Age and Gender Detection in the I-DASH Project

HUGO MEINEDO, L2F - Spoken Language Systems Lab, INESC-ID
ISABEL TRANCOSO, L2F - Spoken Language Systems Lab, INESC-ID and Instituto Superior Técnico

This article presents a description of the INESC-ID Age and Gender classification systems which were developed for aiding the detection of child abuse material within the scope of the European project I-DASH. The Age and Gender classification systems are composed respectively by the fusion of four and six individual subsystems trained with short- and long-term acoustic and prosodic features, different classification strategies, Gaussian Mixture Models-Universal Background Model (GMM-UBM), Multi-Layer Perceptrons (MLP) and Support Vector Machines (SVM), trained over five different speech corpus. The best results obtained by the calibration and linear logistic regression fusion back-end show an absolute improvement of 2% on the unweighted accuracy value for the Age and 1% for the Gender when compared to the best individual frontend systems in the development set. The final age/gender detection system evaluated using a six-hour child abuse (CA) test set achieved promising results given the extremely difficult conditions of this type of video material. In order to further improve the performance in the CA domain, the classification modules were adapted using unsupervised selection of training data. An automatic data selection algorithm using frame-level posterior probabilities was developed. Performance improvement after adapting the classification modules was around 10% relative when compared with the baseline classifiers.

Categories and Subject Descriptors: I.2.7 [**Artificial Intelligence**]: Natural Language Processing

General Terms: Measurement

Additional Key Words and Phrases: Age, gender, fusion of acoustic and prosodic features

ACM Reference Format:
Meinedo, H. and Trancoso, I. 2011. Age and gender detection in the I-DASH project. ACM Trans. Speech Lang. Process. 7, 4, Article 13 (August 2011), 16 pages.
DOI = 10.1145/1998384.1998387 http://doi.acm.org/10.1145/1998384.1998387

1. INTRODUCTION

Paralinguistic analysis is a rapidly emerging field of research due to the constantly growing interest in applications in the fields of Multimedia Retrieval and Human-Machine Interaction. Gender and age detection are two of its tasks. Gender detection is a very useful task for a wide range of applications. In the Spoken Language Systems lab of INESC-ID, the Gender Detection module is one of the basic components of our audio segmentation system [Meinedo 2008], where it is used prior to speaker clustering, in order to avoid mixing speakers from different genders in the same cluster. Gender information is also used for building gender-dependent acoustic modules for speech recognition. In our fully automatic Broadcast News subtitling system, deployed at the national TV channel since March 2008 [Meinedo 2008], gender information is also

This work was partly funded by the European project I-DASH. This work was supported by FCT (INESC-ID multiannual funding) through the PIDDAC Program funds.
Authors' addresses: H. Meinedo, I. Trancoso, L2F—Spoken Language Systems Lab, INESC-ID Lisboa, Rua Alves Redol, 9, 1000-029, Portugal; email: {hugo.meinedo, isabel.trancoso}@inesc-id.pt.

DOI 10.1145/1998384.1998387 http://doi.acm.org/10.1145/1998384.1998387

used to change the color of the subtitles, thus helping people with hearing difficulties to detect which speaker the subtitle refers to, a useful hint that partially compensates for the small latency of the subtitling system. Gender Detection is also a prominent part of our participation in the VIDIVIDEO European project, aiming at the semantic search of audio-visual documents [Bugalho et al. 2009]. In this application, the audio concept "male-voice" may be much easier to detect than the corresponding video-concept "male-speaker."

Most gender classification systems are trained for distinguishing between male and female adult voices alone. In fact, in some applications like Broadcast News (BN) transcription, children's voices are relatively rare, hence justifying their non-inclusion. The difficulties in collecting large corpora of children's voices may also be one of the reasons that most detectors do not attempt a three-class distinction. In some applications, such as the automatic detection of child abuse videos on the Web, however, the detection of children's voices may be specially relevant. This is the main goal of this article and was the target of our participation in the European I-DASH project. I-Dash was a project within the Safer Internet Plus program. It focused on the development of automatic tools to support police professionals in investigations that involve large quantities of child abuse video material. Recent years have shown rapid changes in the way digital media are produced, handled, and distributed. These developments have had a similar impact on the production and distribution of child abuse material. Analogue material has been mostly replaced by digital material. Illicit material is distributed efficiently and commercially via the Internet. While photos are still found in huge amounts, police forces have seen a very steep increase of videos containing child abuse material in the last five years. In the video domain, police forces worldwide are struggling because of lack of tools to handle the huge amounts of video data efficiently. A police investigator working on a child abuse case needs tools to filter irrelevant material at an early stage, to recognize known child abuse and to know which new material needs analysis. For the analysis of new material, the investigator needs tools to search for the presence of children, nudity, or sexual activity. Finally, to link material the investigator must be able to find similarities between videos based on objects such as furniture or tattoos. The I-DASH project addressed the development of tools that makes the above task more efficient, while leaving the investigator in full control.

Our goal in this paper and our role in the I-DASH project was to contribute with whatever cues may be derived from the audio signal to the automatic detection of child abuse (CA). One of the potentially most important audio clues is the presence of children voices. In order to accomplish this goal, our audio analysis tool is composed by several modules performing two tasks. The first task is to identify (filter out) potentially irrelevant material, in this case nonspeech segments. This is the role of the speech/nonspeech detector. The second task is the detection of audio semantic concepts which maybe relevant for the I-DASH project. This is done by applying two classifiers: the age/gender classifier is applied to the segments previously classified as speech and provides a ternary male/female/child classification. The background conditions classifier is mainly used to detect the presence of music.

In the three following sections we discuss the related work (Section 2), introduce the corpora (Section 3), and the speech/nonspeech detector (Section 4). Next we discuss the systems developed for the detection of age (Section 5) and gender (Section 6). A detailed perceptual study is presented in Section 7 to better access human classification capabilities in these tasks. Section 8 presents the evaluation results of the complete audio analysis tool developed for the detection of CA material. Afterwards we present the adaptation of the classification modules using unsupervised selection of training data (Section 9) in order to further improve the performance in the CA domain. Finally we draw some conclusions (Section 10).

