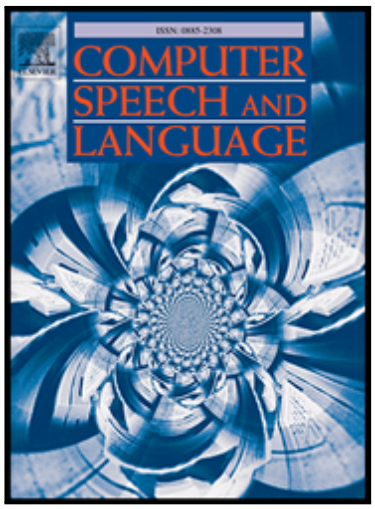
Automatic speaker, age-group and gender identification from children's speech

Safavi, Saeid; Russell, Martin; Jancovic, Peter

DOI:
10.1016/j.csl.2018.01.001

Document Version
Peer reviewed version

Citation for published version (Harvard):
Safavi, S, Russell, M & Jancovic, P 2018, 'Automatic speaker, age-group and gender identification from children's speech', *Computer Speech and Language*, vol. 50, pp. 141-156.
https://doi.org/10.1016/j.csl.2018.01.001



Automatic Speaker, Age-group and Gender Identification from Children's Speech

Saeid Safavi, Martin Russell, Peter Jančovič

PII: S0885-2308(16)30136-X
DOI: 10.1016/j.csl.2018.01.001
Reference: YCSLA 912

To appear in: *Computer Speech & Language*

Received date: 1 May 2016
Revised date: 22 November 2017
Accepted date: 3 January 2018

Please cite this article as: Saeid Safavi, Martin Russell, Peter Jančovič, Automatic Speaker, Age-group and Gender Identification from Children's Speech, *Computer Speech & Language* (2018), doi: 10.1016/j.csl.2018.01.001

Highlights

- First systematic study of acoustic approaches to speaker, gender and age-group identification from speech for children
- First systematic study of the utility of different frequency sub-bands for speaker, gender and age-group identification from speech for children.
- First systematic study of the utility of different acoustic classification techniques, namely GMM-UBM, GMM-SVM and i-vectors, to speaker, gender and age-group identification from speech for children.
- First exploration of the effects of age- and gender-dependent modelling and the onset of male puberty on automatic speaker, gender and age-group identification

Automatic Speaker, Age-group and Gender Identification from Children's Speech

Saeid Safavi, Martin Russell and Peter Jančovič

Department of Electronic, Electrical and Systems Engineering, School of Engineering, University of Birmingham, Birmingham, B15 2TT UK

Abstract

A speech signal contains important paralinguistic information, such as the identity, age, gender, language, accent, and the emotional state of the speaker. Automatic recognition of these types of information in adults' speech has received considerable attention, however there has been little work on children's speech. This paper focuses on speaker, gender, and age-group recognition from children's speech. The performances of several classification methods are compared, including Gaussian Mixture Model - Universal Background Model (GMM-UBM), GMM - Support Vector Machine (GMM-SVM) and i-vector based approaches. For speaker recognition, error rate decreases as age increases, as one might expect. However for gender and age-group recognition the effect of age is more complex due mainly to consequences of the onset of puberty. Finally, the utility of different frequency bands for speaker, age-group and gender recognition from children's speech is assessed.

Keywords: Speaker recognition, gender identification, age-group identification, children's speech, Gaussian Mixture Model (GMM), Support Vector Machine (SVM), i-vector.

1. Introduction

Speech signals contain significant information in addition to their linguistic content, for example information about the speaker's identity, gender, social group, geographical origin, health, smoking habit, height and emotional state. This type of information is referred to as paralinguistic. While speech recognition is concerned with extracting the underlying linguistic message in a speech signal, paralinguistic analysis is concerned with extracting

its paralinguistic information.

For a subset of paralinguistic technologies, including speaker, language, accent and dialect recognition, a set of common tools has been developed. Analysis typically begins by transforming a speech signal into a sequence of spectrum-based feature vectors, for example mel frequency cepstral coefficients (MFCCs) (Davis and Mermelstein (1980)). In the classification stage it is useful to distinguish between acoustic approaches, which assume that a particular class is distinguished by its distribution of acoustic feature vectors, and phonotactic approaches, which exploit higher-level sequential structure (Schuller et al. (2013)). Phonotactic and acoustic methods provide complementary information, which state-of-the-art systems exploit by fusing their output scores (Wong and Sridharan (2001)). A more detailed review of these methods is given in Section 2.

Although there has been considerable work on paralinguistic analysis of adults' speech (Schuller and Batliner (2013); Schuller et al. (2013)), there is relatively little work on this topic for children's speech. The most notable exception is the 2009 paralinguistic challenge (Schuller et al. (2011)) held at Interspeech 2009, which focused on emotion recognition for the German FAU Aibo Emotion Corpus (Batliner et al. (2004)) of spontaneous speech from 10 to 13 year old German children interacting with an AIBO robotic dog. Another example is early work on automatic assessment of goodness of pronunciation in children's speech (Russell et al. (2000)).

The paucity of work in this area on children's speech is surprising because paralinguistic analysis has a number of significant applications that are especially relevant for children. For example, voice-based speaker, gender and age recognition offer novel approaches to safeguarding children engaged in social networking, gaming or general internet browsing. Also, speaker recognition provides an alternative means to identify which child is using a particular piece of educational software or whether a child is present in a school or class.

Children's speech poses particular problems. It is well documented (for example (Lee et al. (1997, 1999); Gerosa et al. (2006, 2007))) that spectral and durational variability are greater for children's speech than for adults' speech, converging to adult levels at around the age of 13 years. However, it is not known whether this is mainly due to developing motor control skills or cognitive issues relating to language acquisition. Because of their shorter vocal tracts, important structure in speech occurs at higher frequencies for children than for adults. Consequently the effects of bandwidth reduction on computer and human speech recognition performance are much more pro-

nounced for children's speech (Russell et al. (2007)). In addition, children's higher fundamental frequencies can result in greater variability in spectrum-based feature vectors, such as MFCCs, used in automatic speech recognition (ASR)(Ghai (2011)).

Unlike paralinguistic analysis, ASR for children's speech has been an active research topic for the past 20 years. Due to the factors cited above, recognition error rates are typically much greater for children's speech, especially for young children, even if the recogniser is trained on age-matched data (Wilpon and Jacobsen (1996)). The availability of ASR systems trained on adult speech and the absence of large corpora of children's speech, has motivated research into the use of adaptation techniques such as vocal tract length normalisation (VTLN) (Lee and Rose (1998)) and pitch normalisation (Ghai (2011)), as well as generic adaptation techniques such as MAP and MLLR adaptation (Elenius and Blomberg (2005); Gerosa et al. (2005)), to enable adult ASR systems to recognise children's speech. In the most recent research, deep neural network (DNN) based ASR has been applied to children's speech (Liao et al. (2015); Serizel and Giuliani (2016)).

In addition to the above, another significant factor that complicates the identification of age-group and gender for older children is the onset of puberty. Rogol et al. (Rogol et al. (2000)) observed wide variations in the timing of the onset of puberty and in the consequent physiologic variations in growth. They estimate the age of the onset of puberty for girls and boys as 11 and 13 years, respectively.

The objectives of the work described in this paper are to establish baselines for paralinguistic analysis of children's speech, focussing on the application of state-of-the-art methods to the identification of a child's identity, gender, and age-group, and to investigate the importance of information in different frequency sub-bands for these tasks.

