A new pitch range–based feature set for speaker age and gender classification

Article *in* Applied Acoustics · April 2015

DOI: 10.1016/j.apacoust.2015.04.013

2 authors:

Buket D. Barkana
University of Bridgeport
69 PUBLICATIONS **151** CITATIONS
SEE PROFILE

Jingcheng Zhou
University of Massachusetts Lowell
10 PUBLICATIONS **8** CITATIONS
SEE PROFILE

Applied Acoustics 98 (2015) 52–61

Contents lists available at ScienceDirect

Applied Acoustics

journal homepage: www.elsevier.com/locate/apacoust

A new pitch-range based feature set for a speaker's age and gender classification

Buket D. Barkana *, Jingcheng Zhou

Department of Electrical Engineering, University of Bridgeport, 221 University Ave., Bridgeport, CT, USA

ARTICLE INFO

Article history:
Received 13 May 2014
Received in revised form 20 April 2015
Accepted 21 April 2015

Keywords:
Age and gender classification
Pitch range
Fundamental frequency
MFCCs

ABSTRACT

This paper presents a pitch-range (PR) based feature set for age and gender classification. The performance of the proposed feature set is compared with MFCCs, energy, relative spectral transform–perceptual linear prediction (RASTA_PLP), and fundamental frequency (F0). Voice activity detection (VAD) is performed to extract speech utterances before feature extraction. Two different classifiers, k-Nearest Neighbors (kNN) and Support Vector Machines (SVM) are used in order to evaluate the effectiveness of the feature sets. Experimental results are reported for the aGender database. Both kNN and SVM classifiers achieved the highest accuracy rates by the proposed PR feature set in age + gender and age classifications. PR features represent the pitch changes over time. In age + gender classification, the class of middle-aged female speaker is recognized with an accuracy of 92.86%, followed by senior female speakers with 83.61%, children with 83.02%, middle-aged male speakers with 73.58%, young female speakers with 67.35%, and senior male speakers with 34.33% by using 3PR features with the SVM classifier. Low classification accuracies are observed for young male speakers.

1. Introduction

Detecting the age and gender of a speaker, given a short speech utterance, is a challenging task, and it is a rapidly emerging field of research because of the continually growing interest in applications of communication, human–computer interface and natural spoken-dialog systems. The understanding of the acoustics of an aging voice will provide a better understanding in a listener's perception of the aging voice. Human–computer interaction (HCI) systems such as a dialog system can be custom-designed, based on the speaker's voice portrait in order to guide the conversation and improve the level of customer satisfaction. The HMIHY system built by AT&T detects age and gender among various other voice signatures [1,2]. Gender detection is also a main part of the VIDIVIDEO European project, which deals with the semantic search of audio–visual documents [3].

There are many factors affecting the performance of such systems. The input speech material can be text-dependent or text-independent. In text-dependent systems, the speaker recites pre-determined words or phrases, whereas in text-independent systems there are no restrictions. Text-dependent systems give better performance than text-independent systems, because the speaker is aware of the task so that s/he is cooperative and consistent [4]. Another factor is the selection of the feature set to be used in the template representation. A desirable feature set should be easy to compute and robust. More importantly for text-independent systems, the feature set should also be unsusceptible to background noise since the speaker is often unaware of the task, and operating surroundings can have background noise and interference. The aim is to obtain a small and efficient set of acoustic features which represent the input pattern for the classification algorithms being trained [5].

This study focuses on the problem of extracting age and gender information from the speaker's voice. The speaker's gender is recognized in many automatic speech recognition (ASR) systems for the purpose of choosing gender-specific acoustic models. Gender classification has been the focus more than age classification in previous studies. Researchers have studied age identification on a small corpus [6,7]. Minematsu et al. stated that humans can identify age groups reliably even across different languages. One of these works in age detection performs a binary classification on a small corpus (elderly versus others) and achieves a 95% accuracy rate [8]. A method for detecting elderly speech based on prosodic features (jitter and shimmer) was proposed in [9].

Spectral and temporal feature sets, Mel-frequency cepstral coefficients (MFCCs), formant frequencies, fundamental frequency (F0),

* Corresponding author.
E-mail addresses: bbarkana@bridgeport.edu (B.D. Barkana), jinzhou@bridgeport.edu (J. Zhou).

http://dx.doi.org/10.1016/j.apacoust.2015.04.013
0003-682X/

energy, RASTA, jitter, shimmer, speech rate, harmony, and zero-crossing rates, were used to analyze speech characteristics in age and gender identification systems in previous studies. MFCCs offer a great deal of linguistic information for speech and speaker recognition applications [10]. Human perception of sound is based on a frequency analysis in the inner ear. MFCCs [11] are the cepstrum representations of this occurrence.