2. RELATED WORK

Children's voices differ from from adult voices in several ways. The differences may be attributed to anatomical and morphological differences in the vocal-tract geometry, less precise control of the articulators, and a less refined ability to control suprasegmental aspects as prosody. These aspects induce major differences in children speech, higher fundamental and formant frequencies, greater spectral variability, slower average speaking rate, higher variability in speaking rate and higher degree of spontaneity [Lee et al. 1997; Potamianos and Narayanan 2007]. It is a well known fact that the fundamental frequency of children's voices is much higher than for adults, where average values of 130Hz for adult males, and 220Hz for adult females can be found. No statistically significant gender difference exists for children below twelve. Children's voices are also known to have much higher formant frequencies (specially for the second and third formants), attaing values above 4 kHz. The boundary values of the phonetic vowel space decrease with age, becoming more compact, and leading to a decrease in dynamic range of the formants values and to a decrease of the variability of spectral values. A 5-year-old child presents values of formants 50% higher than an adult male. Whereas in adults there are typically 3–4 formants in the 0.3–3.2kHz range, for children one can find only 2–3 formants in this range.

These formant characteristics motivate the use of adaptation procedures such as vocal tract length normalization when recognizing children's voices. For a review of methods to help recognition systems deal with children's voices, see Potamianos and Narayanan [2003]. The recognition results may also be severely affected for children's voices, when the input speech has a reduced bandwidth [Russell et al. 2007].

The differences become less marked in the process of growing up. During puberty, the male glottis changes so that the pitch frequency is lowered about one octave. This change sometimes occurs over just a couple of weeks. The pitch drop usually occurs from age eleven to age thirteen and there is no significant pitch change after fifteen. No abrupt changes are observed for girls, where the pitch drop from age seven to age twelve is significant, indicating that the laryngeal growth ends around that age. In another study [Ajmera 2006] it is shown that for male speakers, pitch drops 78% between the ages 12 to 15, and after that there are no significant changes. For female speakers, pitch drops between ages 7 and 12, and stops after. The changes in female speech are more gradual than in male speech, and the main differences become more significant after age 12.

The size of the vocal tract develops somewhat similarly for boys and girls in this age range [Wilpon and Jacobsen 1996]. Potamianos and Narayanan [2003] report an almost linear scaling of formant frequencies with age. The scale presents a significant divergence in male/female after puberty, showing the differences in physical changes between male and female speakers. Another thing that changes with age is the internal control loops of the articulatory system [Sundberg 1987].

Most gender/age classification methods exploit the above mentioned differences in more or less explicit ways. The features most typically found are pitch, formants, Mel-Frequency Cepstral Coefficients (MFCC), Perceptual Linear Prediction Coefficients (PLP), autocorrelation coefficients, linear prediction coefficients (or equivalent), etc. The slower average speaking rate of children is also a motivation for including delta, RASTA (log-RelAtive SpecTral), or any other temporal modeling coefficients in the feature set. This large number of features often motivates the adoption of dimensionality reduction approaches.

Gender classifiers using Gaussian mixture models (GMM), Hidden Markov models (HMM), or multilayer perceptrons (MLP) were proposed and tested with results about 95% of accuracy. Most often, these results concern only male/female distinction, and no

Table I.
Age and gender classes of the aGender corpus, where f and m abbreviate female and male, and x represents children without gender discrimination. The last two columns represent the number of speakers/utterances per set (train and develop).

class	group	age	gender	# train	# develop
1	C hild	7-14	x	68/4406	38/2396
2	Y oung	15-24	f	63/4638	36/2722
3	Y oung	15-24	m	55/4019	33/2170
4	A dult	25-54	f	69/4573	44/3361
5	A dult	25-54	m	66/4417	41/2512
6	S enior	55-80	f	72/4924	51/3561
7	S enior	55-80	m	78/5549	56/3826

children voices are considered. Although frequently adopted for visual gender detection, Support Vector machines (SVM) are not so popular for audio gender detection.

The comparison of the results reported in the literature is hindered by the fact that they have all been obtained with different corpora. A fundamental step in the joint evaluation of this type of classifiers was recently undertaken in the Interspeech 2010 conference, where common corpora were distributed and several gender/age subchallenges were conducted [Schuller et al. 2010; Meinedo and Trancoso 2010].

3. CORPORA

Five different corpora were used to train and evaluate the performance of the developed speech/nonspeech, age and gender detection systems. The aGender corpus [Burkhardt et al. 2007; Schuller et al. 2010], the CMU Kids corpus [Eskenazi et al. 1997], the PF STAR children corpus [Batliner et al. 2005], the BN ALERT corpus [Meinedo 2008] and the I-DASH CA corpus. All corpora were preprocessed in order to boost the energy levels and remove unwanted silence. In order to match the sampling frequency of the aGender corpus audio files, all corpora were downsampled from 16kHz to 8kHz.

3.1. aGender

The aGender corpus [Burkhardt et al. 2007; Schuller et al. 2010] was supplied to the participants of the InterSpeech 2010 Paralinguistic Challenge to assist in the development of speaker age and gender detection systems. It consists of 49 hours of telephone speech, involving 795 speakers, which are divided into train (23h, 471 speakers), development (14h, 299 speakers) and test sets (12h, 25 speakers). In our work, this partitioning was respected with the train set being used for training of age and gender systems and the development set being used for the calibration, fusion and evaluation.

The subjects received written instructions and were asked to call the recording system six times. Each time they were prompted by an automated Interactive Voice Response system to repeat given utterances or produce free content, in response to elicited questions (in German). The recordings of each subject were made using a mobile phone alternating indoor and outdoor to obtain different recording environments. Each of the six recording sessions contained 18 utterances.

