Available online at www.sciencedirect.com

SciVerse ScienceDirect

Computer Speech and Language 27 (2013) 151–167

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Automatic speaker age and gender recognition using acoustic and prosodic level information fusion☆,☆☆

Ming Li*, Kyu J. Han, Shrikanth Narayanan

Signal Analysis and Interpretation Laboratory (SAIL), Viterbi School of Engineering, University of Southern California, Los Angeles, CA 90089, USA

Received 20 April 2011; received in revised form 27 January 2012; accepted 28 January 2012
Available online 8 February 2012

Abstract

The paper presents a novel automatic speaker age and gender identification approach which combines seven different methods at both acoustic and prosodic levels to improve the baseline performance. The three baseline subsystems are (1) Gaussian mixture model (GMM) based on mel-frequency cepstral coefficient (MFCC) features, (2) Support vector machine (SVM) based on GMM mean supervectors and (3) SVM based on 450-dimensional utterance level features including acoustic, prosodic and voice quality information. In addition, we propose four subsystems: (1) SVM based on UBM weight posterior probability supervectors using the Bhattacharyya probability product kernel, (2) Sparse representation based on UBM weight posterior probability supervectors, (3) SVM based on GMM maximum likelihood linear regression (MLLR) matrix supervectors and (4) SVM based on the polynomial expansion coefficients of the syllable level prosodic feature contours in voiced speech segments. Contours of pitch, time domain energy, frequency domain harmonic structure energy and formant for each syllable (segmented using energy information in the voiced speech segment) are considered for analysis in subsystem (4). The proposed four subsystems have been demonstrated to be effective and able to achieve competitive results in classifying different age and gender groups. To further improve the overall classification performance, weighted summation based fusion of these seven subsystems at the score level is demonstrated. Experiment results are reported on the development and test set of the 2010 Interspeech Paralinguistic Challenge aGender database. Compared to the SVM baseline system (3), which is the baseline system suggested by the challenge committee, the proposed fusion system achieves 5.6% absolute improvement in unweighted accuracy for the age task and 4.2% for the gender task on the development set. On the final test set, we obtain 3.1% and 3.8% absolute improvement, respectively.

Keywords: Age recognition; Gender recognition; Prosodic features; Pitch; Harmonic structure; Formant; Polynomial expansion; Maximum likelihood linear regression; UBM weight posterior probability supervectors; GMM; SVM; Sparse representation; Score level fusion

1. Introduction

Automatic recognition of paralinguistic information, such as speaker identity, gender, age range, and emotional state, can guide human computer interaction systems to automatically understand and adapt to different user needs. Likewise

☆ Part of this work (Li et al., 2010) was presented at InterSpeech 2010, special session: Paralinguistic Challenge.
☆☆ This paper has been recommended for acceptance by Björn Schuller, Ph.D.

* Corresponding author. Tel.: +1 213 740 3477.
E-mail addresses: mingli@usc.edu (M. Li), kjhan@us.ibm.com (K.J. Han), shri@sipi.usc.edu (S. Narayanan).
URLs: http://www-scf.usc.edu/ mingli/ (M. Li), http://sail.usc.edu/ (S. Narayanan).

0885-2308/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.csl.2012.01.008

such meta-information can serve as an important analytic in human decision making. For instance, the emerging broad area of behavioral signal processing aims to create quantitative characterization of typical, atypical, and distressed human behavior states across a variety of application domains including in education and health care (Black et al., 2010; Lee et al., 2010). Information about age and gender can be an important ingredient that leads to rich behavioral informatics.

1.1. Background

Identifying the age and gender information of a speaker given a short speech utterance is a challenging task and has gained significant attention recently. Metze et al. (2007) compared four approaches for age and gender recognition from telephone speech; these included a parallel phoneme recognizer system to compare the Viterbi decoding scores for each category-specific phoneme recognizer, a system using dynamic Bayesian networks to combine several prosodic features, a system based solely on linear prediction analysis, and a GMM system based on MFCCs. It was reported in Metze et al. (2007) that the parallel phone recognizer system performs as well as human listeners on long utterances but its performance degrades on short utterances while the system based on prosodic features, such as fundamental frequency (F0), jitter, shimmer and harmonics-to-noise-ratio, has shown relative robustness to the variation of the utterance duration. More recently, novel acoustic features (Ajmera and Burkhardt, 2008), frame and utterance based acoustic-prosodic joint features (Spiegl et al., 2009; Meinedo and Trancoso, 2010; Gajšek et al., 2010; Eyben et al., 2009), lexical features (Wolters et al., 2009) as well as fuzzy SVM modeling (Nguyen et al., 2010) have all been proposed to improve the recognition performance. In Ajmera and Burkhardt (2008), the discrete cosine transform is applied to the cepstral coefficients and the cepstral trajectories corresponding to lower (3–14 Hz) modulation frequencies provide best discrimination. Prosodic features (pitch, energy, formants, vocal tract length warping factor, speaking rate, etc.) and their functionals can also be added to the cepstral features at the frame or utterance level to enhance the performance (Spiegl et al., 2009; Meinedo and Trancoso, 2010; Gajšek et al., 2010; Eyben et al., 2009; Wolters et al., 2009). In addition to the prosodic features, novel lexical level features like word-class frequencies have also been proposed for age recognition purpose (Wolters et al., 2009). In the fuzzy SVM modeling method proposed by Nguyen et al. (2010), a fuzzy membership is assigned as a weight to each training data point to increase the robustness against noise and outliers. Furthermore, techniques from speaker verification and language identification applications such as GMM–SVM mean supervector systems (Bocklet et al., 2008), nuisance attribute projection (NAP) (Dobry et al., 2009), anchor models (Dobry et al., 2009; Kockmann et al., 2010) and Maximum-Mutual-Information (MMI) training (Kockmann et al., 2010) have been successfully applied to speaker age and gender identification tasks to enhance the performance of acoustic level modeling. In Dobry et al. (2009), anchor modeling utilizes a back end SVM to model the distribution of similarity scores between training data and all the anchor speaker models. Due to the different aspects of modeling, combining different classification methods together can often significantly improve the overall performance (Müller and Burkhardt, 2007; van Heerden et al., 2010; Meinedo and Trancoso, 2010; Bocklet et al., 2010; Kockmann et al., 2010; Lingenfelser et al., 2010).

1.2. Specific research motivation and focus

In this paper, we focus on both the acoustic and prosodic level approaches for speaker age and gender identification. As acoustic level approaches, we consider two baseline systems: Gaussian mixture model (GMM) on short-time spectrum based mel-frequency cepstral coefficient (MFCC) features, and support vector machine (SVM) on GMM mean supervectors. We extend the latter (GMM–SVM mean supervector method) by using two kinds of supervectors, namely maximum likelihood linear regression (MLLR) matrix supervector (Stolcke et al., 2005) and UBM weight posterior probability (UWPP) supervector (Li et al., 2010; Zhang et al., 2009; Porat et al., 2010). Generally, in the GMM–SVM mean supervector method, maximum a posteriori (MAP) adaptation is used to adapt the means of a GMM Universal Background Model (UBM), and the corresponding feature vectors are the Gaussian mean supervectors (GSVs) which consist of the stacked adapted means. The MAP adaptation is performed for each utterance using the statistics collected on the UBM and typically only the means are adapted. The idea of MLLR, widely used in automatic speech recognition, is to estimate an affine transformation to adapt the acoustic model parameters of a speaker independent system towards a given speaker. Thus the MLLR matrix itself contains speaker specific characteristics and the entries of this affine transformation matrix can be used as feature supervectors for speaker modeling (Stolcke et al., 2005). MLLR can also

be applied to map the speaker-independent UBM model to a speaker dependent GMM model using the adaptation data (Stolcke et al., 2005, 2007). For each utterance, an affine transformation matrix is calculated to maximize the likelihood of the associated feature vector (Leggetter and Woodland, 1995). The columns of the corresponding MLLR matrices are stacked to form the MLLR matrix supervector and used for SVM modeling (Stolcke et al., 2007). We apply this idea to the age–gender recognition task. For the UWPP supervector modeling method, we utilize the way the expectation-maximization (EM) algorithm updates the weight of each Gaussian component when training an age and gender independent GMM–UBM model. It is shown in Zhang et al. (2009) that the utterances from different speakers generally should get different UWPPs on the same gaussian component. This inspired us to explore the potential to consider the UWPP supervector as a histogram describing the characteristics of different age and gender groups. It is shown in Jebara et al. (2004) that the Bhattacharyya probability product (BPP) kernel outperforms linear kernel on the word frequency feature vector on a text classification task. So we applied the BPP kernel for the SVM modeling on these UWPP supervectors. In summary, we introduce two additional GMM supervectors, namely MLLR and UWPP supervectors, as the features for SVM modeling in age and gender recognition. All these three supervector extraction methods share the same framework of using a GMM–UBM as a front end and therefore combining these approaches is efficient in terms of computational cost.