The paper is organised as follows. Section 2 presents a review of approaches to paralinguistic analysis of adult speech, including speaker, language, regional accent and dialect recognition. Section 3 describes the data that was used and how it was partitioned for speaker, gender and age-group recognition. This section also describes feature extraction and the classification methods that were used. Section 4 investigates the utility of different frequency sub-bands for these technologies. Section 5 presents the results of speaker, gender and age-group recognition experiments with 'full bandwidth' children's speech, and Section 6 presents our conclusions.

2. Paralinguistic Analysis of Adult Speech

The simplest acoustic approach to speaker and language recognition for adults' speech is to model the distribution of feature vectors for a particular class as a Gaussian Mixture Model (GMM) (Reynolds and Rose (1995); Naik et al. (1989)). First, a class-independent model, λ_{UBM} called the Universal Background Model (UBM) is estimated using all of the available training data. A class-dependent GMM λ_c is then created for each class c from the UBM using training data for c and MAP (Gauvain and Lee (1994)) adaptation. In the GMM-UBM approach, classification of an utterance u is based on the probabilities $p(u|\lambda_c)$. In speaker and language identification this probability is used directly, whereas for speaker verification $p(u|\lambda_s)$ is divided by $p(u|\lambda_{UBM})$, and the result is compared with a threshold. Score normalisation techniques, for example T-norm (Auckenthaler et al. (2000)), Z-norm (Wan and Campbell (2000)), and max-norm (Campbell and Karam (2009)) enable the same threshold to be used for all speakers.

The GMM-UBM approach has been superseded by methods that map an utterance u into a vector space and then apply a static classifier such as a Support Vector Machine (SVM) (Campbell et al. (2006)). For example, MAP adaptation of the UBM is performed using u, and the mean vectors of the components of the resulting GMM are concatenated to form a GMM 'supervector' (Campbell et al. (2006)). Inter-session variability, including inter-channel variability or channel factors, is a significant source of error for these techniques. Intersession variability compensation (ISVC) (Vair et al. (2006)) and nuisance attribute projection (NAP) (Solomonoff A. and Campbell (2004)) address this issue by isolating this variability in a linear subspace of the supervector space. In more recent approaches based on Joint Factor Analysis (JFA) (Kenny et al. (2007); Dehak et al. (2011a)), and i-vectors (Dehak et al. (2011b)), the GMM supervector is mapped into a lower-dimensional discriminative total variability space. These methods have been applied to a number of problems, including identification of speaker (Campbell et al. (2009); Kinnunen and Li (2010)), language, regional accent and dialect (Biadsy et al. (2011); DeMarco and Cox (2012); Hanani et al. (2013)), and age and gender (Metze et al. (2007); Bocklet et al. (2008); Dobry et al. (2009); Mahmoodi et al. (2011); Chen et al. (2011); Van Heerden et al. (2010); Bahari and Van Hamme (2011); Bahari et al. (2012)).

Different frequency regions contain different types of linguistic and paralinguistic information. For instance, frequency regions below 600 Hz and

above 3000 Hz are the most informative for speaker identification from adults' speech (Besacier et al. (2000); Safavi et al. (2012); Orman and Arslan (2001)).

In phonotactic approaches, a spoken utterance is transformed into a sequence of tokens from a finite set, and it is assumed that different classes are distinguishable from the statistics of these sequences. In the Phone Recognition followed by Language Modelling (PRLM) and Parallel PRLM (PPRLM) approaches to language identification of Zissman (Zissman (1996)), an automatic phone recognition system transforms the utterance into a sequence of phones, which is analysed using language-dependent statistical language models. In (Hanani et al. (2013)) the tokens are GMM component indices and the component n-gram statistics are collated in a vector which is classified using language-dependent SVMs. Similar methods are employed in the 'ideolect' approach to speaker verification (Doddington (2001)).

Recently, deep neural networks (DNNs) have been employed for speaker (Yamada et al., 2013; Lei et al., 2014; Variani et al., 2014; Sarkar et al., 2014) and language recognition (Song et al., 2013; Lopez-Moreno et al., 2014). One approach is to use a DNN directly as a classifier to discriminate between classes, with the frame-level DNN posteriors being averaged over the duration of the utterance (Lei et al., 2014). An alternative is to use a DNN to extract bottleneck features or posteriors which are then used to train an i-vector classifier (Song et al., 2013; Sarkar et al., 2014).

3. Method

3.1. Speech data

The OGI Kids Speech corpus (Shobaki et al. (2000)) is used in all experiments. It contains examples of spontaneous and read speech from approximately 1100 children (roughly 100 children per grade from kindergarten (5–6 years old) to grade 10 (15–16 years old)), recorded in Portland, Oregon. Prompts were displayed as text on a screen, and a human recording of the prompt was played in synchrony with facial animation using the animated 3D character Baldi (Massaro (1998)). The subject repeated the prompt (read speech) or talked on a chosen topic (spontaneous speech). Recordings were made at 16 kHz sampling rate, 16 bits precision, using a head-mounted microphone. An average session lasted 20 minutes and yielded approximately 63 utterances (8-10 minutes of speech). However, for some subjects much less data is available. In speaker recognition experiments just 144 and 48 seconds

of speech were used per subject for training and testing, respectively, for consistency between subjects. The numbers, ages and genders of the children recorded in the corpus are shown in Table 1.

Table 1: Age-range, gender split and number of children recorded per grade in the OGI Kids Speech Corpus.

Grade	K	1	2	3	4	5	6	7	8	9	10
Age	5-6	6-7	7-8	8-9	9-10	10-11	11-12	12-13	13-14	14-15	15-16
Male	39	58	53	63	47	49	57	46	49	70	76
Fem.	50	31	61	54	45	49	55	51	50	40	30
Total	89	89	114	117	92	98	112	97	99	110	106

3.2. Partitioning of the data

Understanding how the performance of paralinguistic technologies varies with age is another objective of this work. Because only a relatively small amount of data is available for each grade, grades are combined into three age-groups, referred to as AG1, AG2 and AG3. The grade that is likely to differ most from the others is K, because the speech of 5–6 year-olds may exhibit additional variability due to phonological factors associated with language acquisition (Fringi et al. (2015)). Hence the first age-group, AG1, should span as few grades as possible. For speaker recognition, AG1 comprises grades K, 1 and 2. However, because there is relatively little speech from some subjects in grades K, 1 and 2, for gender and age-group identification grade 3 was included in AG1 and the boundaries for AG2 and AG3 were shifted by one grade. Table 2 summarises the allocation of grades to age-group for the different experiments.

Table 2: Definitions of the age-groups used in the experiments.

Experiment	AG1	AG2	AG3
Speaker recognition	K–2	3–6	7–10
Gender identification	K–3	4–7	8–10
Age-group identification	K–3	4–7	8–10

Different experiments require different partitions of the corpus into training and test sets, to achieve a suitable balance between classes or to study

a particular effect. For speaker recognition, different configurations are investigated, but in general 144 seconds of speech per speaker are used for training, and segments of length 10 seconds are used for testing. For gender identification the test set comprises utterances from 687 subjects, each of length 30 seconds, balanced by age and gender, and all recordings from the remaining 413 speakers are used for training. For the age-group identification task a gender balanced test set of recordings from 766 speakers (290, 285 and 191 from AG1, AG2 and AG3, respectively), each of length 30 seconds, is used. The recordings from the remaining 344 speakers were used for training. Details are included in the descriptions of the individual experiments. For building the UBM all the training data available in the corpus is used and we have not used any development set. This is summarized in Table 3.

Table 3: Details of training and test data used in the experiments. s/speaker indicates the approximate duration of speech (in seconds) per speaker. FB, SB(a) and SB(b) stand for full-band, gender-independent sub-band and gender-dependent sub-band experiments, respectively.