Four age groups—Child, Youth, Adult, and Senior—are considered, with borders defined by the suppliers of this corpus. The choice of these borders was not motivated by any physiological aspects that might arise from the development of the human voice with increasing age, but solely on market aspects. In terms of gender, since children are not subdivided into female and male, there are 3 classes: Male, Female, and Child. This results in seven classes, as shown in Table I. The seven classes are combined into

only 4 classes for the purpose of age classification, and into 3 classes for the purpose of gender classification.

3.2. CMU Kids

The CMU Kids corpus [Eskenazi et al. 1997] comprises English sentences read aloud by children from 6 to 11 years old. It was recorded in a controlled environment. It consists of 24 male and 52 female speakers totaling approximately 9 hours. All the available speech data was used as training material both for age and gender systems.

3.3. PF STAR Children

The PF STAR Children corpus [Batliner et al. 2005] was provided by the KTH Research group. Similar to CMU-Kids, this corpus was also recorded in a controlled environment, but includes more diversity of speakers (108 male and 91 female children). The speakers age ranges from 4 to 8 years old. This corpus has 2 types of recordings, each with approximately 9 hours of speech, one recorded with headset and the other with a desktop microphone. As expected, the energy level of the second type of recordings is much lower and some reverberation effects can be perceived. Both types of recordings were used for training age and gender systems. The recordings are mostly composed by children talking after an adult that reads aloud. In terms of speaking style this is very similar to a read speech corpus with a slower speaking rate. The languages of the recordings are British English, German and Swedish.

3.4. BN ALERT

The BN ALERT corpus [Meinedo 2008] was the first European Portuguese Broadcast News corpus. It is composed of recordings from the RTP public TV station. This corpus was used for training speech/nonspech, age/gender and background conditions detection systems, since it is labelled according to the presence of speech, background conditions (clean, music, noise) and has information about the gender of the speakers but not about their age. We used as training data three different sets (train, pilot and devel) consisting of 57 hours with 1182 male and 508 female speakers.

There are very few children segments in Broadcast News. In the BN ALERT corpus there are only around 300 seconds of speech from children which is clearly insufficient, for example, to develop a reliable phonetic classification module or an automatic speech recognizer (ASR). An analysis of the confidence values for some of these children segments when recognized by our best European Portuguese ASR system, which has 18% word error rate, revealed an extremely high word error rate, on average above 60%, much higher than comparable female speech segments and also low word confidence scores. The lack of suitable child speech training data, comparable to male and female in terms of conditions and speech styles normally present in Broadcast News, prevents us from using phonetic classifiers or ASR systems as components in our age or gender detection systems.

3.5. I-DASH CA

The I-DASH CA corpus is composed by audio extracted from CA domain videos, provided by the Dutch Police forces. Although this constitutes a huge amount of data (approximately 1384 hours), none of it was originally hand labeled.

Compared to BN data, these CA recordings have very low quality, being characterized by a very low signal-to-noise ratio, also affected by the distance of the subjects from the video camera microphone. It is also worth pointing out that, unlike BN data, where speakers talk in relatively large utterances, CA videos are mostly characterized by short duration speech events such as screaming and moaning, which makes it much harder

to detect. There is no information regarding the languages spoken in the recordings, since this CA material was mostly obtained from the internet.

For evaluation purposes, a test set was hand labeled, marking segments containing male/female/child voices and marking background music. This test set contains audio from 60 different CA videos with a total duration of almost 6 hours. The amount of speech in these videos is relatively very small: 22 minutes of speech, of which 36% are male voices, 8% female voices, and 55% children voices. In terms of background music, over half of the whole test data contains music (3 hours and 10 minutes).

4. SPEECH/NONSPEECH DETECTOR

Of the different type of cues that can be extracted from the audio signal, the one that was considered most relevant was the identification of child voices. For this purpose, our audio analysis tool first detects if the audio signal contains speech, using a speech/nonspeech segmentation and classification module. Later, the identified speech segments are classified according to the age/gender in order to see if they contain children voices. The speech/nonspeech identifies potentially irrelevant materials (from the acoustic point of view) from the large amount of video data set, thus helping to reduce human labor as much as possible. This is accomplished by dividing the audio signal into smaller segments that contain speech or not. The output of this module is represented by a list with a start and end time for each segment, a tag indicating the classification given to the segment, speech or nonspeech and its confidence value.

This module is composed by three blocks. The first one is the feature extraction block that does the acoustic parametrization of the audio signal, taking into account spectral characteristics, 12th order Perceptual Linear Prediction (PLP) [Hermansky et al. 1992] coefficients plus energy, and temporal characteristics: delta coefficients. These features are then passed to the classification block which is implemented using an artificial neural network of the MLP type [Meinedo 2008]. This neural classifier was trained using the BN ALERT corpus, and also 41 hours of varied music and sound effects to improve the representation of nonspeech audio signals. The output of the trained neural classifier represents the probability of the audio signal containing speech. The third block in the speech/nonspeech detector is a finite state machine that receives as input the probability of the audio signal being speech. This block smoothes the input signal, using a median filter with a small window. This smoothed signal is thresholded and analyzed using a time window (tMin). The finite state machine uses 4 possible states (Probable Nonspeech, Nonspeech, Probable Speech, and Speech). If the input audio signal has a probability of speech above a given threshold the finite state machine is put into Probable Speech state. If after a given time interval (tMin) the average speech probability is above an average given confidence value, the machine changes to the Speech state. Otherwise it goes to the Nonspeech state. The finite state machine generates segment boundaries for nonspeech segments larger than the resolution of the median window. Additionally, nonspeech segments larger than tMin are discarded. The tMin value is an open parameter of the system and was optimized to maximize the nonspeech detected. The speech/nonspeech multilayer perceptron classifier was trained with 19 epochs of backpropagation stochastic mode.