Another extension considered at the acoustic level is sparse representation of supervectors. In our recent work (Li and Narayanan, 2011), sparse representation computed by l^1-minimization with quadratic constraints was proposed to replace SVM in GMM mean supervector modeling and was demonstrated to achieve better performance compared to SVM in the robust talking face video verification task. The sparse representation classification is extended to model the low dimensional i-vectors (Dehak et al., 2011) in the speaker verification task and has been shown to be complementary with SVM modeling (Li et al., 2011). This inspired us to exploit the sparse representation approach to perform speaker age and gender identification on GMM supervectors. We first construct an over-complete dictionary using all the training supervector samples, and then calculate the sparsest linear representation via l^1-minimization for each test supervector sample. If the test sample is from the nth class, the test sample should have a sparse representation whose nonzero entries concentrate mostly on the dictionary samples from the nth class. Therefore, the membership of the sparse representation in the over-complete dictionary itself captures the discriminative information given sufficient training samples and large number of classes on the training data (Wright et al., 2008; Li and Narayanan, 2011; Li et al., 2011). It is similar to the anchor models (Dobry et al., 2009; Kockmann et al., 2010) that utilize the distribution of similarity scores between testing sample and all the anchor speaker models but without the backend SVM training process. Sparse representation modeling does not require training process which makes it suitable for large scale online adaptive learning. For SVM, if we want to add even one additional training sample, the whole model needs to be re-trained which is computationally expensive and prohibitive. However, the sparse representation approach demands a large amount of memory space due to the over-complete dictionary which can limit the training sample numbers and slow down the recognition process for large databases. Thus, when we exploit this sparse representation framework, GMM supervectors with good discriminative capability and small dimensionality are preferred. Among the three aforementioned supervector types (GMM mean, MLLR and UWPP), UWPP supervector has the smallest dimensionality (the number of components in GMMs) with comparable discriminative power. Thus, it is in this context that we examine the validity and feasibility of using sparse representation on GMM UWPP supervectors for the age and gender identification task.

We develop prosodic level approaches as well. As shown by Schötz (2007) in a phonetic study of speaker age, features related to speech rate, sound pressure level (SPL), fundamental frequency (F_0) and their temporal variations appear to be important correlates of speaker age. These prosodic features have been extensively studied for speaker age and gender recognition at both utterance and frame levels. Our extension is to model such prosodic information at the syllable level of a speech signal. One issue in prosodic information modeling at the syllable level is that, for each utterance, the number of syllables as well as the number of voiced speech segments are not fixed which makes the dimensionality of feature vectors vary. Therefore, statistical measures, such as mean, median, standard deviation and percentile, of these syllable-level voiced-segment-based prosodic feature vectors along all the segments are calculated and concatenated with other utterance level feature vectors (Bocklet et al., 2010; Schuller et al., 2010; van Heerden et al., 2010; Müller and Burkhardt, 2007; Spiegl et al., 2009). For instance, one of our three baseline subsystems (SVM–SMILE) is based on 450-dimensional utterance level features including acoustic, prosodic and voice quality information (Schuller et al., 2010). Similarly, prosodic features can also be combined with frame based features, such as MFCC and Perceptual Linear Predictive (PLP) features, to generate acoustic-prosodic joint frame level feature

vectors (Meinedo and Trancoso, 2010; Gajšek et al., 2010). However, each syllable size segment has its own prosodic patterns which might be averaged out by such aforementioned statistical measures along all the segments. Furthermore, the frame level prosodic features cannot effectively capture the dynamic variations along the whole syllable segment. Thus, in this paper, we assume each syllable segment to be independent and then model the age and gender dependent prosodic information directly at the syllable level using SVM. In testing, these syllable segment based score vectors are averaged at the utterance basis to generate the final decision. Moreover, rather than the statistical measures of prosodic information (Eyben et al., 2009), we employed polynomial expansion on all the interested contours and the duration (Dehak et al., 2007a,b) as our low level descriptors for each voiced speech segment. In our approach, not only pitch, energy, and formant contours which were explored in Dehak et al. (2007b) are adopted, but contours of frequency domain harmonic structure energy (Cao et al., 2007) are also considered for analysis. Since the contours can be reconstructed by these polynomial expansion coefficients, this feature set can capture rich prosodic information as well as the full picture of the variations within each segment. Common prosodic features, such as F_0, jitter, shimmer, formant, or even speech rate, can be captured or derived by our contour-based low-level descriptors. The proposed prosodic features and SVM modeling can provide complementary information to the acoustic level modeling at the score level fusion stage.

1.3. Summary of contributions

In summary, we address the speaker age and gender recognition problem with acoustic and prosodic level information fusion. The contributions are as follows: (1) At the acoustic level, we apply two additional GMM supervectors, namely MLLR and UWPP supervectors, as the features for SVM modeling. (2) The notion of sparse representation is introduced for GMM UWPP supervector modeling which is suitable for large scale online adaptive learning due to its property of no new training effort required. (3) Contours of pitch, time domain energy, frequency domain harmonic structure energy and formant for each syllable unit in every voiced speech segment are mapped into polynomial expansion coefficients and duration as the prosodic features. Then we model the age and gender dependent prosodic information directly at the syllable level using SVM. (4) Score level fusion of the proposed four subsystems (GMM–MLLR–SVM, GMM–UWPP–SVM, GMM–UWPP–Sparse representation, SVM–Prosody) as well as the three baseline subsystems (GMM, GMM–Mean–SVM, SVM–SMILE) is performed to improve the overall performance.

The remainder of the paper is organized as follows. A description of the corpus and classification task is provided in Section 2. Each subsystem as well as the score level fusion method is explained in Section 3. Experimental results and discussions are presented in Section 4. The conclusions are provided in Section 5.

2. Corpus and classification task

The database used to evaluate the proposed approach is the aGender database (Burkhardt et al., 2010; Schuller et al., 2010). The task is to classify a speaker's age and gender class which is defined as follows: children <13 years (C), young people 14–19 years (YF/YM), adults 20–54 years (AF/AM), and seniors >55 years (SF/SM), where F and M indicate female and male, respectively. We employed a Czech phoneme recognizer (Schwarz et al., 2006) to perform the voice activity detection (VAD) by simply dropping all frames that are decoded as silence or speaker noises. The mean and standard deviation of speech duration per data sample after VAD in the training and development data sets of the aGender database are 1.13 ± 0.86 s and 1.14 ± 0.87 s, respectively. Thus it is a short length speech utterance database. The training data set of the aGender database (472 speakers, 32,527 utterances) was used for model training while the development data set from the aGender database (300 speakers, 20,549 utterances) was used as the evaluation set of each subsystem as well as the fusion system in this paper. Finally, the testing data set from the aGender database (17,332 utterances) was evaluated. The details about the aGender database and the evaluation methods are provided in Schuller et al. (2010).

3. Methods

The overview of the proposed approach is shown in Fig. 1. In this section, we first introduce our three baseline subsystems (Sections 3.1–3.3) and then present the details of our proposed four new subsystems (Sections 3.4–3.7). Finally, the description of score level fusion is provided (Section 3.8).