	Training set		Test set	
	Num. of speakers	Duration (s/speaker)	Num. of speakers	Duration (s/speaker)
Speaker rec. (FB)	864	144	864	10
Speaker rec. (SB(a))	359	144	359	10
Speaker rec. (SB(b))	864	144	864	10
Speaker rec. (Class)	30	144	30	10
Gender ID (FB)	413	All	687	30
Gender ID (SB)	413	All	687	30
Age-group ID (FB)	344	All	766	30
Age-group ID (SB)	344	All	766	30

3.3. Signal analysis

Silence was discarded using energy-based speech activity detection. Frames of length 20 ms (10 ms overlap, Hamming window) were extended to 512 samples and a DFT was applied, giving a frequency resolution of 31.25 Hz. The resulting magnitude spectrum was passed to a bank of 24 Mel-spaced triangular filters, spanning frequencies from 0 Hz to 8000 Hz. Table 4 shows the

Table 4: The center frequencies for 24 Mel-spaced critical-band filters.

Filter Number	1	2	3	4	5	6	7	8
Center Frequency	156	281	406	500	625	750	875	1000
Filter Number	9	10	11	12	13	14	15	16
Center Frequency	1125	1281	1437	1625	1843	2062	2343	2656
Filter Number	17	18	19	20	21	22	23	24
Center frequency	3000	3375	3812	4312	4906	5531	6281	7093

centre frequency of each filter, quantized to the nearest 31.25 Hz (the cut-off frequencies of a filter are the centre frequencies of the adjacent filters). For the full bandwidth experiments the outputs of all 24 filters were transformed into 19 static plus 19 Δ and 19 Δ^2 MFCCs.

3.4. Classification

3.4.1. Automatic classification

Three types of classifier are considered, GMM-UBM (Reynolds et al. (2000); Campbell et al. (2006)), GMM-SVM (Campbell et al. (2006)) and i-vector (Singer et al. (2012); Dehak et al. (2011b)) based methods. All identification and verification experiments were text-independent.

In the GMM-UBM approach (Reynolds et al. (2000)), class-dependent GMMs are obtained by MAP adaptation (Gauvain and Lee (1994)) of the component mean vectors of the UBM ("class" is speaker, gender or age-group), using 144 seconds of data per speaker for speaker recognition, and all of the training data for gender and age-group identification. Speaker verification followed the NIST methodology (NIST (2010)), whereby a test utterance is compared with eleven models, the "true" model plus ten randomly selected impostor models. For age-group identification, gender-dependent and gender-independent models are created. For speaker verification the scores were normalised using "max-norm" (Campbell and Karam (2009)).

In the GMM-SVM system, the mean vectors of class-specific GMMs were concatenated into supervectors (Campbell et al. (2006)). These supervectors were used to build one SVM per class, treating that class as the 'target' and the others as 'background'. For age and gender recognition, one supervector was created per speaker. In speaker recognition, the training data was split into nine 16 second segments to create 9 supervectors per speaker.

The third approach uses i-vectors. This assumes that a GMM supervector μ can be decomposed as $\mu = m + Tw$, where m is the UBM mean supervector, T is a linear mapping from a low-dimensional 'total variability subspace' W into the supervector space, and $w \in W$ is an 'i-vector', sampled from a standard normal distribution. Two approaches to i-vector scoring were employed. In the first (Singer et al. (2012)), linear discriminant analysis (LDA) is applied to W, and each class c is represented by the mean m_c of length normalized i-vectors for that class in the LDA sub-space. At the recognition stage, the score for each class c is the dot product of the normalized LDA test i-vector with m_c (the cosine of angle between the test i-vector and m_c). This approach is referred to as 'i-vector'. The second i-vector method, applied to speaker recognition, uses probabilistic LDA (PLDA) (Prince and Elder (2007)). Before applying PLDA, the i-vectors are length-normalised and whitened (Garcia-Romero and Espy-Wilson (2011)). Recognition is based on the log-likelihood ratio between same versus different speaker hypotheses (Garcia-Romero and Espy-Wilson (2011)). This is referred to as 'i-vector-PLDA'. In all experiments, the size of the T matrix and the i-vector dimension after LDA were set, empirically, to 400 and 300, respectively. The T matrix was learned using all training data.

In NIST evaluations (NIST (2010)) test utterances are 10, 20 or 30 seconds long. We used 10 second test utterances for speaker identification and verification and 30 second utterances for age and gender identification.

3.4.2. Experiments with human listeners

Human gender and age-group identification performance for children's speech was also measured, using test utterances from the automatic classification experiments. Twenty listeners, mainly research students, participated in the evaluations. Since few of these had children they cannot be considered to be "expert" listeners. Each participant listened to an average of 34 and 38 utterances, each of duration 10 seconds, for gender-ID and age-group-ID, respectively, in a quiet room using the same computer and headphones.

4. Experimental Results I: Sub-Bands

The purpose of this section is to identify the most important frequency bands for speaker, gender and age-group identification from children's speech. During feature extraction (Section 3.3) the short-term magnitude spectrum

is passed through a bank of 24 Mel-spaced triangular filters, spanning frequencies from 0 Hz to 8000 Hz. Sub-band classification experiments were conducted using speech data comprising the outputs of groups of 4 adjacent filters. We considered 21 overlapping sub-bands, where the N^{th} 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 MFCCs plus 4 Δ and 4 Δ^2 parameters.

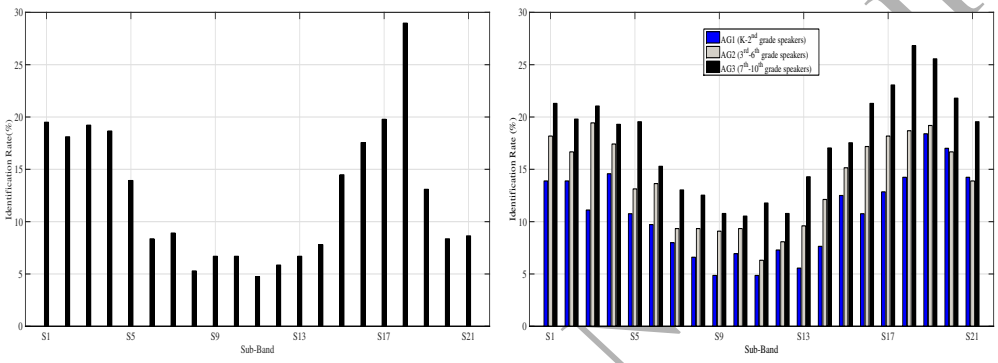
Feature warping, which improves performance in adult speaker recognition (Pelecanos and Sridharan (2001)), was applied to transform the short-term distribution of features to be normally distributed, using a 3 second sliding window. Systems were built for each of the 21 sub-bands using 64-component GMM-UBM systems (the choice of 64 components was determined empirically on a evaluation set disjoint from the test set). It will be seen (Section 5) that GMM-UBM systems do not give best performance in full-band experiments. However, the purpose of the present study is to measure the relative contribution, rather than optimize performance, of sub-bands for speaker-, age-group- and gender-identification.

4.1. Speaker recognition

Results of sub-band speaker identification experiments, using a test set of 10 second recordings from 359 speakers balanced for age and gender, are presented in Figure 1(a). The sub-band speaker identification accuracies vary between 5% and 34%. A second test set, comprising 288 recordings of duration 10 seconds each, from each of AG1, AG2 and AG3 (Section 3.1), was used to investigate the effect of age on sub-band recognition accuracy. Figure 1(b) shows sub-band speaker identification rates for children in AG1, AG2, and AG3. In most cases, the identification rate is highest for the older children, and decreases for younger children.