5. AGE DETECTOR

The goal of our age detection system is to detect the age of the speakers in four separate classes: Child, Young, Adult, Senior: C, Y, A, S. The age detector was trained using the aGender corpus, our only source that contains age information for the Young, Adult and Senior classes and also the additional child corpora (CMU KIDS and PF STAR).

The developed age detection systems output the results in one of these seven classes which are then combined to produce the required four age classes. This is achieved by

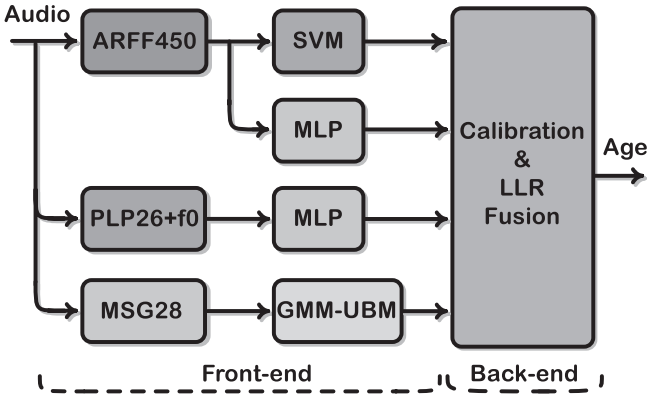

Fig. 1. Age detection system.

adding the output probability scores of female and male for each of the age classes, Young, Adult and Senior. See Table I for the age/gender class definition ($\{1, \ldots, 7\}$).

$$\begin{aligned} p(C) &= p(1) \\ p(Y) &= p(2) + p(3) \\ p(A) &= p(4) + p(5) \\ p(S) &= p(6) + p(7). \end{aligned} \tag{1}$$

Our approach for the age detection system (Figure 1) uses several separate age detection frontend systems, that take advantage of different features, classification paradigms, and training datasets. The output scores of each of these frontend systems is then calibrated and combined to produce the final system output. The motivation for having several frontends with different properties is that the diversity will improve the combination and ultimately will lead to a more robust age detection system.

5.1. FrontEnds

In this section we describe the developed frontend age detection systems.

The first two frontends used as training data the features described in Schuller et al. [2010], here denoted as "ARFF450," obtained in the aGender training set. These features were given to the participants of the InterSpeech 2010 Paralinguistic Challenge. They were obtained with the open-source Emotion and Affect Recognition toolkit's feature extracting backend openSMILE [Eyben et al. 2009]. 1582 acoustic features and transliteration, including those capturing nonlinguistic characteristics, are obtained in total by systematic "brute-force" feature (over)generation in three steps: first, the 38 low-level descriptors shown in Table II are extracted and smoothed by simple moving average low-pass filtering. Next, their first order regression coefficients are added. Then, 21 functionals are applied (cf. Table II). However, 16 zero-information features (e.g., minimum F0, which is always zero) are discarded. Finally, the two single features F0 number of onsets and turn duration are added. Due to the size of the aGender corpus, a limited set is provided, consisting of 450 features. This is reached by reducing the number of descriptors from 38 to 29, and that of functionals from 21 to 8.

Two different classification paradigms that take as input the "ARFF450" features were used. Support Vector Machines, for which we used the toolkit LibSVM,[1] and Multi-Layer Perceptrons, for which we used our own simulator [Meinedo 2008]. The

[1]http://www.csie.ntu.edu.tw/cjlin/libsvm/index.html

Table II.
ARFF450 feature sets: 38 low-level descriptors with regression coefficients, 21 functionals. Details in the text. A indicates those used only for the TUM AVIC baseline. Abbreviations: LSP: line spectral pairs, Q/A: quadratic, absolute.

Descriptors	Functionals
PCM loudness -	Position max./min. -
MFCC [0-14]	arith. mean, std. deviation
log Mel Freq. Band [0-7] -	skewness, kurtosis
LSP Frequency [0-7]	lin. regression coeff. 1/2 -
F0	lin. regression error Q/A -
F0 Envelope	quartile 1/2/3 -
Voicing Prob.	quartile range 2-1/3-2/3-1 -
Jitter local	percentile 1/99
Jitter consec. frame pairs	percentile range 99-1
Shimmer local	up-level time 75/90 -

SVM frontend uses a linear kernel, and the MLP front-end uses a fully connected feed-forward architecture with two hidden layers of 50 sigmoidal units and softmax outputs.

The third front-end extracts from the audio every 10ms a frame with 12th order Perceptual Linear Prediction (PLP) [Hermansky et al. 1992] coefficients plus energy plus deltas plus pitch (F0). The slower average speaking rate of children and senior relative to adults is a strong motivation for including delta, PLP and other temporal modeling coefficients in the feature set. Experiments with higher order PLP did not lead to improved results, possibly because of the small quantities of training data when compared with usual speaker recognition evaluation campaigns which typically use thousands of hours of speech material. The same applies to the use of double-deltas in our feature set. This frontend takes advantage not only of acoustic, but also of prosodic features (pitch) both at frame level (short term features), and at utterance level (long term functional ARFF features).

The fourth frontend extracts from the audio a frame of 28 static modulation spectrogram [Kingsbury et al. 1998] features.

For the third and fourth frontends, the classification paradigm used was the MLP which takes as input context seven contiguous frames of features and has two hidden layers of 100 and 50 units. This configuration of hidden units was the one that achieved the better classification scores.

The second, third, and fourth frontends share the same training data (the aGender dataset) and the same classification paradigm (MLP). The three frontends use different feature extraction methods (ARFF450, PLP26+F0 and MSG28). This enables the comparison of the performance of the different feature sets for this task.

The fifth front-end uses again the aGender training set and uses the MSG28 features. The classification paradigm used was different from the first ones, Gaussian Mixture Models—Universal Background Model (GMM-UBM) with 1024 mixtures. After training the UBM, each of the age class GMMs was created by performing five iterations of Maximum a Posteriori (MAP) adaptation.