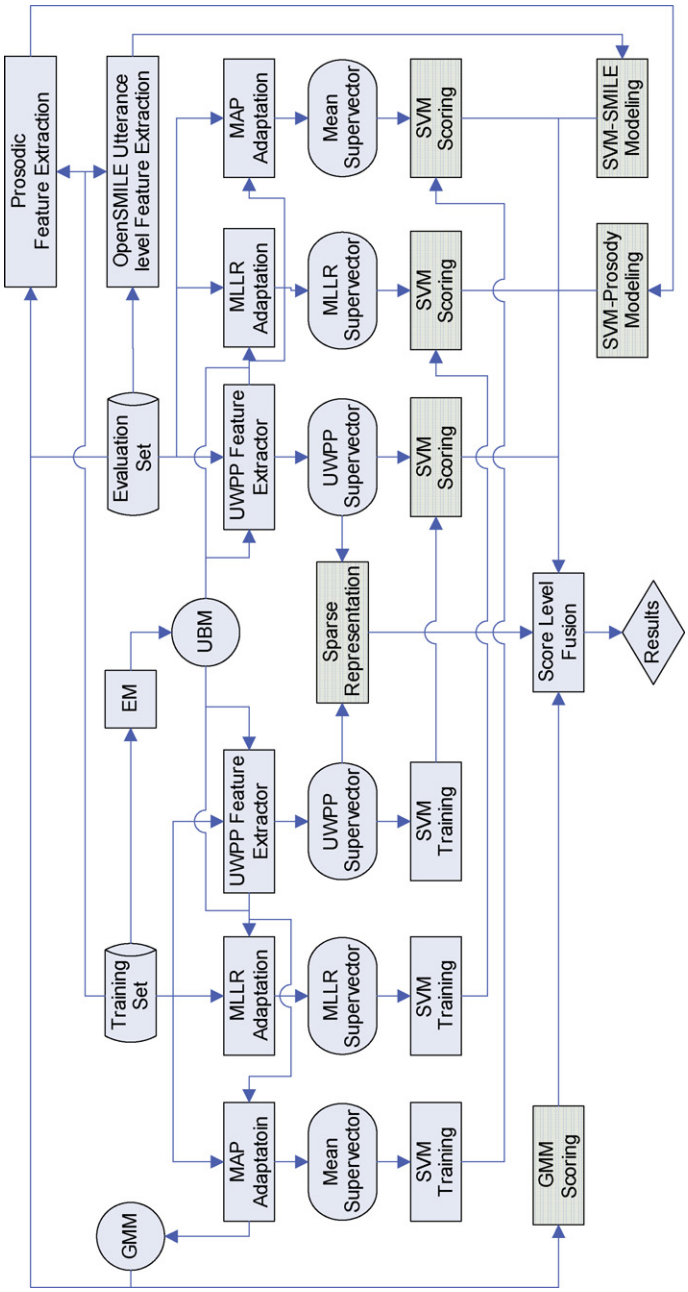

Fig. 1. System overview.

3.1. GMM baseline system

The features in this system are 13-dimensional MFCCs (including C0) and their first and second order derivatives, which result in 39-dimensional coefficients per frame. The window size and window shift for each frame is 20 ms and 10 ms, respectively. After voice activity detection, non-speech frames were eliminated and the 39-dimensional MFCC features were extracted. Cepstral mean subtraction and variance normalization were performed to normalize the MFCC features to zero mean and unit variance on a per-utterance basis. In the proposed work, a UBM in conjunction with a MAP model adaptation approach (Reynolds et al., 2000) was utilized to model different age and gender classes in a supervised manner. All the data in the training set were used to train a M-component UBM, and MAP adaptation was performed using the training set data for each age and gender class. The GMMs were modeled with diagonal covariance matrices and only the means of the GMMs were adapted with a relevance factor of 12. The fast scoring procedure devised by Reynolds et al. (2000) with top 4 Gaussian components was applied on the GMM scoring stage.

3.2. GMM–Mean–SVM baseline system

The feature extraction and UBM training were done in the GMM baseline system (Section 3.1). The means of Gaussian components were adapted by MAP adaptation for each training set and evaluation set utterance. Then the corresponding GMM mean supervectors, created by concatenating the mean vectors of all the Gaussian components, were modeled by SVM. The supervector was normalized by the corresponding standard deviation and weight to fit the supervector kernel (Campbell et al., 2006b). We arbitrarily added one dummy dimension with value 1 at the head of each mean supervector so that all the support vectors can be collapsed down into a single model vector and each target score can be calculated by a simple inner product which makes this framework computationally efficient (Campbell et al., 2006a). The dimension of the mean supervector from a 512 components GMM is $1 + (39 \times 512) = 19,969$. In addition, there are more than 30,000 utterances in the training set which makes the SVM training data set too large to be handled efficiently. Instead of directly training a multi-class SVM classifier using all the high dimensional supervectors, we adopted a two stage framework (Li et al., 2007) which can solve the practical limitation of computer memory demanded by large database training. First, the training data set of the aGender database (471 speakers) was divided into 2 parts: data from the last 20 speakers in alphabetical order of each age and gender class was used for back end SVM training (140 speakers) and the rest of the data (331 speakers) was used for front end SVM training. Then, based on the supervector samples from the front end SVM training set, multiple binary age and gender group based discriminative classifiers (in our case, 21 1vs1 classifiers for 7 classes) were trained and employed to map the mean supervectors into SVM score vectors (Li et al., 2007). SVMTorch (Collobert and Bengio, 2001) with linear kernel was used here as the front end modeling method due to its efficiency and capability to handle large scale training. Since the scoring function for each binary model is just an inner product (Campbell et al., 2006a), the mapping from supervectors to score vectors is computationally efficient. Furthermore, a back end SVM classifier was trained using LIBSVM (Chang and Lin, 2001) to model the probability distribution of each target age and gender class in the score vector space using the back end SVM training set supervector samples. The LIBSVM toolkit was employed in the second stage due to its capability of generating the decision scores in the posterior probability format which is easier to fuse with other subsystems.

3.3. SVM–SMILE baseline system

The SVM baseline system (Schuller et al., 2010) was provided by the 2010 Paralinguistic Challenge, which is based on 450-dimensional acoustic and prosodic features per utterance. We employed the LIBSVM toolkit to model the features provided by the challenge committee. The details of feature extraction and SVM modeling are presented in (Eyben et al., 2009; Schuller et al., 2010). Since various kinds of features, such as MFCC, line spectral pairs frequency, voicing probability, F0, F0 envelop, jitter, and shimmer, etc., are included, this system can capture age and gender information at both acoustic and prosodic levels. Moreover, combining this utterance level feature based system with our frame level MFCC feature based systems can potentially further improve the performance.

3.4. GMM–UWPP–SVM system

For each utterance in the training and evaluation sets, UWPP feature extraction is performed using the UBM. Given a frame-based MFCC feature $\boldsymbol{x_t}$ and the GMM–UBM λ with M Gaussian components (each component is defined as λ_i),

$$\lambda_i = \{w_i, \boldsymbol{\mu_i}, \boldsymbol{\Sigma_i}\}, \quad i = 1, \ldots, M, \tag{1}$$

the occupancy posterior probability is calculated as follows:

$$P(\lambda_i|\boldsymbol{x_t}) = \frac{w_i p_i(\boldsymbol{x_t}|\boldsymbol{\mu_i}, \boldsymbol{\Sigma_i})}{\Sigma_{j=1}^{M} w_j p_j(\boldsymbol{x_t}|\boldsymbol{\mu_j}, \boldsymbol{\Sigma_j})}. \tag{2}$$

This posterior probability can also be considered as the fraction of this feature $\boldsymbol{x_t}$ coming from the ith Gaussian component which is also denoted as partial counts. The larger the posterior probability, the better this Gaussian component can be used to represent this feature vector. The UWPP supervector is defined as follows:

$$UWPP_{supervector} = \boldsymbol{b} = [b_1, b_2, \ldots, b_M] \tag{3}$$

$$b_i = \frac{y_i}{T} = \frac{1}{T}\Sigma_{t=1}^{T} P(\lambda_i|\boldsymbol{x_t}). \tag{4}$$

Eq. (4) is exactly the same as the weight updating equation in the expectation-maximization (EM) algorithm in GMM training. The mixing coefficient b_i is equal to the fraction of data points assigned to the corresponding ith GMM component. Considering the GMM as a generative model which generates all the T independent frames of feature vectors, $y_i = \Sigma_{t=1}^{T} P(\lambda_i|\boldsymbol{x_t})$ is the total frames being drawn from the corresponding ith GMM component.

In order to model the UWPP supervector using SVM, we extend the kernel from the traditional linear kernel (Zhang et al., 2009) to the Bhattacharyya probability product (BPP) kernel (Jebara et al., 2004).

$$k_{linear}(P, P') = (\boldsymbol{b})^t \boldsymbol{b'} \tag{5}$$

$$k_{BPP}(P, P') = (\sqrt{\boldsymbol{b}})^t \sqrt{\boldsymbol{b}'} \tag{6}$$

Let the square root of UWPP supervector $\boldsymbol{\theta} = [b_1^{1/2}, b_2^{1/2}, \ldots, b_M^{1/2}]$ and $\boldsymbol{\theta'} = [b_1'^{1/2}, b_2'^{1/2}, \ldots, b_M'^{1/2}]$ be two input feature vectors to the SVM, then the BPP kernel is just the standard linear kernel on the square root UWPP supervectors:

$$k_{BPP}(P, P') = (\boldsymbol{\theta})^t \boldsymbol{\theta'}. \tag{7}$$

A multi-class SVM classifier based on the BPP kernel was employed for UWPP supervector modeling using all the training set data. LIBSVM (Chang and Lin, 2001) with probabilistic output training was adopted.