By analogy with adult speech (Besacier et al. (2000), Orman and Arslan (2001), Safavi et al. (2012)), it is useful to partition the spectrum into frequency regions. We consider four regions, B1 to B4, comprising sub-bands 1–5 (0 to 1.13 kHz), 6–14 (0.63 kHz to 3.8 kHz), 15–18 (2.1 kHz to 5.53 kHz), and 19–21 (3.4 kHz to 8 kHz), respectively. The most useful bands for speaker recognition are B1, corresponding to individual differences in primary vocal tract resonances and nasal speech sounds, and B3, corresponding to high-frequency speech sounds. The importance of fricatives (hence region B3) for speaker recognition was noted in (Parris and Carey (1994)), where it is hypothesized that this may be due to individual differences in the anatomy of the dental arch. Frequency regions similar to B1 to B4 were identified in

(Besacier et al. (2000), Orman and Arslan (2001)) for adult speaker recognition. However, the frequencies at which these bands occur in children's speech are increased by approximately 38% (B1), 21% (B2) and 11% (B3) relative to adults. One would expect the bands to occur at higher frequencies for young children, where formants and other structures will occur at higher frequencies due to their shorter vocal tracts.



(a) (b)

Figure 1: Speaker identification accuracy for each frequency sub-band for the GMM-UBM system: independent of age (a) and for age-groups AG1, AG2 and AG3 (b).

4.2. Gender identification

Figure 2 shows results, averaged across age-groups, of sub-band gender-ID experiments using age-dependent sub-band GMM-UBM systems. The test set comprises 687 recordings of child speech, each of duration 30 seconds, balanced by age and gender, and all recordings from the remaining speakers were used for training. For most sub-bands, accuracy is better than chance. The most useful sub-bands are from 8 to 12 (0.9 kHz to 2.6 kHz), which corresponds, approximately, to the location of the second formant. This region also provides the best gender-ID performance for adults (Wu and Childers (1991)).

Figure 3 shows gender-ID accuracy per sub-band by age-group. For sub-bands 1–10 and 20–21 there is a strong relationship between accuracy and age, with AG1 achieving the poorest and AG3 the best results. Indeed, for AG1, the performance is close to chance for sub-bands up to 7 (1.4 kHz). By contrast, the performance for sub-bands 10, 11 and 12 is comparable to

the full bandwidth performance for AG3. For sub-bands 12–19 the picture is less clear and the variations in performance for the different age-groups is less marked.

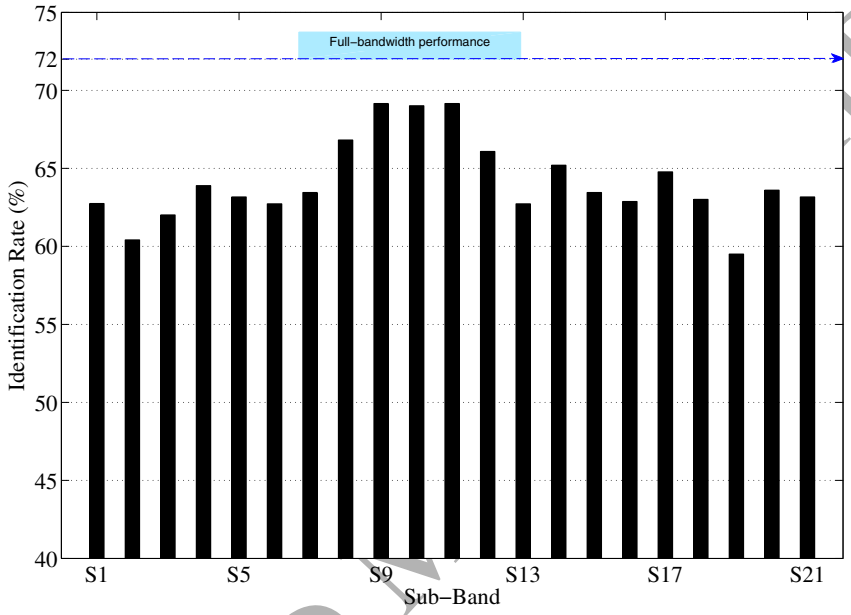


Figure 2: Gender-ID accuracy for different frequency sub-bands averaged over all age-groups obtained by the GMM-UBM system.

We speculated that poor performance for sub-bands up to 11 for young children (AG1) might be due to the wide spacing of their pitch harmonics, compared with the narrow low frequency Mel-scale filters (Ghai (2011)). In automatic speech recognition, this is mitigated by broadening these filters so that the equivalent rectangular bandwidth is 300 Hz (Ghai (2011)). The results of applying this to the filters in sub-bands 1–11 are shown as cyan bars in Figure 3, which indicates significantly improved performance.

4.3. Age-group identification

All sub-band age-group-ID experiments use gender-independent GMM-UBM systems and the test set of 30 second recordings from 766 speakers (290, 285 and 191 from AG1, AG2 and AG3, respectively), balanced for gender.

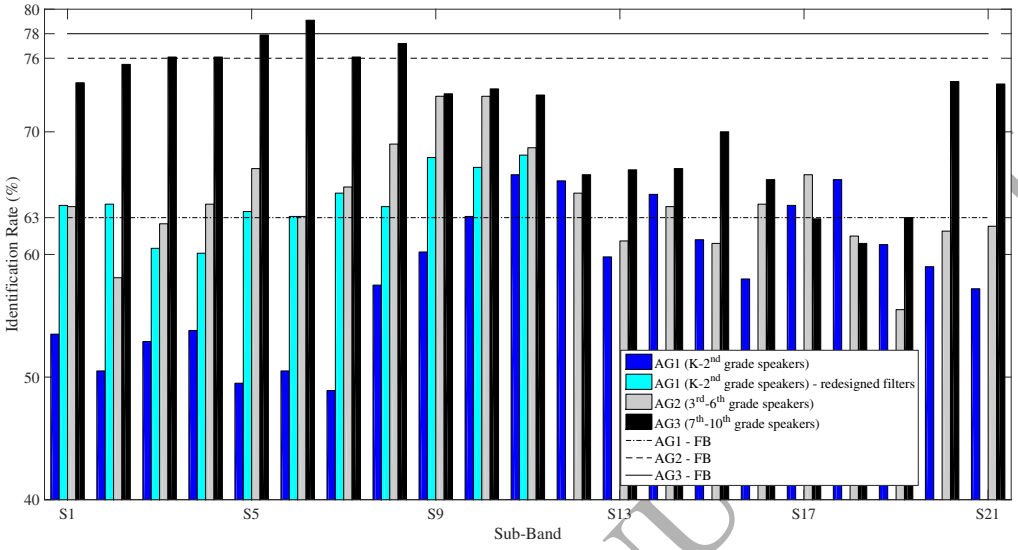
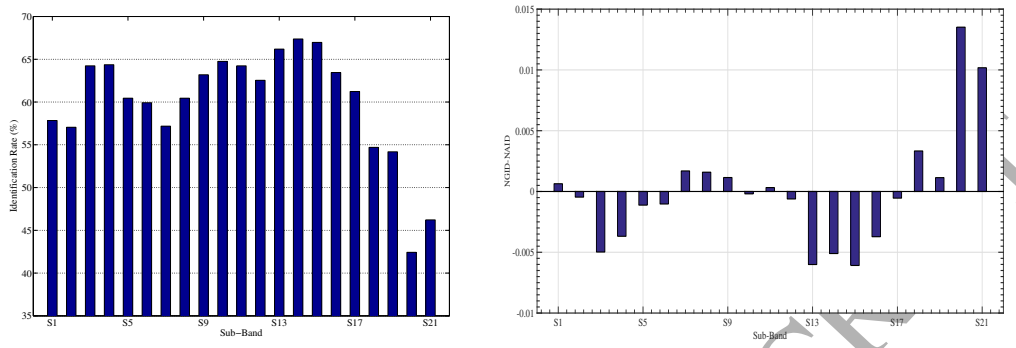


Figure 3: Gender-ID accuracy for different frequency sub-bands for each of the three children age-groups obtained by the GMM-UBM system. The horizontal lines indicate the full-bandwidth (FB) gender-ID accuracy for each age-group.