Finally, the last two front-ends developed for detecting age used all available training data that has age labels, that is, the aGender training set plus the CMU Kids, PF STAR children head-mount and desktop sets. These last three sets are herein simply referred to as "child" data. By using the additional child data, we are promoting diverseness and ultimately better combination scores. The sixth frontend uses PLP26+F0 features and an MLP classifier and the seventh uses MSG28 features and a GMM-UBM classifier.

Other pairings of features and machine learning methods were tested but the reported configurations were the ones which lead to better combination results in the

Table III. Age Results Obtained in the aGender Development Set

Front-ends $\{1,\ldots,7\} \rightarrow \{C, Y, A, S\}$		% UA	% WA
a	SVM - ARFF450 - aGender	47.4	47.0
b	MLP - ARFF450 - aGender	46.7	48.1
c	MLP - PLP26+F0 - aGender	48.6	47.9
d	MLP - MSG28 - aGender	46.0	46.2
e	GMM-UBM - MSG28 - aGender	39.3	44.0
f	MLP - PLP26+F0- aGender+child	49.2	47.5
g	GMM-UBM - MSG28 - aGender+child	39.7	44.5
Fusions $\{1,\ldots,7\} \rightarrow \{C, Y, A, S\}$		% UA	% WA
h	a + b	47.9	48.6
i	f + g	50.2	49.1
j	a + b + f + g	51.2	50.6

aGender development set. The motivation for having frontends with different properties is that the diversity improves the combination and ultimately will lead to a better final system.

5.2. Calibration and Fusion Backend

Linear logistic regression fusion and calibration of the developed frontend systems has been done with the FoCal Multiclass Toolkit.[2] The output log-likelihood ratio (llr) scores from this fusion backend were later converted into probabilities, which is more meaningful in terms of human analysis. This was achieved by scaling the scores to produce confidence values with Expression (2).

$$p(score(t)) = \frac{e^{score(t)}}{\sum_k e^{score(k)}}. \tag{2}$$

Several experiments of fusing the different frontends were tested. The ones that obtained better results are presented in Table III and discussed in the next section.

5.3. Results

Table III summarizes the results obtained in the aGender development set by the different frontends individually and by the combination of them using the calibration and llr fusion back-end. Results are expressed in terms of Unweighted and Weighted Accuracy on average per class (% UA and %WA). The former (%UA) is the relevant measure since the distribution among different classes is not well balanced [Schuller et al. 2010].

Comparing the different frontends we observe that ARFF450 and MSG28 features obtained similar results ("b" and "d" frontends), which are lower than the results obtained by using PLP26+F0 features (frontend "c"). We also observe that in this task the GMM-UBM obtained worse results when compared with MLP and SVM classifiers. By comparing the performance of the "c" and "f" frontends, we observe that expanding the training set with additional child data was benefic. The same is true for the MSG28 features and GMM-UBM classifiers of "e" and "g" frontends. The best individual frontend "f" in terms of % UA combines short term acoustic and prosodic features (PLP26+F0) with an MLP classifier and uses an expanded training set.

The lower part of Table III represents the results obtained by the calibration and fusion of several frontends. The fusion identified with letter "h" combines the first two frontends which were trained using the ARFF450 features. The second fusion, "i", combined the two frontends that were trained using the aGender and the additional child data (frontends "f" and "g"). Finally, the best results were obtained by combining

[2]http://niko.brummer.googlepages.com/focalmulticlass

Table IV. Age Confusion Matrix Results for the Final Fusion System Obtained in the aGender Development Set

	C	Y	A	S
C	**57.5**	24.0	11.1	7.4
Y	8.5	**50.7**	24.4	16.4
A	2.6	24.2	**39.9**	33.3
S	2.5	12.6	28.1	**56.7**

four frontends, the two ones that used ARFF450 features ("a" and "b") and the two ones that used all training data that had age group labels ("f" and "g"). Fusion of these four systems represents an improvement of 2% absolute over the best individual frontend.

An analysis of Table IV shows that there are some confusions between neighboring classes especially in Adult/Young and Adult/Senior. This mix is somewhat expected, since in these age groups it is more difficult to establish a clear border and as such will ultimately lead to overlaps. We also see that although we are using very diverse speech material between classes in terms of languages and speaking styles, the final system is capable of detecting all four classes with more or less the same performance (around 50% accuracy) which means that the developed age detection systems is tolerant to variations in languages and speaking styles.

6. GENDER DETECTOR

The Gender detector is required to detect the gender of the speakers in three separate classes, child, female and male $\{x, f, m\}$. Since this is a much more well behaved task with a smaller number of classes and clearer borders between them, one expects to obtain higher accuracies. Another factor that might lead to better performance of our gender detection system is that we have available a larger training set with gender labels. As training data for some of the frontends we used the BN ALERT corpus [Meinedo 2008], which not only has more speaker variability, but also has more diverse audio background conditions. This increased variability may lead to a more robust gender detection system.

Our developed approach (Figure 2) for the gender detection system uses several separate frontends which again take advantage of different features, classification paradigms and different training datasets. The output scores of each of these frontend systems is then calibrated and combined to produce the final system output. The following section describes in detail each of the developed frontend systems.

6.1. Frontends

For the gender detection system, eight independent frontends were developed and tested. The first two front-ends used here were developed for detecting the age group and were described in detail in section 5. The first one is the fusion that used frame level features PLP26+F0 and MSG28 (denoted as "i" in Table III and Table V) and the second one is the best fusion (denoted as "j" in Table III and Table V). In order to use these age detection systems it was necessary to convert its seven class output probability scores into the three class gender scores $\{1, \ldots, 7\} \rightarrow \{x, f, m\}$ by summing the female age probability scores together and the male age probability scores together. Again, the child class score is directly the score from age class 1. The other six developed frontend systems output directly the scores in the required three classes.