3.5. GMM–UWPP–Sparse representation system

Given N_j training samples $\boldsymbol{A_j}$ from the jth class, we construct the over-complete dictionary $\boldsymbol{A}$ using all the N training samples from K classes ($\Sigma_{j=1}^{K} N_j = N$):

$$\boldsymbol{A} = [\boldsymbol{A_1 A_2}, \ldots, \boldsymbol{A_K}] \tag{8}$$

$$= [\boldsymbol{s_{11}}, \boldsymbol{s_{12}}, \ldots, \boldsymbol{s_{1N_1}}, \boldsymbol{s_{21}}, \boldsymbol{s_{22}}, \ldots, \boldsymbol{s_{2N_2}}, \ldots, \boldsymbol{s_{K1}}, \boldsymbol{s_{K2}}, \ldots, \boldsymbol{s_{KN_K}}]. \tag{9}$$

Each sample $\boldsymbol{s_{ij}}$ is an M dimensional square root UWPP supervector $\boldsymbol{\theta}$ and automatically satisfies the unit l^2 norm property. Unit l^2 normalization on the original form of UWPP supervectors may violate the fundamental constraint of $\Sigma_{i=1}^{M} b_i = 1$.

$$\|\boldsymbol{\theta}\|_2 = \|[b_1^{1/2}, b_2^{1/2}, \ldots, b_M^{1/2}]\|_2 = \Sigma_{i=1}^{M} b_i = 1 \tag{10}$$

If no online training strategy is adopted, the over-complete dictionary is fixed throughout the entire testing progress. In order to achieve the sparse representation on an under-determined problem, $N_j \ll N, \forall j$ and $M < N$ need to be satisfied. In this work, we model the age and gender information together as a seven class ($K = 7$) joint recognition problem. Thus

$N_j \approx N/7$ and $N_j \ll N$, $\forall j$ is satisfied. Compared to the GMM mean supervectors, the UWPP supervector has notably smaller dimensionality M which not only strengthens the adoption of the sparse representation but also reduces the memory usage dramatically.

For any test square root UWPP supervector sample $\boldsymbol{y} \in \mathbb{R}^M$ (unit l^2 norm automatically holds), we want to use the over-complete dictionary $\boldsymbol{A}$ to linearly represent $\boldsymbol{y}$ (i.e., $\boldsymbol{y} = \boldsymbol{Ax}$) in a sparse way. If $\boldsymbol{y}$ is from the jth class, then $\boldsymbol{y}$ will approximately lie in the linear span of training samples in $\boldsymbol{A_j}$ (Wright et al., 2008). In order to be more robust against small, possibly, dense noise, we constrain the distance measure between the test sample $\boldsymbol{y}$ and the linear combination of training samples to be smaller than ϵ (Wright et al., 2008) which results in a standard convex optimization problem (l^1-minimization with quadratic constraints):

$$\text{Problem A}: \quad min\|\boldsymbol{x}\|_1 \quad \text{subject to} \quad \|\boldsymbol{Ax} - \boldsymbol{y}\|_2 \leq \epsilon \tag{11}$$

If we replace the square root UWPP supervectors $\boldsymbol{s_{ij}}$ and $\boldsymbol{y}$ into the original UWPP supervector form, the Euclidean distance in Eq. (11) is exactly the Hellinger's distance.

For each class j ($j = 1, \ldots, K$), let $\delta_j : \mathbb{R}^N \to \mathbb{R}^N$ be the characteristic function which selects the coefficients only associated with the jth class (Wright et al., 2008). For $\boldsymbol{x} \in \mathbb{R}^N$, $\delta_j(\boldsymbol{x}) \in \mathbb{R}^N$ is a new vector whose nonzero entries are the only entries in the N_j elements of $\boldsymbol{x}$ that corresponds to $\boldsymbol{A_j}$. Now based on the sparse representation $\boldsymbol{x}$, we proposed two decision criteria for verification purpose, namely l^1 norm ratio and l^2 residual ratio. A larger score represents a higher likelihood for the testing sample to be from the associated class.

$$l^1\text{norm ratio}: \quad \alpha_j = \|\delta_j(\boldsymbol{x})\|_1 / \|\boldsymbol{x}\|_1 \tag{12}$$

$$l^2\text{residual ratio}: \quad \beta_j = \frac{\sum_{i=1, i \neq j}^{K} \|\boldsymbol{y} - \boldsymbol{A}\delta_i(\boldsymbol{x})\|_2}{\|\boldsymbol{y} - \boldsymbol{A}\delta_j(\boldsymbol{x})\|_2} \tag{13}$$

Due to large variabilities, the test sample $\boldsymbol{y}$ could be partially corrupted. Thus an error vector $\boldsymbol{e}$ was introduced to explain the variability (Wright et al., 2008):

$$\boldsymbol{y} = \boldsymbol{y_0} + \boldsymbol{e} = \boldsymbol{Ax_0} + \boldsymbol{e} \tag{14}$$

So the original optimization problem becomes the following form:

$$\text{Problem B}: \quad min\|z\|_1 \quad \text{subject to} \quad \|\boldsymbol{Bz} - \boldsymbol{y}\|_2 \leq \epsilon \tag{15}$$

$$\boldsymbol{B} = [\boldsymbol{A} \quad \boldsymbol{I}] \in \mathbb{R}^{M \times (N+M)}, \qquad \boldsymbol{z} = [\boldsymbol{x}^t \quad \boldsymbol{e}^t]^t \in \mathbb{R}^{(N+M)} \tag{16}$$

If the error vector $\boldsymbol{e}$ is sparse and has no more than $M/2$ nonzero entries, the new sparse solution $\boldsymbol{z}$ is the true generator according to (14) (Wright et al., 2008). Finally, we redefine the two decision criterions based on the new sparse solution $\hat{\boldsymbol{z}} = [\hat{\boldsymbol{x}}^t \quad \hat{\boldsymbol{e}}^t]^t$.

$$l^1\text{norm ratio}: \quad \alpha_j = \|\delta_j(\hat{\boldsymbol{x}})\|_1 / \|\hat{\boldsymbol{x}}\|_1 \tag{17}$$

$$l^2\text{residual ratio}: \beta_j = \frac{\sum_{i=1, i \neq j}^{K} \|\boldsymbol{y} - \hat{\boldsymbol{e}} - \boldsymbol{A}\delta_i(\hat{\boldsymbol{x}})\|_2}{\|\boldsymbol{y} - \hat{\boldsymbol{e}} - \boldsymbol{A}\delta_j(\hat{\boldsymbol{x}})\|_2} \tag{18}$$

Fig. 2 demonstrates the sparse solution and ratio scores ((17) and (18)) of one random test utterance from the development set under the problem B (15) scenario. First, we can observe from the top sub-figure that the solution is sparse and the majority of non-zero coefficients of vector X (the first N dimension of Z) are associated with the correct dictionary index ($1 - 4407$ for class 1). Second, both the L1 norm ratio and L2 residual ratio show that this random sample belongs to class 1 clearly which demonstrates the effectiveness of the proposed sparse representation subsystem.

3.6. GMM–MLLR–SVM system

For each utterance in the training set and the evaluation set, a MLLR adaptation was performed to map the speaker-independent UBM model to a speaker dependent model (Stolcke et al., 2005, 2007). Statistics on the UBM model were gathered from the available adaptation data and used to calculate a linear regression based transformation for the UBM mean vectors (Leggetter and Woodland, 1995). For each utterance, an affine transformation matrix was calculated to maximize the likelihood of the corresponding feature vector. The columns of the corresponding MLLR matrices