The recordings from the remaining 344 speakers were used for training. The use of gender-independent models is justified in Section 5.3.

Figure 4(a) presents average age-group-ID accuracy for each sub-band. Even for narrow frequency regions, performance is above chance in most cases. The best performance is achieved using sub-bands 13–16, while sub-bands 18–21 are the least useful for age-group-ID.

Figure 4(b) contrasts the usefulness of sub-bands for age-group-ID and gender-ID. The figure was obtained by normalising the data in Figures 2 and 4(a) so that in both cases the sum of the values over all of the sub-bands is 1. The normalised age-group-ID results were then subtracted from the normalised gender-ID results to obtain Figure 4(b) (a similar procedure is described in (Safavi et al. (2012))). Positive and negative values in Figure 4(b) indicate sub-bands which are more useful for gender-ID and age-group-ID, respectively. The most useful sub-bands for age-group-ID, relative to gender-ID, are 3 and 4 (281 Hz to 625 Hz), and from 13 to 16 (1.62 kHz to 3 kHz). Thus, while gender-ID appears to make use of similar information to speaker recognition, age-group-ID is more similar to speech recognition or accent ID (Safavi et al. (2012)).



(a) (b)

Figure 4: Age-group-ID accuracy (a) and the differences between normalized gender-ID and age-group-ID (NGID–NAID) accuracy (b) for each frequency sub-band.

5. Experimental Results II: Full bandwidth

5.1. Speaker recognition

Table 5 shows full-bandwidth speaker verification and identification results for age-groups AG1 to AG3 (288 children per group, 10 second test utterances), using 1024 component GMM-UBM, 64-component GMM-SVM systems and i-vector-PLDA. The number of GMM components was determined empirically using a separate evaluation set which is disjoint from the test set.

In most cases verification EER decreases and identification accuracy increases with age. For example, for the GMM-UBM system the EER drops by 70% from 2.1% for AG1 to 0.67% for AG3. The corresponding increase in identification rate is 38%. The i-vector-PLDA system achieves the best overall performance.

Comparison with results for adult speech is complicated by the focus of current research on conversational telephone speech with challenging channel and noise conditions. However, GMM-UBM speaker identification error rates less than 0.5% and verification equal error rates of 0.24% have been reported for TIMIT (Reynolds (1995)). Although these are plausible limiting values for the scores in Table 5, they suggest that even for children in AG3 speaker recognition is more challenging than for adults' speech.

Table 6 shows the results of simulated identification and verification of

Table 5: Speaker verification equal error rate (EER) and identification accuracy for three different age-groups of children (AG1, AG2 and AG3).

	GMM-UBM (1024)	GMM-SVM (64)	i-vector (PLDA)
Verification – EER (%)			
AG1 (K-2^{nd})	2.10	1.80	**1.59**
AG2 (3^{rd}-6^{th})	1.33	1.21	**0.72**
AG3 (7^{th}-10^{th})	0.67	0.64	**0.53**
Identification – Acc (%)			
AG1 (K-2^{nd})	62.15	75.00	**78.16**
AG2 (3^{rd}-6^{th})	80.56	89.24	**91.13**
AG3 (7^{th}-10^{th})	85.71	93.26	**95.38**

a child in a classroom, using i-vector-PLDA and 10 second test utterances. The experiment uses 12 "classrooms" of 30 children, 4 for each age-group, balanced across gender. The results follow the expected pattern, with identification accuracy increasing and verification EER decreasing with age.

Table 6: Speaker verification EER and identification accuracy for three different age-groups (AG1, AG2 and AG3) obtained using the i-vector-PLDA system for the case of a child in a classroom scenario.

Classroom scenario	Verification EER (%)	Identification Acc (%)
AG1 (K-2^{nd})	1.89	86.39
AG2 (3^{rd}-6^{th})	1.39	93.61
AG3 (7^{th}-10^{th})	0.97	97.34

5.2. Gender identification

5.2.1. Age-independent and age-dependent modelling

Age-independent GMM-UBM, GMM-SVM and i-vector gender-ID systems were evaluated on 30 second test utterances from 687 speakers balanced by age and gender. In the age-dependent systems a separate system was created for each age-group. Table 7 shows that the age-dependent systems

perform best, with accuracy increasing from 67.39%, for the age-independent GMM-UBM system, to 79.18% for the age-dependent GMM-SVM system. In both cases the GMM-SVM systems achieve the best results. The numbers of GMM components in the age-independent and -dependent GMM-UBM, GMM-SVM and i-vector systems were 1024, 512, and 256, and 512, 256 and 128, respectively.

Table 7: Gender-ID accuracy using age-independent and age-dependent GMM-UBM, GMM-SVM, and i-vector systems.

System	Gender-ID rate (%) (age-independent)	Gender-ID rate (%) (age-dependent)
GMM-UBM	67.39	71.76
GMM-SVM	**77.44**	**79.18**
i-vector	74.26	72.54

The application of intersession variability compensation (ISVC) (Vair et al. (2006)) resulted in only small improvements. Specifically, the accuracies of the age-independent and -dependent systems improved from 67.39% to 69.29%, and 71.76% to 72.81%, respectively.

5.2.2. Effect of age on gender-identification performance

Table 8 shows the performance of the age-dependent GMM-SVM system according to child age and gender. For AG1 and AG2, performance for boys is poorer than girls. The accuracy for AG2 is considerably higher than for AG1 (the performance increase is 7.34%), but for AG3 the accuracy is unexpectedly low. We speculated that this is because the boys in AG3 fall into two subsets, according to whether or not their voices have broken as a consequence of reaching puberty, and boys whose voices have not broken may achieve a better match with the "girls" model. It has been estimated (Rogol et al. (2000)) that puberty starts at a skeletal (biological) age of 11 years in girls and 13 years in boys.

Table 9(a) is the confusion matrix for AG3 for the GMM-UBM system from Table 8, which uses two gender-dependent models. It indicates that 28.5% of boys are misclassified as girls, compared with only 11.5% of girls misclassified as boys. The focus on GMM-UBMs is due to the limited data available in the next experiment.

Table 8: Gender-ID accuracies (in %) for the age-dependent GMM-SVM system for boys ("B"), girls ("G") and overall ("Av."), and overall ("Av.") accuracy for human listeners, for each age group.

Age group	GMM-SVM			Human
	B	G	Av.	Av.
AG1	73.33	80.86	76.80	60.48
AG2	79.50	88.70	84.14	70.49
AG3	77.69	72.13	75.91	70.90

A subset of the boys data in AG3 was split into two separate classes: those whose voices have broken (BB), or remained unbroken (BU), each containing recordings from 18 speakers. A human listener allocated the boys' data to these two classes. This data, together with recordings from 18 of the girls in AG3 (G), was used to create 3 GMM-UBM sub-group models, corresponding to (BB), (BU) and (G). The number of GMM components was reduced to 128 because of the limited data.

Table 9(b) shows the confusion matrix for AG3 for the three gender sub-group GMM-UBM systems. The overall gender identification rate for AG3 is 87.43%, compared with 78.53% using two gender models, and the number of boys misclassified as girls is reduced to zero. Girls are misclassified as boys with unbroken voices 3.5 times more often than as boys with broken voices. Gender-ID accuracy for girls decreases, because more girls are confused with boys with unbroken voices, suggesting that an appropriate partition of the girls' data may also improve performance.