Zhan et al. compared the performance of Linear predictive coefficient (LPC), Linear-prediction cepstral coefficient (LPCC), MFCCs, and Bark-frequency cepstral coefficient (BFCC) feature sets by using a Gaussian mixture model (GMM), SVM, Multi-layer perception (MLP), kNN, and DS fusion classifiers in gender recognition. They reported that DS fusion classifiers achieved the highest gender classification accuracy of 93.07% female and 92.88% male, while the SVM classifier achieved 92.61% female and 92.28% male accuracy rates by using the MFCC feature set. It is also reported that LPC and LPCC sets achieved the lowest total classification accuracies for all classifiers [12].

Hu et al. proposed a two-level classifier with a pitch-based gender identification method to overcome the complexity of MFCC-based gender classification [13]. The first-stage classified the gender when pitch clearly indicates the gender of the speaker by using a threshold-based decision rule. The second-stage GMM classifier was used for undetermined speakers or difficult cases. They reported an accuracy rate of 98.65% for the TIDIGITS dataset.

Several speaker's age and gender classification studies are carried out by using aGender corpus. Ming et al. [38] proposed a method which combines five different acoustic level modeling methods as Gaussian Mixture Model (GMM) based on MFCC features, GMM–SVM mean supervector, GMM–SVM maximum likelihood linear regression (MLLR) supervector, GMM SVM Tandem supervector, and SVM baseline subsystems using 450-dimensional feature vectors including prosodic features. Their fusion system achieved 52.7% unweighted accuracy for the joint age and gender (age + gender) classification task and outperformed the GMM–MFCC system and SVM baseline, respectively, by 9.6% and 8.2%. Metze et al. [39] a comparative study in age and gender classification using aGender telephone speech corpus. They also compared the classification results with human performance on the same data. Four automatic classification methods, a parallel phone recognizer; dynamic Bayesian networks to combine prosodic features; linear prediction analysis; and GMM based on MFCC features are compared. Overall achieved accuracies were reported as 54%, 40%, 27%, and 42%, respectively. Overall classification accuracy by human listeners was reported as 55% for the aGender corpus. The classification of speakers' age and gender is a challenging task.

The calculation of MFCCs requires a large amount of storage space and has a high computation complexity. The performance of MFCC features is greatly affected by noisy recording environments. Although MFCCs are currently used for age and especially gender identification, temporal features of speech utterances may provide better information for age identification systems. We propose a PR feature set based on time–domain analysis. The proposed age and gender classification is based on three main steps: (1) pre-processing, which applies voice activity detection (VAD), (2) feature extraction, including MFCCs, Energy, RASTA_PLP, F0, and PR feature sets, and (3) classification. kNN and SVM classifiers are used in this stage to test the performance of selected combinations of the feature sets as well as individually is tested.

2. Database

In this work, data was taken from the aGender corpus [14,26]. It was supplied by the Interspeech 2010 Paralinguistic Challenge organization to support the development of speaker age and gender detection systems. The corpus consists of 49 h of telephone speech, stemming from 795 speakers, which are divided into a train (23 h, 471 speakers), development (14 h, 299 speakers) and test sets (12 h, 175 speakers) [14]. Four age groups make up the database: children, 7–14 years old (C); young-aged, 15–24 years old (YF/YM); middle-aged, 25–54 years old (MF/MM); and seniors, 55–80 years old (SF/SM). This choice was not motivated by physiological aspects that arise from the development of the human voice with increasing age, but rather it was based on market applications. Children are not subdivided as male and female speakers.

3. Methodology

Fig. 1 has three main steps. (1) The pre-processing step contains VAD to separate conversational speech from silence. There are many types of VAD algorithms. In this work, energy and zero-crossing rates [16] are used for VAD. (2) The feature extraction step calculates the feature sets used in classification. In addition to the calculation of well-known feature sets (MFCC, RASTA_PLP, F0), the proposed pitch-range (PR) feature set has been calculated. (3) The classification step uses kNN and SVM classifiers. kNN is a simple statistical learning algorithm in which an entity is classified by its neighbors. Computation time is short. SVM is a sophisticated supervised learning algorithm that requires training to determine the hyper-plane needed to separate classes accurately. The computation time of SVM can be long for multiple-class problems such as age and gender classification.

3.1. Pre-processing

The performance of speech processing applications is strongly affected by the quality of the speech signal. Although the speech signal is usually high-pass filtered to remove undesired low frequency components in practical speech applications, we do not do this in order to preserve spectral information that might be useful in age and gender classification. VAD is used in speech signal processing fields with the purpose of enhancing the quality of speech [15,16] before the feature extraction process. Short-time energy and zero-crossing measurements can be used in VAD. Fig. 2 shows short-time energy and zero-crossing rate of a speech utterance belonging to a child. A rectangular window of duration 32 ms (260 samples) is used with a 50% overlap.