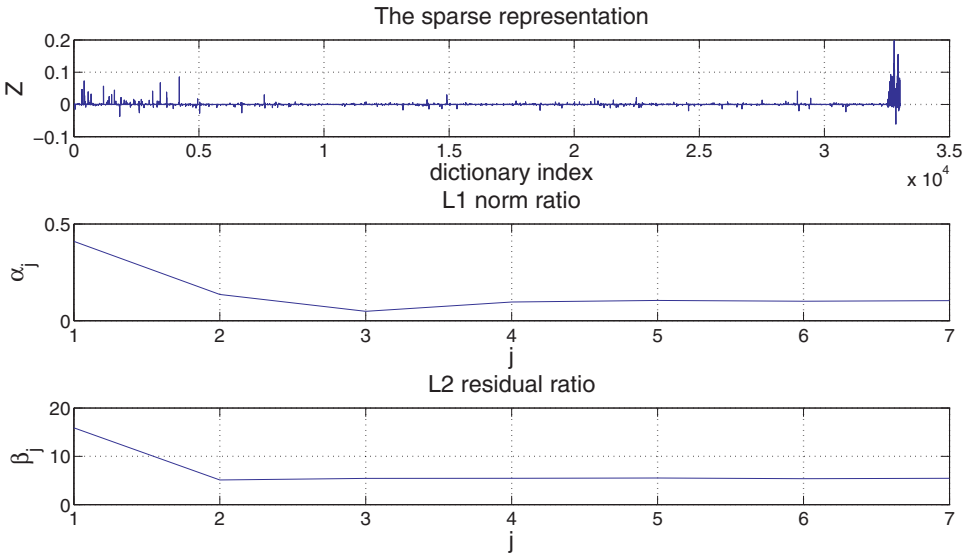

Fig. 2. Illustration of the sparse solution and ratio scores of utterance 1383_1_a11383s18 (class 1) using problem B (15) setting.

were stacked to form the MLLR matrix supervector and used for SVM modeling. Since the dimension of the MLLR matrix supervector is $39 \times 40 = 1560$ which is considerably smaller than the dimension of GMM mean supervector, we used all the supervectors from the training set to train a multi-class SVM classifier and performed scoring on the evaluation set. In order to increase the system robustness, Linear Discriminant Analysis (LDA) was employed to perform dimension reduction on the MLLR supervector space. Since there are K ($K = 7$) classes in our joint age and gender classification task, the dimensionality of the MLLR supervector after LDA is $K - 1$. Finally, a linear kernel multi-class SVM classifier was trained by LIBSVM (Chang and Lin, 2001) with probabilistic outputs.

3.7. SVM–Prosody system

It has been shown in Metze et al. (2007) and Bocklet et al. (2010) that prosodic features have better robustness to the variation of the utterance duration and the score level fusion has slightly superior performance than the feature level fusion. Thus, in this work, we used contour-based low-level descriptors to extract the prosodic information and then performed the multi-class SVM modeling directly at the syllable level. Finally, the acoustic and prosodic information were fused at the score level.

Given a speech utterance after the voice activity detection (VAD), we first performed sub-harmonic summation based pitch extraction (Li et al., 2008; Cao et al., 2007), formant extraction by WaveSurfer toolkit (Sjölander and Beskow, 2000), time domain energy calculation and frequency domain harmonic energy computation (Hermes, 1988; Cao et al., 2007) at a 32 ms window with 10 ms shift frame basis. The energy E_n of the nth harmonic frequency of the F0 is defined as follows:

$$E_n = \frac{1}{2\rho f_0} \int_{nf_0-\rho f_0}^{nf_0+\rho f_0} P(f)df \tag{19}$$

where $P(f)$ is the Short-Time Fourier Transform (STFT) power spectrum of an arbitrary frame, f_0 is the fundamental frequency and ρ is a spectral normalized width factor with 0.05 value.

To enhance the robustness, logarithm of the feature values was applied and log energy was normalized on an utterance basis by subtracting the maximum value of the entire utterance (Dehak et al., 2007a). The continuous long voiced segments were located using pitch values and unvoiced frames were discarded. Then, within each voiced speech segment, we segment the long prosodic contours into syllable-like regions in the same way as Dehak et al. (2007a) by detecting the valley points of the log energy contour. Fig. 3 shows an example of syllable-like region segmentation using pitch and log energy contour information. Short syllable units with less than 60 ms duration were discarded. In Fig. 3, we can see that in total there are 9 valid syllable-like units (S1–S9) in 5 continuous voiced segments. Generally,

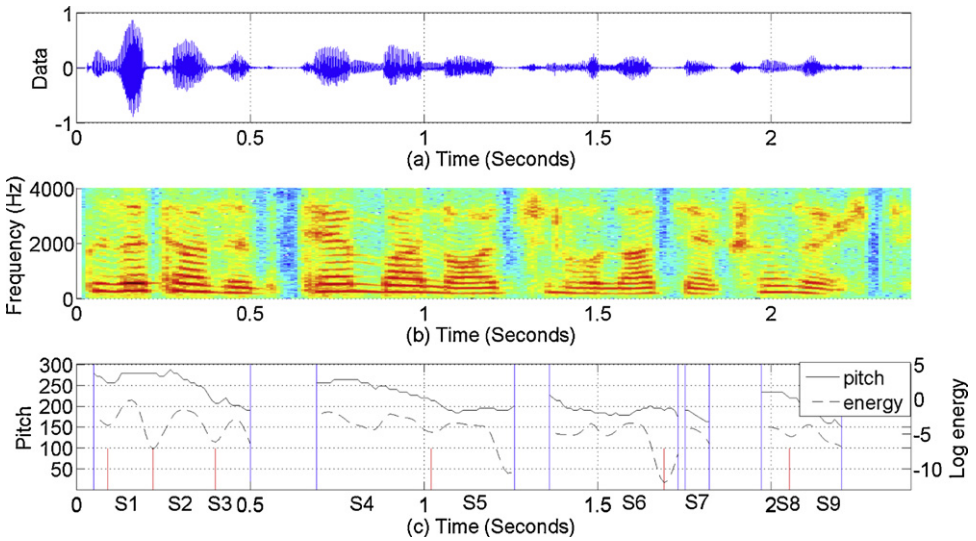

Fig. 3. Syllable-like region segmentation using pitch and log energy contour information (Dehak et al., 2007a). (a) Waveform, (b) spectrogram (c) pitch and log energy contour within each voiced speech segment. The blue segment lines were the start and end time for each continuous voiced speech segment. For each continuous voiced speech segment, the red vertical lines marked the valley points of log energy contour which also served as the boundaries of adjacent syllable units. Short syllable units with less than 60 ms duration were discarded. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

Table 1
The order and dimension for prosodic features.

Prosodic feature type	Order	Dimension
Log pitch	1	6
Log energy	1	6
Log harmonic energy	10	60
Log formant	4	24
Unit duration	1	1
Total		97

these valley points are the boundary of each syllable-like unit. We limit the minimum length of the unit to be 6 frames due to the constraint of six term Legendre polynomial expansions. Tables 1 and 2 summarize the prosodic feature extraction parameters.

Furthermore, for each syllable-like unit, the contours for pitch, time domain energy, frequency domain harmonic structure energy and formants were normalized into $[-1, 1]$ time scale and approximated by a six term Legendre polynomial expansion (Lin and Wang, 2005; Dehak et al., 2007a). As shown by Fig. 4, the reconstructed contours fit the original contours very well on the normalized $[-1, 1]$ time scale. Thus, the expansion coefficients can capture the mean, slope, curvature and many other detailed information about the contour. As shown in Table 1, in total we have 97-dimensional prosodic feature vectors for each syllable-like unit in voiced speech segments. A multi-class SVM

Table 2
Parameters for pitch and formant extraction.

Frame size	32 ms
Frame shift	10 ms
Lowest F_0	50 Hz
Highest F_0	350 Hz
Saliency threshold	0.5
Preemphasis factor	0.97
FFT length	512

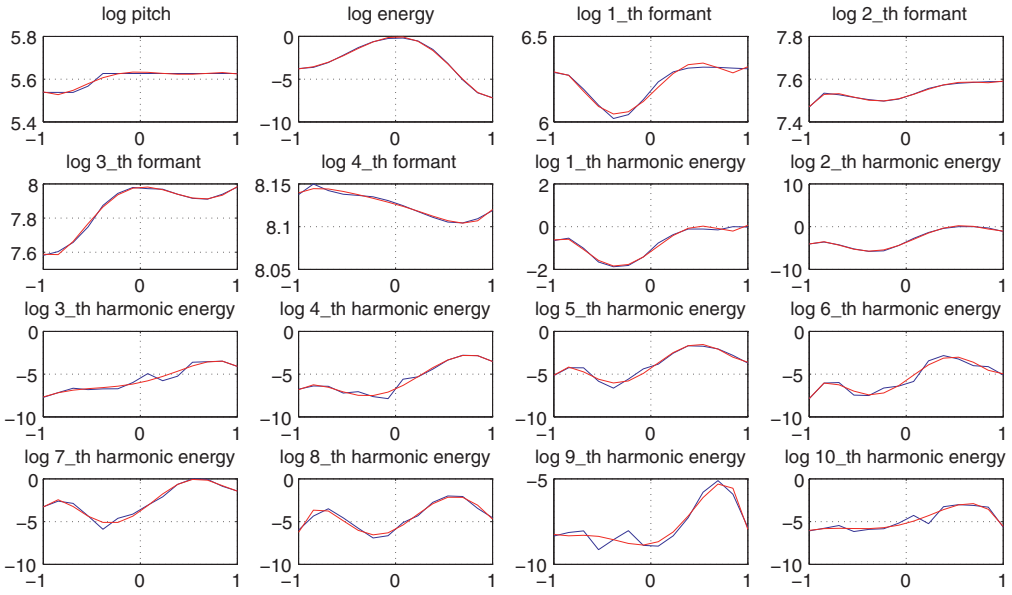

Fig. 4. The original and the polynomial expansion reconstructed contours of S1 segment in Fig. 3. The blue curves are the original contours and the red curves are the reconstructed contours based on the Legendre polynomial expansion coefficients. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