Table 9: Confusion matrix for gender identification (in %) for age-group AG3 when using (a) a single model for boys, and (b) two separate models for boys, corresponding to broken B_B and unbroken B_U voices.

(a)

	B	G
B	71.5	28.5
G	11.5	88.5

(b)

	B_B	B_U	G
B_B	96.2	3.8	0
B_U	5.8	94.2	0
G	6.5	23.0	70.5

5.2.3. Human performance

The final column of Table 8 shows gender-ID accuracies achieved by human listeners. The average performance over all age-groups is 66.96%. Evidently humans, as well as computers, have difficulty identifying the gender of a child from his or her speech.

5.3. Age-group identification

5.3.1. Gender-independent and -dependent modelling

Table 10 presents age-group-ID results for GMM-UBM, GMM-SVM, and i-vector systems. The test set is the same set of 766 recordings, each of length 30 seconds, that was used in the sub-band age-group-ID experiments. The first two rows of the table compare the age-group-ID accuracies obtained using gender-dependent and -independent GMM-UBM systems, with the latter giving better performance. This may be due to the smaller amount of training data available for gender-dependent age modelling, or because for very young children gender does not provide significant benefit for age-group-ID. Subsequent experiments use gender-independent modelling.

The next two rows of Table 10 present age-group-ID accuracy obtained with gender-independent GMM-SVM and i-vector systems. The number of GMM components in the GMM-UBM, GMM-SVM, and i-vector systems is 1024, 512 and 256, respectively, and the dimension of the i-vector total variability subspace is 400. These parameters were set empirically using a separate evaluation set. The best performance, 82.62%, was obtained using the i-vector system.

Table 10: age-group-ID recognition rate (in %) obtained by the gender-independent GMM-UBM, GMM-SVM and i-vector systems and gender-dependent GMM-UBM system.

System	age-group-ID rate (%)
GMM-UBM (gender dep.)	71.76
GMM-UBM (gender indep.)	82.01
GMM-SVM (gender indep.)	79.77
i-vector (gender indep.)	**82.62**

Based on the results in section 4, we also conducted age-group-ID experiments using speech that was band-limited to 5.5 kHz. For the gender-independent i-vector systems, this resulted in an age-group-ID accuracy of 85.77%, compared with 82.62% for full-bandwidth speech.

Table 11 is the confusion matrix for the i-vector system with 5.5 kHz bandwidth speech. Each row corresponds to a grade and shows the percentages of children in that grade who were classified as being in AG1, AG2, and AG3. The thin horizontal lines indicate the AG1, AG2, and AG3 boundaries. The table shows similar characteristics for boys and girls up to the 7^{th} grade, with the majority of errors near age-group boundaries. At the boundary between AG1 and AG2, 10% of 3^{rd} grade boys (AG1) are incorrectly classified as AG2, and 33% of 4^{th} grade boys (AG2) are incorrectly classified as AG1. For girls, the corresponding figures are 8% and 39%. For 7^{th} grade, (AG2) 45% of boys and 24% of girls are classified as being in AG3, while for 8^{th} grade, (AG3) 29% of girls are classified as AG2, but none of the boys are misclassified. The inconsistency between the results for boys and girls at the AG2-AG3 boundary may be because AG3 contains speech from a number of boys whose voices have broken. It may be that gender-dependent modelling is needed for AG3, even though it is not advantageous overall, or that, as in the case of gender-ID it is necessary to build separate models for AG3 boys whose voices have or have not broken.

Table 11: Confusion matrix for age-group identification (in %) for three age-groups, obtained with the i-vector system using 5.5 kHz band-limited speech.

Age-		Male			Female		
Group	Grade	AG1	AG2	AG3	AG1	AG2	AG3
AG1	K	100	0	0	100	0	0
	1^{st}	100	0	0	100	0	0
	2^{nd}	97.43	2.56	0	97.87	2.12	0
	3^{rd}	85.71	10.20	4.08	92.10	7.89	0
AG2	4^{th}	33.33	60.60	6.06	38.70	61.29	0
	5^{th}	8.57	82.85	8.57	11.42	82.85	5.71
	6^{th}	6.97	81.39	11.62	10.00	80.00	10.00
	7^{th}	0	54.83	45.16	2.70	72.97	24.32
AG3	8^{th}	0	0	100	0	29.03	70.96
	9^{th}	2.17	6.52	91.30	5.00	20.00	75.00
	10^{th}	0	0	100	0	30.00	70.00

5.3.2. Human performance

The age-group-ID rates for each age-group for human listeners are presented in Table 12. The results indicate that most confusions arise for test utterances in AG2. Only 50.8% of these are correctly identified. The confusion between AG1 and AG3 is small, with only 1.8% and 3.8% of the test utterances from AG1 and AG3 being incorrectly identified as AG3 and AG1, respectively. The average performance over all age-groups is 67.54%.

Table 12: Confusion matrix for age-group-ID (in %) for three age-groups, obtained by human listeners.

	Model		
Test data	AG1	AG2	AG3
AG1	**81.2**	16.9	1.8
AG2	25.5	**50.8**	23.6
AG3	3.8	24.4	**71.7**

6. Conclusions

The objectives of this paper are to investigate, and provide performance benchmarks for, automatic speaker, gender and age-group identification for children's speech, and to identify the regions of the spectrum that contain the most important information for these three classification problems.

The OGI "Kids" corpus was used in all of the experiments. In order to study the effects of age, it was partitioned into three age-groups, AG1 (5–8 years), AG2 (8–12 years) and AG3 (12–16 years).

Three different types of classifier were considered, namely GMM-UBM, GMM-SVM and i-vectors with PLDA-based classification. Gender and age-group identification rates for human listeners are also presented.

For speaker identification and verification for children, the most useful frequencies are up to 1.1 kHz (B1) and between 2.1 kHz and 5.5 kHz (B3). These are similar to the bands identified in (Safavi et al. (2012)) and elsewhere for adults, increased by approximately 38% (B1) and 11% (B3) for children. For gender identification the most useful frequencies are between 0.9 kHz and 2.6 kHz. This corresponds approximately to the location of the second formant for vowels, which again is consistent with results for adults'

speech (Wu and Childers (1991)). The most useful frequencies for age-group identification from child speech are between 1.6 kHz and 3 kHz.

The results for gender identification show very poor performance for sub-bands 1–11 for the youngest children (AG1). Since the relationship between the mel and Hertz scales is approximately linear over this region, the equivalent rectangular bandwiths of the triangular mel scale filters is small compared with the wide spacing of spectral harmonics for these children, due to high fundamental frequency. This increases variability in the high-order MFCCs (Ghai (2011)). The problem was alleviated by setting the minimum bandwidth of the mel scale filters to be 300 Hz. The same problem does not seem to occur for speaker or age-group identification, presumably because in these cases fundamental frequency is relevant for distinguishing between classes.

For full-bandwidth speech the best speaker recognition performance is obtained with i-vectors combined with a PLDA classifier. This achieves verification equal error rates (identification accuracies) of 1.59%, 0.72% and 0.53% (78.16%, 91.13% and 95.38%) for AG1, AG2 and AG3, respectively.

For gender identification the situation is more complex. The best full-bandwidth gender identification accuracy is 79.18% obtained with an age-independent GMM-SVM system. However, the relationship between accuracy and age is not straightforward. For a GMM-UBM system, accuracy increases from 76.8% for AG1 to 84.1% for AG2, but then drops to 75.9% for AG3. This is because AG3 includes boys whose voices have and have not broken. When the training set for boys was partitioned into two sets, corresponding to broken and unbroken voices, and two separate boys' gender models were created, gender identification accuracy for AG3 rose to 87.4%. The average gender identification performance for human listeners over the three age-groups is 67%, which is poorer than the best automatic system.

