rather than in girls, motivated us to split the data of boys in AG3 into two separate classes: boys whose voice has broken, denoted as B_B, and whose voice is still unbroken, denoted as B_U. The labeling of the boys data into these two classes was performed by a human listener. Some further details on the resulting training and testing data are described in Section 2, where it is denoted as TS3. Since this resulted in reduced amounts of training data for each class, the GMM-UBM system for each of the AG3 gender sub-group consisted of 128 mixture components. When using the three gender subgroup models, the average GI rate for AG3 was 87.43%. This is an improvement of 8.9% from 78.53% achieved by the system using two gender models (and each consisting of 256 mixture components) as presented in Table 4. Table 5 (b) presents the confusion matrix corresponding to this experiment. It can be seen that the amount of gender confusions from boys to girls is reduced to zero. The amount of confusion from girls to boys with unbroken voices is much larger than to boys with broken voices, which is what would be expected.

<table>
<thead>
<tr>
<th>Age group</th>
<th>GMM-UBM</th>
<th>GMM-SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>AG1</td>
<td>40.00</td>
<td>90.00</td>
</tr>
<tr>
<td>AG2</td>
<td>69.67</td>
<td>82.25</td>
</tr>
<tr>
<td>AG3</td>
<td>70.00</td>
<td>88.52</td>
</tr>
</tbody>
</table>

Table 5: Confusion matrix for gender identification (in %) for age group AG3 when using for boys a single model (a) and two separate, broken B_B and unbroken B_U, models (b).

4.2.3. Effect of intersession variability compensation

We also investigated the effect of intersession variability compensation (ISVC). These experiments were performed using the GMM-UBM system, and both age-independent and age-dependent modelling. Both systems achieved only small performance improvements when ISVC was applied, specifically, the age-independent system improved from 67.39% to 69.29% and the age-dependent system improved from 71.76% to 72.81%.

4.3. Human GI for children’s speech

In addition to the computer GI experiments presented in the previous sections, we also performed experiments on GI by human listeners. The test set consisted of the same 687 test utterances used in the computer GI experiments. Twenty listeners participated in the experimental evaluations. Each participant listened to 34 utterances on average. The length of each utterance was 10 seconds. All human listening tests were performed in a quiet room using the same PC and headphones.

The GI rates for each age group achieved by human listeners are presented in Table 6. The average performance over all age groups was 66.96%.

<table>
<thead>
<tr>
<th>Age group</th>
<th>Human GI rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AG1</td>
<td>60.48</td>
</tr>
<tr>
<td>AG2</td>
<td>70.49</td>
</tr>
<tr>
<td>AG3</td>
<td>70.90</td>
</tr>
</tbody>
</table>

5. Conclusions

This paper presented the results of experiments in gender identification (GI) for children’s speech using the OGI kids speech corpus. A study of the utility of different narrow frequency bands has shown that the frequency region 0.9–2.6 kHz is the most useful by average over all age groups. The analysis of the results separately for each of the three age groups showed that the performance trend is different for each age group. For the AG3 children (13–16 years), the frequencies up to 1.8 kHz and above 3.8 kHz provide the best performance which is around 75%. For AG2 children (9–13 years) performance does not vary largely across the frequency bands. For AG1 children (5–9 years), the performance is close to chance up to 1.4 kHz, then increases to 65% for sub-bands around 1.8 kHz and fluctuates around 60% for frequencies above 2.3 kHz. The effect of using age-dependent gender modelling and using the GMM-UBM and GMM-SVM systems was examined using the full-bandwidth experiments. The age-independent GMM-SVM system outperformed the GMM-UBM system by nearly 10%. The age-dependent gender models gave 4.23% and 1.74% GI improvement in the case of the GMM-UBM and GMM-SVM system, respectively. The full-bandwidth results for each age group were analysed. This showed unexpectedly low performance for the AG3 children. An investigation confirmed that this was due to the fact that the ‘boys’ category in AG3 includes both boys with broken and unbroken voices, depending on whether or not the child has entered puberty. Consequently, speech from boys whose voices have not broken may achieve a better match with the ‘girls’ acoustic model. The data of boys in AG3 were divided into two separate groups, boys with broken and unbroken voices. The use of the three gender classes provided a GI rate of 87.34% for the AG3 children, which was an improvement of 8.81% from using only two gender classes. The application of intersession variability compensation was explored but experiments showed only little improvement. Human GI experiments were also conducted and the average performance over all age groups was 66.96%, which is lower than the performance achieved by machine.

6. References

speech, 2012.
[6] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan, “Paralinguistics in speech and lan-
[8] E. S. Parris and M. J. Carey, “Language independent gender iden-
tification,” IEEE Int. Conf. on Acoustics, Speech, and Signal Proc-
(ICAASSP), no. 865-688, 1996.
[14] P. Jančovič and M. Köküer, “Incorporating the voicing informa-
tion into HMM-based automatic speech recognition in noisy envi-
1608, 2008.
[16] M. Li, C.-S. Jung, and K. J. Han, “Combining five acoustic level modeling methods for automatic speaker age and gender recogni-
[22] K. Shobaki, J. P. Hosom, and R. A. Cole, “The OGI kids’ speech corpus and recognizers,” Int. Conf. on Spoken Language Process-
ing, 2000.