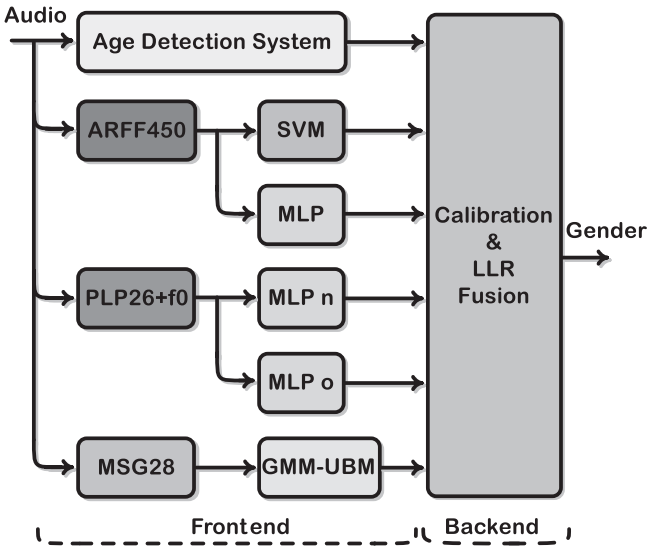

Fig. 2. Gender detection system.

Table V. Gender Results Obtained in the aGender Development Set

Front-ends $\{x, f, m\}$		% UA	% WA
i	Age Detection $\{1, \ldots, 7\} \rightarrow \{x, f, m\}$	80.1	89.1
j	Age Detection $\{1, \ldots, 7\} \rightarrow \{x, f, m\}$	80.6	89.3
k	SVM - ARFF450 - aGender	76.9	86.4
l	MLP - ARFF450 - aGender	76.5	86.5
m	MLP - PLP26+F0 - aGender	78.0	88.8
n	MLP - PLP26+F0 - aGender+child	78.9	89.5
o	MLP - PLP26+F0 - aGender+child+BN	82.2	88.2
p	GMM-UBM - MSG28 - aGender+child+BN	75.9	84.1
Fusions $\{x, f, m\}$		% UA	% WA
q	i + n + o + p	82.8	86.6
r	j + k + l + n + o + p	83.1	86.9

$$\begin{aligned} p(x) &= p(1) \\ p(f) &= p(2) + p(4) + p(6) \\ p(m) &= p(3) + p(5) + p(7). \end{aligned} \tag{3}$$

The third and fourth gender frontends, respectively "k" and "l" in Table V, used as training data the ARFF450 features obtained in the aGender training set. Two different classification paradigms were used: SVM (linear kernel), and MLP. The latter was used without input context since a single frame of features represents the whole audio file [Schuller et al. 2010]. Our experiments used a fully connected feed-forward architecture with two hidden layers of 100 and 50 sigmoidal units, and softmax outputs. The fifth frontend "m" also uses an MLP classifier but instead receive PLP26+F0 features as input data and has seven input context frames and two hidden layers of 350 units.

The other three frontends developed for detecting speakers gender used additional training data besides the aGender corpus. The sixth frontend (denoted "n" in Table V) also used the other "child" corpora (CMU Kids and PF STAR children head-mount and desktop sets). The seventh and the eighth frontends (denoted "o" and "p" in Table V)

used all available training data that has gender labels, that is, all of the above (aGender, CMU Kids and PF STAR children head-mount and desktop sets) plus the BN ALERT training, pilot and development sets.

In terms of feature extraction and classification methods, the sixth "n" and the seventh "o" frontends extract from the audio every 10ms a frame with 12th order PLP [Hermansky et al. 1992] coefficients plus energy plus deltas plus pitch (F0). Both use MLP classifiers with seven input context frames and two hidden layers of 350 units.

The eighth frontend "p" extracts from the audio a frame of 28 static modulation spectrogram [Kingsbury et al. 1998] features. A GMM-UBM with 1024 mixtures is employed. After training the UBM, each of the gender class GMMs was created by performing five iterations of Maximum a Posteriori (MAP) adaptation.

6.2. Calibration and Fusion Backend

Similar to the age detection system, linear logistic regression fusion and calibration of the independent frontend systems has been done with the FoCal Multiclass Toolkit.[3] The output scores from this fusion backend were converted to probability confidence values in order to be more meaningful when a human analysis is required. The conversion was achieved using Expression (2).

Several experiments of fusing the different front-ends were tested. The ones that obtained better results are presented in Table V and discussed in the results section.

6.3. Results

Table V summarizes the results obtained in the aGender development set by the eight independent gender frontends and by their combinations using the calibration and llr fusion backend. Results are expressed in terms of accuracy (% UA which is the relevant measure and % WA [Schuller et al. 2010]).

Comparing the different frontends we observe that the two age detection systems are also very good at detecting the gender. The front-ends with ARFF450 features, "k" and "l" have lower accuracies than the ones obtained by using PLP26+F0 features in frontend "m," which is similar to what happened in the age results. Here we also observe that in this task GMM-UBM obtained worse results when compared with MLP classifiers. By comparing the performance of the "m," "n" and "o" frontends we observe that expanding the training set with additional data (child data and also and Broadcast News male and female speech data in the "o" frontend) was benefic . The best individual frontend "o" in terms of % UA combines short term acoustic and prosodic features (PLP26+F0) with an MLP classifier and uses the expanded training set.

The lower part of Table V represents the results obtained by the calibration and fusion of several frontends. The first fusion identified with letter "q" combines four frontends trained only with frame level features, PLP26+F0 and MSG28. The first age detection fusion "i" and the three frontends that used expanded training data (aGender + child + BN), "n," "o," and "p." Finally, the best results were obtained by combining six frontends, the best age detection system "j," the two ones that used ARFF450 features, "k" and "l" and the three ones that used expanded training data with gender labels, "n", "o" and "p". Fusion of all six systems represent an improvement of 1% absolute over this best single frontend.