For age-group identification with full-bandwidth speech, the best performance, 83%, is obtained with a gender-independent i-vector system. Reducing the speech bandwidth to 5.5 kHz (motivated by the sub-band age-group identification results) increases the accuracy to 85.8%. The corresponding figure for human listeners is 67.5%.

In summary, our results emphasize the importance of taking properties of children's speech into account in the design of automatic speaker-, gender- and age-group-identification systems. These include the occurrence of structure at higher frequencies, the effects of higher fundamental frequency on feature extraction, and the effects of the onset puberty.

Finally, speaker and language recognition, in particular, are the subjects of considerable ongoing research. It will be interesting to discover how well new techniques that emerge are able to accommodate the particular challenges of children's speech.

References

Auckenthaler, R., Carey, M. J., Lloyd-Thomas, H., 2000. Score normalisation for text-independent speaker verification systems. Digital Signal Processing 10 (1-3), 42–54.

Bahari, M., Van Hamme, H., 2011. Speaker age estimation and gender detection based on supervised non-negative matrix factorization. In: Proc. IEEE Workshop on Biometric Measurements and Systems for Security and Medical Applications. pp. 1–6.

Bahari, M. H., McLaren, M., van Hamme, H., van Leeuwen, D. A., 2012. Age estimation from telephone speech using i-vectors. In: Proc. Interspeech, *Portland, Oregon, USA*. ISCA.

Batliner, A., Hacker, C., Steidl, S., Noth, E., D'Arcy, S., Russell, M., Wong, M., 2004. 'you stupid tin box' - children interacting with the aibo robot: A cross-linguistic emotional speech corpus. In: Proc. 4th Int. Conf. on Language Resources and Evaluation LREC 2004 *Lisbon, Portugal*. pp. 171–174.

Besacier, L., Bonastre, J. F., Fredouille, C., 2000. Localization and selection of speaker-specific information with statistical modeling. Speech Communication 31 (2-3), 89–106.

Biadsy, F., Hirschberg, J., Ellis, D. P. W., 2011. Dialect and accent recognition using phonetic-segmentation supervectors. In: Proc. Interspeech, *Florence, Italy*. pp. 745–748.

Bocklet, T., Maier, A., Bauer, J. G., Burkhardt, F., Noth, E., 2008. Age and gender recognition for telephone applications based on GMM supervectors and support vector machines. Proc. IEEE-ICASSP, *Las Vegas, Nevada, USA*, 1605–1608.

Campbell, J., Shen, W., Campbell, W., Schwartz, R., Bonastre, J.-F., Matrouf, D., 2009. Forensic speaker recognition. IEEE Signal Processing Magazine 26 (2), 95–103.

Campbell, W., Karam, Z. N., 2009. A framework for discriminative SVM/GMM systems for language recognition. In: Proc. Interspeech, *Brighton, UK*.

Campbell, W., Sturim, D., Reynolds, D., Solomonoff, A., 2006. SVM based speaker verification using a GMM supervector kernel and NAP variability compensation. In: Proc. IEEE-ICASSP, *Toulouse, France*. Vol. 1. pp. I–I.

Chen, C.-C., Lu, P.-T., Hsia, M.-L., Ke, J.-Y., Chen, O.-C., August 2011. Gender-to-age hierarchical recognition for speech. In: Proc. IEEE Int. Midwest Symposium on Circuits and Systems. pp. 1–4.

Davis, S. B., Mermelstein, P., 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. on Acoustics, Speech and Signal Processing 28 (4), 357–366.

Dehak, N., Kenny, P., Dehak, R., Dumouchel, P., Ouellet, P., 2011a. Front-end factor analysis for speaker verification. IEEE Trans. on Audio, Speech, and Language Processing 19 (4), 788–798.

Dehak, N., Kenny, P., Dehak, R., Dumouchel, P., Ouellet, P., 2011b. Front-end factor analysis for speaker verification. IEEE Trans. on Audio, Speech, and Language Processing 19 (4), 788–798.

DeMarco, A., Cox, S. J., 2012. Iterative classification of regional british accents in i-vector space. In: Proc. Symposium on Machine Learning in Speech and Language Processing.

Dobry, G., Hecht, R. M., Avigal, M., Zigel, Y., 2009. Dimension reduction approaches for SVM based speaker age estimation. In: Proc. Interspeech, *Brighton, UK*. ISCA, pp. 2031–2034.

Doddington, G., 2001. Speaker recognition based on idiolectal differences between speakers. In: Proc. Eurospeech '01, *Aalborg, Denmark*.

Elenius, D., Blomberg, M., 2005. Adaptation and normalization experiments in speech recognition for 4 to 8 year old children. In: Proc. Interspeech, *Lisbon, Portugal*. pp. 2749–2752.

Fringi, E., Lehman, J., Russell, M., 2015. Evidence of phonological processes in automatic recognition of children's speech. In: Proc. Interspeech, *Dresden, Germany*. pp. 1621–1624.

Garcia-Romero, D., Espy-Wilson, C., 2011. Analysis of i-vector length normalization in speaker recognition systems. In: Proc. Interspeech, *Florence, Italy*. pp. 249–252.

Gauvain, J.-L., Lee, C., 1994. Maximum a-posteriori estimation for multivariate Gaussian mixture observations of Markov chains. IEEE Trans. Speech and Audio Processing 2, 291–298.

Gerosa, M., Giuliani, D., Brugnara, F., 2005. Speaker adaptive acoustic modeling with mixture of adult and childrens speech. In: Proc. Interspeech, *Lisbon, Portugal*. pp. 2193–2196.

Gerosa, M., Giuliani, D., Brugnara, F., 2007. Acoustic variability and automatic recognition of children's speech. Speech Communication 49, 847–860.

Gerosa, M., Lee, S., Giuliani, D., Narayanan, S., 2006. Analysing children's speech: An acoustic study of consonants and consonant-vowel transition. In: Proc. IEEE-ICASSP, *Toulouse, France*. Vol. 1. pp. 393–396.

Ghai, S., 2011. Addressing Pitch Missmatch for Children's Automatic Speech Recognition. Ph.D. thesis, Indian Institute of Technology Guwahati.

Hanani, A., Russell, M., Carey, M., 2013. Human and computer recognition of regional accents and ethnic groups from British English speech. Computer Speech Language 27 (1), 59–74.

Kenny, P., Boulianne, G., Ouellet, P., Dumouchel, P., 2007. Joint factor analysis versus eigenchannels in speaker recognition. IEEE Trans. on Audio, Speech, and Language Processing 15 (4), 1435–1447.

Kinnunen, T., Li, H., 2010. An overview of text-independent speaker recognition: From features to supervectors. Speech Communication 52, 12–40.

Lee, S., Potamianos, A., Narayanan, S., 1997. Analysis of children's speech: duration, pitch and formants. In: Proc. Eurospeech, *Rhodes, Greece*.

Lee, S., Potamianos, A., Narayanan, S., 1999. Acoustics of children's speech: Developmental changes of temporal and spectral parameters. Journal of the Acoustical Society of America 105 (3), 1455–1468.

Lee, S., Rose, R., 1998. A frequency warping approach to speaker normalization. In: Proc. IEEE-ICASSP, *Seattle, WA*. Vol. 6. pp. 49–60.