with linear kernel was employed as the classifier and feature vectors from all the syllable-like units were used for training. Suppose we have N training utterances and we have C_n syllable-like units for the *nth* utterance, there are totally $\sum_{n=1}^{N} C_n$ samples with dimensionality of 97. We normalized each dimension of the feature vector to be within $[-1, 1]$, and then performed multi-class SVM training using LIBSVM with linear kernel and probability modeling function enabled. In the testing phase, suppose we have C_{test} syllable-like units for an arbitrary testing utterance. For every syllable like unit, we performed the SVM scoring independently. Finally, the average score vector along all those C_{test} syllable-like units was computed as the final subsystem result.

3.8. Score level fusion

Due to the limited amount of training data, we simply employed the weighted summation fusion approach with parameters tuned by cross validation. Let there be F input subsystems (as shown in Fig. 1 and Table 6, $F = 7$ in this work) where the ith subsystem outputs its own posterior probability vector $l_i(\boldsymbol{x})$ for every trial. Then the fused score vector $\acute{l}(\boldsymbol{x})$ is given by:

$$\acute{l}(\boldsymbol{x}) = \sum_{i=1}^{F} \eta_i l_i(\boldsymbol{x}) \tag{20}$$

The weight, η_k, can be tuned by validation data. For the GMM baseline subsystem, log-likelihood normalization was adopted to map the log likelihood scores into the posterior probabilities. Within all the K target classes ($K = 7$ in our joint age-gender classification task), suppose the GMM log-likelihood score of an arbitrary testing utterance $\boldsymbol{x}$ on the kth target class is $s_k(\boldsymbol{x})$, then the approximated posterior probability score under flat prior assumption is defined as follows:

$$l(\boldsymbol{x}) = \frac{e^{s_k(\boldsymbol{x})}}{\sum_{k=1}^{K} e^{s_k(\boldsymbol{x})}} \tag{21}$$

When the evaluation was performed on the testing set of the aGender database, both the training and development sets were used for modeling and the weight vector η_i, $(i = 1, \ldots, F)$ was exactly the same as the one tuned on the development set. It is worth noting that other advanced score fusion approaches, like the logistic regression method in

Table 3
Performance of subsystems based on different GMM sizes evaluated on the development set.

ID.System	GMM size task									
	128		256		512					
	Age and gender		Age and gender		Age and gender		Age		Gender	
	UA	WA	UA	WA	UA	WA	UA	WA	UA	WA
1.GMM	42.8	42.2	43.5	43.1	**45.8**	**45.3**	47.5	49.3	78.0	83.7
2.GMM–Mean–SVM	42.0	42.6	42.2	42.8	**42.6**	**43.2**	46.1	45.6	75.7	82.5
4.GMM–UWPP–SVM (linear kernel)	38.3	39.1	40.1	40.7	**41.5**	**42.2**	45.0	44.9	74.5	82.1
6.GMM–MLLR–SVM	38.8	39.3	39.7	40.1	**39.8**	**40.2**	44.1	43.7	72.5	79.4

Bold font is to highlight the best configuration for each subsystem.

the popular FoCal toolkit (Brümmer, 2007), can be adopted here to increase the performance which is a topic for our future work.

4. Experimental results

The development set was used to evaluate the performance of each subsystem as well as of the fusion approach. Finally, performance on the testing set is also reported. Both unweighted accuracy (UA) and weighted accuracy (WA) on average per class (UA/WA, weighting with respect to number of instances per class) for each of the 3 different classification tasks (7 class age and gender {C,YF,YM,AF,AM,SF,SM}, 4 class age {C,Y,A,S} and 3 class gender {C,F,M}) are presented. The details of these 3 tasks as well as the evaluation method are provided in Section 2 and Schuller et al. (2010).

4.1. GMM based subsystems

Table 3 shows the results of GMM based subsystems evaluated against the GMM size. It can be observed from Table 3 that the bigger the GMM size, the better the performance. Other than the GMM and GMM–UWPP–SVM (linear kernel) subsystems that achieved significant improvement with 512 GMMs compared to 128 GMMs (the Z-test p values of the null hypothesis that systems based on these 2 GMM sizes equally performed in the WA accuracy for the age and gender task are <0.0001), the GMM–Mean–SVM and GMM–MLLR–SVM subsystems only had moderate accuracy gain. In addition, all the 3 GMM supervector based subsystems did not achieve superior performance compared to the traditional GMM baseline. This might be because the utterance duration is too short to perform good quality GMM adaptations on the UBM. In the training stage, the GMM baseline used all the training data for each class which is sufficient for the MAP adaptation. However, in the GMM supervector based SVM systems, the amount of data for generating each supervector sample is just one short utterance. Continuing to increase the GMM size will make the dimension of GMM supervectors too high for efficiently modeling using SVM. Thus, the rest of this section will focus on the evaluation results on the 512 component GMMs.

4.2. SVM and sparse representation on the GMM UWPP supervectors

In Table 4, the results of different setups for the SVM and sparse representation subsystems on the GMM UWPP supervectors are shown. First, we can see that BPP kernel outperforms linear kernel in the SVM modeling on the UWPP supervectors. Furthermore, L2 residual ratio based criteria has better performance compared to the L1 norm ratio for the sparse representation modeling. And Hellinger's distance based quadratic constraint yields significant boost in the sparse representation subsystem compared to the baseline Euclidean distance. This might be because the square root UWPP supervectors automatically satisfy the unit l^2 norm and the Hellinger's distance is closely related to the Bhattacharyya's affinity between distributions (Jebara et al., 2004). Finally, we can observe from Table 4 that the sparse representation modeling with L2 ratio on square root UWPP supervectors yields competitive results compared

Table 4
Performance of GMM UWPP supervector modeling evaluated on the development set with 512 GMM.

Methods	Kernel or ratio	Supervector	Age and gender		Age		Gender	
			UA	WA	UA	WA	UA	WA
SVM	linear	UWPP	41.5	42.2	45.0	44.9	74.5	82.1
SVM	BPP[a]	UWPP	**42.4**	**42.9**	45.8	45.7	74.6	82.2
Sparse representation	L1	UWPP	36.9	38.2	41.7	42.3	72.1	79.8
Sparse representation	L2	UWPP	38.5	38.7	43.1	42.2	73.7	80.4
Sparse representation	L1	Square root UWPP	38.3	39.7	42.8	43.5	72.8	80.9
Sparse representation	L2	Square root UWPP	**41.4**	**41.4**	45.3	44.4	75.3	82.4

Bold font is to highlight the best configuration for each subsystem.
[a] BPP kernel denotes the Bhattacharyya probability product kernel.

to the other GMM–SVM supervector subsystems. Considering there is no multi-class training work required for sparse representation, this approach can serve as a kind of online adaptive learning method.

4.3. SVM–Prosody subsystem

Performance of the SVM–Prosody subsystem based on different configurations of prosodic features is reported in Table 5. It is shown that the pitch feature by itself achieved 34.6% unweighted accuracy (UA) for the age and gender joint modeling while adding duration and time domain energy information only improved the performance to 35.1% UA. Therefore, the pitch information plays a dominant role in the prosodic feature sets. Furthermore, adding frequency domain harmonic energy and formant contour information yields significant improvement from 35.1% to 39.3%. This validates the effectiveness of the proposed low level prosodic descriptors. Thus, in our SVM–Prosody subsystem for the age and gender recognition task, pitch, spectral harmonic energy and formant seem to be the most effective and important prosodic features. We can observe from Table 5 that extending traditional prosodic contours to spectral domain feature contours yields high gains in performance which matches the situation in speaker verification domain (Dehak et al., 2007b; Kockmann and Burget, 2008).