Inspection of Table VI reveals that the biggest misclassifications come from the child (x) and female (f) classes. This is somewhat expected, since these two types are more similar than the male gender. In fact, the male class (m) detection has excellent results. We suspect that this is also because the training sets are not balanced, and there is more male training data. We also see that, although we are using very diverse speech

[3]http://niko.brummer.googlepages.com/focalmulticlass

Table VI. Gender Confusion Matrix Results for the Final Fusion System Obtained in the aGender Development Set

	x	f	m
x	**70.5**	20.4	9.1
f	14.9	**83.8**	1.3
m	1.7	3.3	**95.0**

material between classes in terms of languages and speaking styles, the final system is capable of detecting all three classes with very good performance which means that the developed gender detection systems is not biased by variations in language or speaking styles.

7. PERCEPTUAL TEST

The results of the two preceding sections in terms of age and gender automatic classification should be compared with the ones obtained by human subjects in similar conditions. This comparison is specially relevant when we take into account the borders between the 4 age groups of the aGender corpus. For this purpose, we set up a small human benchmark with only 50 samples of this corpus, randomly chosen among recordings of the 7 classes. The subjective test was done only by 3 native Portuguese listeners with limited knowledge of the target language. The raters were asked to classify each sample according to gender and age group.

The test revealed great difficulties in distinguishing between neighboring classes. In terms of gender detection the human listeners obtained classification accuracies between 80 and 86%. The majority of confusions (61.5%) occurred between child and female voices. The other misclassifications, distinguishing between child and male and distinguishing between female and male both represented 19.2% of the total errors.

In terms of age group detection the human listeners obtained much lower accuracies, between 40% and 46%. The majority of misclassifications involved neighbourging classes, as expected: adult and senior (35.6%), young and adult (28.7%), child and young (18.4%). In particular, senior female voices were always confused with adult female voices, although senior male voices seemed much easier to identify. Creaky voices and slower speaking rates were major cues in the identification of senior voices. On the other hand, young male voices were almost always confused with adult male voices.

This small test enabled us to have a much better understanding of how good human listeners are at assessing gender and age from recorded telephone voice samples. With this test we conclude that human listeners had difficulties distinguishing between neighboring classes specially in the age detection, which gives us a different perspective on the automatically obtained results.

8. EVALUATION USING CA DATA

For evaluating the results of the speech/nonspeech module on CA data, the frame Classification Error Rate (CER) measure was used [Meinedo 2008]. The CER measure is the ratio between the misclassified speech frames and the total number of audio frames. The CER in our CA test set is 4.0% with an average precision (at recall 10%) of 100%. These results may be considered very good given the demanding conditions of the CA genre.

Although the results for age/gender detection shown in the previous section were very promising, they are not so good for the CA genre. For the age/gender classification the % UA result in our CA test set is 62.2% and the % WA is 68.1%, with an average precision of 100% for both male and female and 59.5% for child. Part of these errors may be attributed to misclassification by the speech/nonspeech detector. For the most

part this may be attributed to the extremely difficult conditions such as the presence of background music and reverberation due to the recording by the (distant) microphone of the video camera.

9. AUTOMATIC SELECTION OF TRAINING DATA

Accurate user feedback in appropriate quantities is always a good way to improve classification in data driven methods such as the ones used by the Audio Analysis tool. In our case we concluded that an indication that a video clip contains CA material is not enough for the purpose of retraining the audio models, as the relevant cues may derive only from the image signal. Furthermore our experience showed that small amounts of hand labeled material were not enough to achieve an increase in accuracy of the new models and that imprecise user feedback could contaminate the data and make the training more unstable, usually producing worse models than the baseline ones. To circumvent these difficulties, we investigated a way of improving the Audio Analysis tool using in domain CA data and without requiring expensive and time consuming user feedback. The idea is to use the I-DASH CA corpus data in an unsupervised fashion, that is, use the baseline Audio Analysis tool to automatically classify the data and use for training just the audio frames which have a high degree of confidence in the classification. We have experience in using this technique to improve audio phonetic classification systems with very good results without the need for more hand labeled data, provided that the baseline system has a reasonable performance [Meinedo 2008]. This type of unsupervised training is used by several other authors in acoustic modeling of automatic speech recognition systems [Wessel and Ney 2005; Ma et al. 2006; Wang et al. 2007]. A maximum entropy criterion is also used [Wu et al. 2007] instead of confidence measure to select the unsupervised training material.

The first step was to classify the 1384 hours of audio from the I-Dash CA corpus using the speech/nonspeech detector. The second step was to choose the segments classified as speech with a confidence above 90%. This is a conservative threshold which cuts out the majority of the audio data but has the advantage of introducing very few classification mistakes in the training since we are only choosing the segments where the tool has a very high degree of confidence in the classification. From this selection procedure approximately 75 hours of speech data were used to enrich the training set. To complement the increase of speech material in the training set, we also chose 75 hours of nonspeech data also with a confidence above 90%. These new 150 hours of automatically annotated audio data were used to retrain the Speech/Nonspeech classifier and the Gender classifier.

This unsupervised training is an iterative process where after the initial selection phase new models are trained and then the whole data is classified again using the new models. The process continues until the new trained models no longer provide an increase in accuracy.

9.1. Results

Table VII summarizes the results obtained by the two classifiers, for each of the three iterations of unsupervised training with automatically selected data. The results are also graphically displayed in Figures 3 and 4. The retraining with in domain data provided a welcome decrease in CER of the speech-nonspeech classifier, which was already quite low. The third iteration produced no further improvements. Concerning the gender classifier, unsupervised training brought a relative improvement of around 10%.

Part of the errors present in the gender detector may be attributed to misclassification by the speech/nonspeech detector. For the most part this may be attributed to the

Table VII. Classification Error Rate (CER) in the CA Test Set

iteration	Speech/Non-Speech	Gender/Age
baseline	4.0	31.4
1st	3.8	29.1
2nd	3.6	28.4
3rd	3.6	28.3

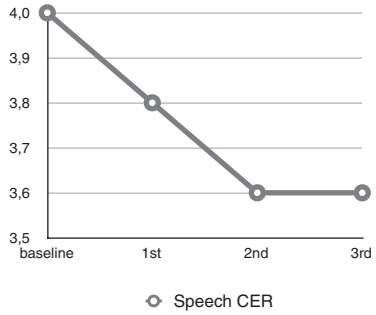

Fig. 3. CA test set results (Speech-nonspeech).