Lei, Y., Scheffer, N., Ferrer, L., McLaren, M., 2014. A novel scheme for speaker recognition using a phonetically-aware deep neural network. In: Proc. IEEE-ICASSP, *Florence, Italy*. pp. 1714–1718.

Liao, H., Pundak, G., Siohan, O., Carroll, M. K., Coccaro, N. Jiang, Q.-M., Sainath, T. N., Senior, A., Beaufays, F., Bacchiani, M., 2015. Large vocabulary automatic speech recognition for children. In: Proc. Interspeech, *Dresden, Germany*.

Lopez-Moreno, I., Gonzalez-Dominguez, J., Pichot, O., Martinez, D., Gonzalez-Rodriguez, J., Moreno, P., 2014. Automatic language identification using deep neural networks. In: Proc. IEEE-ICASSP, *Florence, Italy*. pp. 5374–5378.

Mahmoodi, D., Soleimani, A., Marvi, H., Razzazi, F., Taghizadeh, M., Mahmoodi, M., 2011. Age estimation based on speech features and support vector machine. In: Proc. Computer Science and Electronic Engineering Conf. pp. 60–64.

Massaro, D. W., 1998. Perceiving Talking Faces: From Speech Perception to a Behavioral Principle. MIT Press: Cambridge, MA.

Metze, F., Ajmera, J., Englert, R., Bub, U., Burkhardt, F., Stegmann, J., Muller, C., Huber, R., Andrassy, B., Bauer, J., Littel, B., 2007. Comparison of four approaches to age and gender recognition for telephone applications. In: Proc. IEEE-ICASSP, *Honolulu, Hawai*. Vol. 4. pp. IV–1089–IV–1092.

Naik, J., Netsch, L., Doddington, G., 1989. Speaker verification over long distance telephone lines. In: Proc. IEEE-ICASSP, *Glasgow, Scotland*. pp. 524–527 vol.1.

NIST, 2010. 2003 NIST Speaker Recognition Evaluation LDC2010S0. National Institute of Standards and Technology, (Speech Group), Gaithersburg, MD.

Orman, O., Arslan, L., 2001. Frequency analysis of speaker identification. In: Proc. 2001: A Speaker Odyssey, *Crete, Greece*.

Parris, E. S., Carey, M. J., 1994. Discriminative phonemes for speaker identification. In: Proc. Int. Conf. on Spoken Lang. Proc., *Yokohama, Japan*.

Pelecanos, J., Sridharan, S., 2001. Feature warping for robust speaker verification. Proc. 2001: A Speaker Odyssey, *Crete, Greece*.

Prince, S., Elder, J., 2007. Probabilistic linear discriminant analysis for inferences about identity. In: IEEE Int. Conf. on Computer Vision. pp. 1–8.

Reynolds, D., 1995. Speaker identification and verification using gaussian mixture speaker models. Speech Communication 17, 91–108.

Reynolds, D., Rose, R., 1995. Robust text-independent speaker identification using Gaussian mixture speaker models. IEEE Trans. on Speech and Audio Proc. 3, 72–83.

Reynolds, D. A., Quatieri, T. F., Dunn, R. B., 2000. Speaker verification using adapted Gaussian mixture models. Digital Signal Processing 10 (1-3), 19–41.

Rogol, A. D., Clark, P. A., Roemmich, J. N., 2000. Growth and pubertal development in children and adolescents: effects of diet and physical activity. The American journal of clinical nutrition 72 (2), 521s–528s.

Russell, M., Series, R., Wallace, J., Brown, C., Skilling, A., 2000. The star system: an interactive pronunciation tutor for young children. Computer Speech and Language 14 (2), 161–175.

Russell, M. J., D'Arcy, S., Qun, L., 2007. The effects of bandwidth reduction on human and computer recognition of children's speech. IEEE Signal Processing Letters 14 (12), 1044–1046.

Safavi, S., Hanani, A., Russell, M. J., Jančovič, P., Carey, M., 2012. Contrasting the effects of different frequency bands on speaker and accent identification. IEEE Signal Processing Letters 19 (12), 829–832.

Sarkar, A. K., Do, C.-T., Le, V.-B., Barras, C., 2014. Combination of cepstral and phonetically discriminative features for speaker verification. IEEE Signal Process. Lett. 21 (9), 1040–1044.

Schuller, B., Batliner, A., 2013. Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing. Wiley.

Schuller, B., Batliner, A., Steidl, S., Seppi, D., 2011. Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. Speech Communication 53 (9-10), 1062–1087.

Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C., Narayanan, S., 2013. Paralinguistics in speech and language - state-of-the-art and the challenge. Computer Speech and Language, Special issue on Paralinguistics in Naturalistic Speech and Language 27, 4–39.

Serizel, R., Giuliani, D., 2016. Deep-neural network approaches for speech recognition with heterogeneous groups of speakers including children. to appear in Natural Language Engineering.

Shobaki, K., Hosom, J. P., Cole, R. A., 2000. The OGI kids' speech corpus and recognizers. Proc. Int. Conf. on Spoken Lang. Proc., *Beijing, China*.

Singer, E., Torres-Carrasquillo, P. A., Reynolds, D. A. McCree, A., Richardson, F. amd Dehak, N., Sturim, D. E., 2012. The MITLL NIST LRE 2011 language recognition system. In: Proc. Odyssey'12, The Speaker and Language Recognition Workshop, *Singapore*. ISCA, pp. 209–2015.

Solomonoff A., Quillen, C., Campbell, W., 2004. Channel compensation for svm speaker recognition. In: Proc. Odyssey'04, The Speaker and Language Recognition Workshop, *Toledo, Spain*. p. 219226.

Song, Y., Jiang, B., Bao, Y., Wei, S., Dai, L.-R., 2013. I-vector representation based on bottleneck features for language identification. Electron. Lett., 1569–1580.

Vair, C., Colibro, D., Castaldo, F., Dalmasso, E., Laface, P., 2006. Channel factors compensation in model and feature domain for speaker recognition. In: Proc. Odyssey'06, The Speaker and Language Recognition Workshop, *San Juan, Puerto Rico*. pp. 1–6.

Van Heerden, C., Barnard, E., Davel, M., van der Walt, C., van Dyk, E., Feld, M., Muller, C., 2010. Combining regression and classification methods for improving automatic speaker age recognition. In: Proc. IEEE-ICASSP, *Dallas, TX, USA*. pp. 5174–5177.

Variani, E., Lei, X., McDermott, E., Lopez-Moreno, I., GonzalezDominguez, J., 2014. Deep neural networks for small footprint text-dependent speaker verification. In: Proc. IEEE-ICASSP, *Florence, Italy*. pp. 4080–4084.

Wan, V., Campbell, W., 2000. Support vector machines for speaker verifcation and identifcation. In: Proc. of the 2000 IEEE Signal Processing Society Workshop on Neural Networks for Signal Processing. Vol. 2. pp. 775–784.

Wilpon, J., Jacobsen, C., 1996. A study of speech recognition for children and the elderly. In: Proc. IEEE-ICASSP, *Atlanta, GA*.

Wong, E., Sridharan, S., 2001. Fusion of output scores on language identification system. In: Workshop on Multilingual Speech and Language, Processing, Aalborg Denmark.

Wu, K., Childers, D. G., 1991. Gender recognition from speech. Part II: Fine analysis. Journal of the Acoust. Soc. Am. 90 (4), 1841–1856.

Yamada, T., Wang, L., Kai, A., 2013. Improvement of distant-talking speaker identification using bottleneck features of dnn. In: Proc. Interspeech, *Lyon, France*. pp. 3661–3664.

Zissman, M., January 1996. Comparison of four approaches to automatic language identification of telephone speech. IEEE Trans. on Speech and Audio Proc. 4 (1), 31–44.