4.4. Score level fusion

It is shown in Table 6 that the GMM (system 1) and SVM–SMILE (system 3) baseline subsystems outperformed the other subsystems in terms of both UA and WA accuracy. However, by combining all the 7 different methods together, significant improvements (Z-test $p < 0.00001$) were achieved for all the 3 classification tasks. Now let us focus on the unweighted accuracy for the age and gender 7 classes joint recognition task. Firstly, fusing systems 4 and 5 achieved a 1.4% improvement against system 4 alone, which is interesting because fusing the two UWPP supervector based subsystems surpasses the GMM mean supervector system that solely achieves the best performance among the three supervector based subsystems using SVM (system 2, 4, and 6). Secondly, fusing all the GMM supervector based subsystems together into system 10 achieved 46.4% UA accuracy which outperformed the GMM and SVM–SMILE baseline subsystems. This confirms that these GMM supervectors are complementary with one another. Thirdly, fusing

Table 5
Performance of SVM–Prosody subsystem evaluated on the development set.

Pitch	Duration	Energy	Harmonic energy	Formant	Age and gender		Age		Gender	
					UA	WA	UA	WA	UA	WA
√					34.6	39.1	37.4	42.1	71.9	84.5
√	√				34.9	39.1	39.2	42.1	74.0	82.9
√	√	√			35.1	39.2	39.3	42.2	74.1	83.0
√	√	√	√		38.3	40.0	42.0	42.8	73.7	81.5
√	√	√	√	√	**39.3**	**40.7**	43.1	43.5	74.7	82.1

Bold font is to highlight the best configuration for each subsystem.

Table 6
Performance of each subsystem and score level fused systems evaluated on the development set with 512 GMMs.

ID	System	Age and gender		Age		Gender	
		UA	WA	UA	WA	UA	WA
1	GMM	45.8	45.3	47.5	49.3	78.0	83.7
2	GMM–Mean–SVM	42.6	43.2	46.1	45.6	75.7	82.5
3	SVM–SMILE	44.5	44.9	47.2	46.6	77.5	85.1
4	GMM–UWPP–SVM	42.4	42.9	45.8	45.7	74.6	82.2
5	GMM–UWPP–Sparse representation	41.4	41.4	45.3	44.4	75.3	82.4
6	GMM–MLLR–SVM	39.8	40.2	44.1	43.7	72.5	79.4
7	SVM–Prosody	39.3	40.7	43.1	43.5	74.7	82.1
8	Weighted sum fuse 4 + 5	43.8	44.2	47.2	46.9	76.2	83.5
9	Weighted sum fuse 4 + 5 + 6	44.7	45.3	48.2	47.8	77.0	84.2
10	Weighted sum fuse 2 + 4 + 5 + 6	46.4	47.1	49.8	49.4	78.5	85.3
11	Weighted sum fuse 1 + 2 + 4 + 5 + 6	47.4	47.8	50.8	50.0	79.0	85.3
12	Weighted sum fuse 1 + 2 + 4 + 5 + 6 + 7	49.0	49.8	51.9	51.1	81.1	87.4
13	Weighted sum fuse 1 + 2 + 3 + 4 + 5 + 6 + 7	**50.3**	**51.1**	**52.8**	**52.2**	**81.7**	**88.2**
	Kockmann et al. (2010) [a]	53.9	54.2	56.0	55.3	81.6	87.1
	Meinedo and Trancoso (2010) [b]			51.2	50.6	83.1	86.9

Bold font is to highlight the best configuration for each subsystem.
[a] The age sub-challenge winner of the INTERSPEECH 2010 Paralinguistic Challenge.
[b] The gender sub-challenge winner of the INTERSPEECH 2010 Paralinguistic Challenge.

Table 7
Confusion matrix for 7 class age and gender task on the development set for system 13.

	C	YF	YM	AF	AM	SF	SM
C	**61.0**	16.9	7.5	4.9	2.0	6.7	1.0
YF	16.4	**57.1**	0.8	15.8	0.3	9.0	0.6
YM	0.3	0.8	**49.4**	1.0	21.9	3.2	23.5
AF	5.5	26.6	1.8	**33.8**	0.4	31.3	0.6
AM	0.1	0.0	29.2	1.1	**27.1**	2.0	40.5
SF	7.1	11.4	1.5	22.9	0.9	**53.9**	2.2
SM	0.2	0.1	11.5	0.2	16.2	2.0	**69.7**

Bold font is to highlight the best configuration for each subsystem.

the GMM baseline with the previous system 10 further increased the UA accuracy to 47.4%. So far, all the information in the fusion are all coming from the acoustic level MFCC features with the GMM framework. Fourthly, we fused the acoustic level information with the prosodic level information together to achieve our final fusion system 13. It needs to be pointed out that this significant improvement (around 3% from system 11 to 13) is achieved by exploiting the syllable and utterance level prosodic information. Thus, score level fusion of the acoustic and prosodic information together enhanced the performance significantly. Finally, our fusion system 13 achieved 52.8% UA (52.2% WA) and 81.7% UA (88.2% WA) on the development set for the age recognition and gender recognition tasks, respectively.

The confusion matrix for the 7 class age and gender task on development set for system 13 is shown in Table 7. First, we can see that children, youth and senior groups perform better than adult group. This might be due to the relatively large age range (20–54 years) for the adult group. Second, the children class has bigger confusion with the female youth (YF) class than with the male youth (YM) class. This might be because children have relatively similar voices to female youths. Third, the main confusion comes from speakers with the same gender of other age groups. For example, Female Adult (AF) group has big confusion with Female Youth (YF) and Female Senior (SF) groups. This result is consistent with the big gap between age classification accuracy and gender classification accuracy in Table 6.

Table 8 demonstrates the comparison of final performances (system 13) on the official test set of the challenge. It is shown that the proposed system achieved competitive results compared to other participating systems. Finally, from the confusion matrix of system 13 shown in Tables 9 and 10, we can find out that the most difficult age group to recall is

Table 8
Comparison of the final performances evaluated on the official testing set.

Methods	Task			
	Age		Gender	
	UA	WA	UA	WA
Schuller et al. (2010)	48.9	46.2	81.2	84.8
Nguyen et al. (2010)	49.1		81.7	
Porat et al. (2010)	43.1	39.8		
Gajšek et al. (2010)			82.8	87.3
Kockmann et al. (2010) [a]	52.4	51.2	83.1	85.7
The proposed method	**52.0**	**49.5**	**85.0**	**88.4**

Bold font is to highlight the best configuration for each subsystem.
[a] The age sub-challenge winner of the INTERSPEECH 2010 Paralinguistic Challenge.

Table 9
Confusion matrix for age task on test set.

	C	Y	A	S
Children	**71.0**	15.8	5.5	7.8
Youths	7.3	**41.8**	26.2	24.7
Adults	2.2	19.1	**25.3**	53.4
Seniors	4.0	9.8	16.3	**70.0**

Bold font is to highlight the best configuration for each subsystem.

Table 10
Confusion matrix for gender task on test set.

	C	F	M
Children	**71.2**	25.2	3.7
Females	9.9	**88.5**	1.6
Males	0.8	3.9	**95.3**

Bold font is to highlight the best configuration for each subsystem.

still the adult group and the biggest confusion pair for gender recognition is child and female which match our findings on the development set.

5. Conclusions

In this work, we addressed the speaker age and gender recognition problem with acoustic and prosodic level information fusion. The contributions are as follows: (1) At the acoustic level, we introduced two additional GMM supervectors, namely MLLR and UWPP supervectors, as features for SVM modeling. (2) Sparse representation was introduced for GMM square root UWPP supervector modeling which is suitable for large scale online adaptive learning due to its property of no training effort required. (3) Contours of pitch, time domain energy, frequency domain harmonic structure energy and formant for each syllable unit in every voiced speech segment are mapped into polynomial expansion coefficients as our novel prosodic features and modeled directly at the syllable level using SVM. Experimental results showed that pitch, spectral harmonic energy and formant are the most effective and important prosodic features in our case. (4) The proposed four subsystems have been demonstrated to be effective and show competitive results in classifying different age and gender groups. (5) Score level fusion of all the subsystems was shown to improve the overall performance significantly. Future work includes investigating the GMM-SVM Constrained MLLR supervector method, combining other prosodic or phonetic level methods, and validating the results with a relatively larger and longer duration database. Moreover, rather than applying existing approaches from speaker verification and language

identification domains, investigation on novel features and algorithms specifically targeted for the detection of age and gender states are also very important.

References

Ajmera, J., Burkhardt, F., 2008. Age and gender classification using modulation cepstrum. In: Proc. Odyssey, p. 025.

Black, M., Katsamanis, A., Lee, C., Lammert, A., Baucom, B., Christensen, A., Georgiou, P., Narayanan, S., 2010. Automatic classification of married couples' behavior using audio features. In: Proc. INTERSPEECH, pp. 2030–2033.