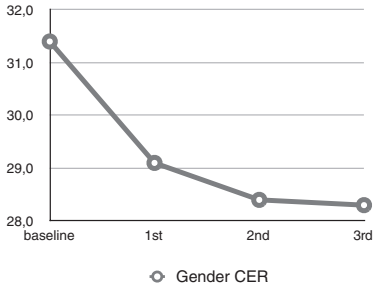

Fig. 4. CA test set results (Gender/Age).

previously mentioned extremely difficult conditions, such as the presence of background music and reverberation, and the very short duration of important audio events.

10. CONCLUSIONS

This article presented the age and gender classification systems that compose the audio analysis tool in the European I-DASH project. Our participation in this project involved the development of an age detection system, an area where we had no prior experience, and resulted in the development of a much improved age/gender detection system.

These age and gender classification systems are composed respectively by the fusion of four and six individual subsystems trained with short and long term acoustic and prosodic features, different classification paradigms (GMM-UBM, MLP and SVM) and different speech corpora. The complementary nature of these different approaches boosted their combination performance. The best results obtained by the calibration and linear logistic regression fusion backend show an absolute improvement of 2% on the unweighted accuracy value for the age evaluation and 1% for the age/gender evaluation when compared to the best individual frontend systems. The final age/gender detection system evaluated using our CA test set achieved promising results given the extremely difficult conditions of the CA video material. Unsupervised retraining brought a 10% relative improvement without requiring expensive hand labeled data.

This is specially significant if one takes into account the great difficulties in manually labeling CA data.

ACKNOWLEDGMENTS

The authors would like to thank Mats Blomberg and Daniel Elenius for letting us use the KTH PF STAR children corpus, and Felix Burkhardt and colleagues for the aGender corpus. We would also like to thank the reviewers for their very helpful comments.

REFERENCES

AJMERA, J. 2006. Effect of age and gender on LP smoothed spectral envelope. In *Proceedings of the Speaker and Language Recognition Workshop.* IEEE Odyssey 2006.

BATLINER, A., BLOMBERG, M., D'ARCY, S., ELENIUS, D., GIULIANI, D., GEROSA, M., HACKER, C., RUSSELL, M., STEIDL, S., AND WONG, M. 2005. The PF STAR childrens speech corpus. In *Proceedings of Interspeech*.

BUGALHO, M., PORTELO, J., TRANCOSO, I., PELLEGRINI, T., AND ABAD, A. 2009. Detecting audio events for semantic video search. In *Proceedings of InterSpeech*.

BURKHARDT, F., ECKERT, M., JOHANNSEN, W., AND STEGMANN, J. 2007. A database of age and gender annotated telephone speech. In *Proceedings of the Language and Resources Conference (LREC).*

ESKENAZI, M., MOSTOW, J., AND GRAFF, D. 1997. The CMU kids corpus. In *Linguistic Data Consortium*. Philadelphia, PA.

EYBEN, F., WOELLMER, M., AND SCHULLER, B. 2009. openEAR - introducing the munich open-source emotion and affect recognition toolkit. In *Proceedings of the International Conference on Affective Computing and Intelligent Interaction (ACII)*.

HERMANSKY, H., MORGAN, N., BAYA, A., AND KOHN, P. 1992. RASTA-PLP speech analysis technique. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* .

KINGSBURY, B. E., MORGAN, N., AND GREENBERG, S. 1998. Robust speech recognition using the modulation spectrogram. *Speech Comu. 25*, 117–132.

LEE, S., POTAMIANOS, A., AND NARAYANAN, S. 1997. Analysis of children's speech: Duration, pitch and formants. In *Proceedings of the EuroSpeech.*

MA, J., MATSOUKAS, S., KIMBALL, O., AND SCHWARTZ, R. 2006. Unsupervised training on large amounts of broadcast news data. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*.

MEINEDO, H. 2008. Audio pre-processing and speech recognition for broadcast news. Ph.D. thesis, IST, Lisboa, Portugal.

MEINEDO, H. AND TRANCOSO, I. 2010. Age and gender classification using fusion of acoustic and prosodic features. In *Proceedings of Interspeech*.

POTAMIANOS, A. AND NARAYANAN, S. 2003. Robust recognition of children's speech. *IEEE Trans. Speech Audio Process. 11,* 6, 603–616.

POTAMIANOS, A. AND NARAYANAN, S. 2007. A review of the acoustic and linguistic properties of children's speech. In *Proceedings of the International Workshop on Multimedia Signal Processing (MMSP).*

RUSSELL, M., D'ARCY, S., AND QUN, L. 2007. The effects of bandwidth reduction on human and computer recognition of children's speech. *IEEE Signal Process. Lett.*, 1044–1046.

SCHULLER, B., STEIDL, S., BATLINER, A., BURKHARDT, F., DEVILLERS, L., MUELLER, C., AND NARAYANAN, S. 2010. The interspeech 2010 paralinguistic challenge. In *Proceedings of Interspeech*.

SUNDBERG, J. 1987. *The Science of the Singing Voice*. Northern Illinois University Press, Dekalb Illinois.

WANG, L., GALES, M. J. F., AND WOODLAND, P. C. 2007. Unsupervised training for mandarin broadcast news and conversation transcription. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*.

WESSEL, F. AND NEY, H. 2005. Unsupervised training of acoustic models for large vocabulary continuous speech recognition. In *IEEE Trans. Speech Audio Process.*

WILPON, J. AND JACOBSEN, C. 1996. A study of speech recognition for children and the elderly. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*.

WU, Y., ZHANG, R., AND RUDNICKY, A. 2007. Data selection for speech recognition. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*.

Received June 2010; revised November 2010; accepted January 2011