Bocklet, T., Maier, A., Bauer, J., Burkhardt, F., Nöth, E., 2008. Age and gender recognition for telephone applications based on GMM supervectors and support vector machines. In: Proc. ICASSP, pp. 1605–1608.

Bocklet, T., Stemmer, G., Zeissler, V., Nöth, E., 2010. Age and gender recognition based on multiple systems – early vs. late fusion. In: Proc. INTERSPEECH, pp. 2830–2833.

Brümmer, N., 2007. Focal multi-class: toolkit for evaluation, fusion and calibration of multi-class recognition scorestutorial and user manual. Software available at http://sites.google.com/site/nikobrummer/focalmulticlass.

Burkhardt, F., Eckert, M., Johannsen, W., Stegmann, J., 2010. A database of age and gender annotated telephone speech. In: Proc. 7th International Conference on Language Resources and Evaluation (LREC), pp. 1562–1565.

Campbell, W., Campbell, J., Reynolds, D., Singer, E., Torres-Carrasquillo, P., 2006a. Support vector machines for speaker and language recognition. Computer Speech & Language 20, 210–229.

Campbell, W., Sturim, D., Reynolds, D., Solomonoff, A., 2006b. SVM based speaker verification using a GMM supervector kernel and NAP variability compensation. In: Proc. ICASSP, pp. 97–100.

Cao, C., Li, M., Liu, J., Yan, Y., 2007. Multiple f0 estimation in polyphonic music. In: Third Music Information Retrieval Evaluation eXchange (MIREX).

Chang, C.C., Lin, C.J., 2001. LIBSVM: A Library for Support Vector Machines. Software available at http://www.csie.ntu.edu.tw/cjlin/libsvm.

Collobert, R., Bengio, S., 2001. SVMTorch: support vector machines for large-scale regression problems. The Journal of Machine Learning Research 1, 143–160.

Dehak, N., Dumouchel, P., Kenny, P., 2007a. Modeling prosodic features with joint factor analysis for speaker verification. IEEE Transactions on Audio, Speech, and Language Processing 15, 2095–2103.

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

Dehak, N., Kenny, P., Dumouchel, P., 2007b. Continuous prosodic features and formant modeling with joint factor analysis for speaker verification. In: Proc. INTERSPEECH, pp. 1234–1237.

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

Eyben, F., Wollmer, M., Schuller, B., 2009. OpenEAR introducing the Munich open-source emotion and affect recognition toolkit. In: Affective Computing and Intelligent Interaction and Workshops, ACII, pp. 1–6.

Gajšek, R., Žibert, J., Justin, T., Štruc, V., Vesnicer, B., Mihelič, F., 2010. Gender and affect recognition based on GMM and GMM-UBM modeling with relevance MAP estimation. In: Proc. INTERSPEECH, pp. 2810–2813.

van Heerden, C., Barnard, E., Davel, M., van der Walt, C., van Dyk, E., Feld, M., Müller, C., 2010. Combining regression and classification methods for improving automatic speaker age recognition. In: Proc. ICASSP, pp. 5174–5177.

Hermes, D., 1988. Measurement of pitch by subharmonic summation. Journal of the Acoustical Society of America 83, 257–264.

Jebara, T., Kondor, R., Howard, A., 2004. Probability product kernels. The Journal of Machine Learning Research 5, 819–844.

Kockmann, M., Burget, L., 2008. Contour modeling of prosodic and acoustic features for speaker recognition. In: Proc. Spoken Language Technology Workshop, IEEE, pp. 45–48.

Kockmann, M., Burget, L., Černocký, J., 2010. Brno university of technology system for interspeech 2010 paralinguistic challenge. In: Proc. INTERSPEECH, pp. 2822–2825.

Lee, C., Black, M., Katsamanis, A., Lammert, A., Baucom, B., Christensen, A., Georgiou, P., Narayanan, S., 2010. Quantification of prosodic entrainment in affective spontaneous spoken interactions of married couples. In: Proc. INTERSPEECH, pp. 793–796.

Leggetter, C., Woodland, P., 1995. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. Computer Speech and Language 9, 171.

Li, M., Cao, C., Di Wang, P., Fu, Q., Yan, Y., 2008. Cochannel speech separation using multi-pitch estimation and model based voiced sequential grouping. In: Proc. INTERSPEECH, pp. 151–154.

Li, M., Jung, C.S., Han, K.J., 2010. Combining five acoustic level methods for automatic speaker age and gender recognition. In: Proc. INTERSPEECH, pp. 2826–2829.

Li, M., Narayanan, S., 2011. Robust talking face video verification using joint factor analysis and sparse representation on GMM mean shifted supervectors. In: Proc. ICASSP, pp. 1481–1484.

Li, M., Suo, H., Wu, X., Lu, P., Yan, Y., 2007. Spoken language identification using score vector modeling and support vector machine. In: Proc. INTERSPEECH, pp. 350–353.

Li, M., Zhang, X., Yan, Y., Narayanan, S., 2011. Speaker verification using sparse representations on total variability i-vectors. In: Proc. INTERSPEECH.

Lin, C., Wang, H., 2005. Language identification using pitch contour information. In: Proc. ICASSP, pp. 601–604.

Lingenfelser, F., Wagner, J., Vogt, T., Kim, J., André, E., 2010. Age and gender classification from speech using decision level fusion and ensemble based techniques. In: Proc. INTERSPEECH, pp. 2798–2801.

Meinedo, H., Trancoso, I., 2010. Age and gender classification using fusion of acoustic and prosodic features. In: Proc. INTERSPEECH, pp. 2818–2821.

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. ICASSP, pp. 1089–1092.

Müller, C., Burkhardt, F., 2007. Combining short-term cepstral and long-term pitch features for automatic recognition of speaker age. In: Proc. INTERSPEECH, pp. 2277–2280.

Nguyen, P., Le, T., Tran, D., Huang, X., Sharma, D., 2010. Fuzzy support vector machines for age and gender classification. In: Proc. INTERSPEECH, pp. 2806–2809.

Porat, R., Lange, D., Zigel, Y., 2010. Age recognition based on speech signals using weights supervector. In: Proc. INTERSPEECH, pp. 2814–2817.

Reynolds, D., Quatieri, T., Dunn, R., 2000. Speaker verification using adapted Gaussian mixture models. Digital Signal Processing 10, 19–41.

Schötz, S., 2007. Acoustic analysis of adult speaker age. Speaker classification I. Lecture Notes in Computer Science, 88–107.

Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Mueller, C., Narayanan, S., 2010. The INTERSPEECH 2010 paralinguistic challenge. In: Proc. INTERSPEECH, pp. 2794–2797.

Schwarz, P., Matejka, P., Cernocky, J., 2006. Hierarchical structures of neural networks for phoneme. In: Proc. ICASSP, pp. 325–328. Software available at http://speech.fit.vutbr.cz/software/phoneme-recognizer-based-long-temporal-context.

Sjölander, K., Beskow, J., 2000. Wavesurfer-an open source speech tool. In: Proc. ICSLP, pp. 464–467.

Spiegl, W., Stemmer, G., Lasarcyk, E., Kolhatkar, V., Cassidy, A., Potard, B., Shutn, S., Song, Y., Xu, P., Beyerlein, P., Harnsberger, J., Nöth, E., 2009. Analyzing features for automatic age estimation on cross-sectional data. In: Proc. INTERSPEECH, pp. 2923–2926.

Stolcke, A., Ferrer, L., Kajarekar, S., Shriberg, E., Venkataraman, A., 2005. MLLR transforms as features in speaker recognition. In: Proc. INTERSPEECH, pp. 2425–2428.

Stolcke, A., Kajarekar, S., Ferrer, L., Shrinberg, E., Int, S., Park, M., 2007. Speaker recognition with session variability normalization based on MLLR adaptation transforms. IEEE Transactions on Audio, Speech, and Language Processing 15, 1987–1998.

Wolters, M., Vipperla, R., Renals, S., 2009. Age recognition for spoken dialogue systems: Do we need it? In: Proc. INTERSPEECH, pp. 1435–1438.

Wright, J., Yang, A., Ganesh, A., Sastry, S., Ma, Y., 2008. Robust face recognition via sparse representation. IEEE Transactions on Pattern Analysis and Machine Intelligence 31, 210–227.

Zhang, X., Suo, H., Zhao, Q., Yan, Y., 2009. Using a kind of novel phonotactic information for SVM based speaker recognition. IEICE Transactions on Information and Systems 92, 746–749.