An accurate and robust VAD plays an important role in the performance of the classifier. The theory of short-time energy and zero-crossing rate is briefly given below. The short-time energy is calculated as:

$$E = \sum_{n=0}^{N-1} |x(n)|^2 \tag{1}$$

where N is the window duration and $x(n)$ is the speech signal. Short-time zero-crossing rate is calculated as:

$$Z = \sum_{n=0}^{N-1} |sgn[x(n)] - sgn[x(n-1)]| \tag{2}$$

where $sgn[x(n)] = 1$ if $x(n) \geqslant 0$ and $sgn[x(n)] = 1$ if $x(n) < 0$.

3.2. Feature Extraction

Feature extraction is the process of calculating parameters that represent the characteristics of the input signal, whose output will have a direct and strong influence on the performance of classification systems. In this study, five different feature sets are calculated. They are MFCCs + energy (12 + 1 *features*), PLP-RASTA (13 *features*),

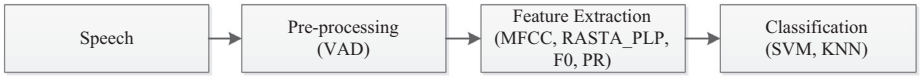

Fig. 1. Age and gender classification system.

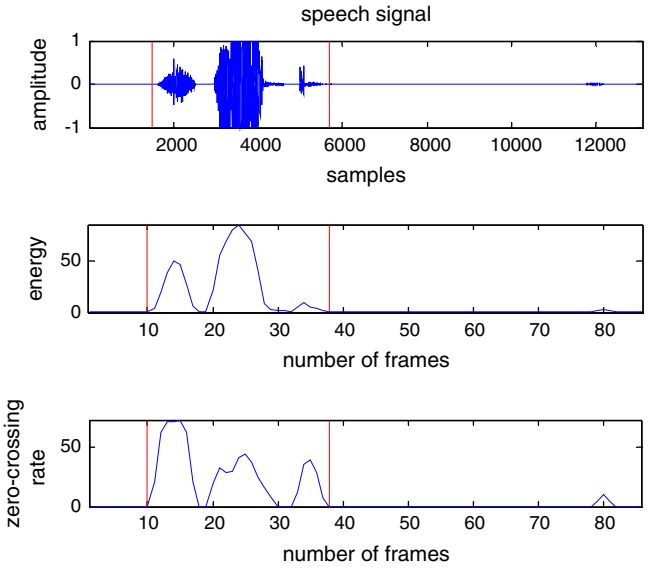

Fig. 2. Voice activity detection by energy and zero-crossing rate.

F0 (*one feature*), 3PR (6 *features*), and 20PR (40 *features*). Statistical analysis of feature sets, MFCCs and 3PR, is performed to understand the differentiative ability of the feature sets for seven aGender classes. Since MFCC features are studied extensively in literature [41–43] we do not include its low order moments, mean and standard deviation, and only include its high order moments analysis, skewness and kurtosis here. For 3PR set, both low-order and high-order moments are included. The reason of analyzing high-order moments of the feature sets lies in the fact that low-order moments describe a distribution of scores unrelatedly of whether inferential statistics are subsequently invoked. It is satisfactory when the data distribution is established as normal. However, in most cases there are some degree of skewness (asymmetry) and kurtosis (peakedness) in data distribution. Therefore the skewness and kurtosis analysis of the presented feature sets may portray the efficiency of them in age and gender classification.

3.2.1. MFCC

MFCCs have become steadfastly recognized as the basic feature vector for most speech and acoustic pattern recognition problems [4]. Their success is due to the ability to exemplify the speech amplitude spectrum in a concise form. A voice generated by a speaker is filtered by the shape of the vocal tract articulators such as the tongue, teeth, and nasal cavity. The resonance characteristics of the voice are affected by this shape. If the shape is controlled precisely, this should give a precise illustration of the phoneme being formed. The shape of the vocal tract shows itself in the envelope of the short time power spectrum, and the purpose of MFCCs is to accurately represent this envelope [17]. The steps for calculating MFCC features are shown in Fig. 3. The first step is pre-emphasis. Then, the speech signal is divided into frames, typically by adding a windowing function at a fixed time span. Normally, a Hamming window is used as the window function to remove edge effects so that a cepstral feature vector for each frame is generated. The next step is to apply a Discrete Fourier Transform (DFT) to each frame. Then, the values of power obtained from this spectrum

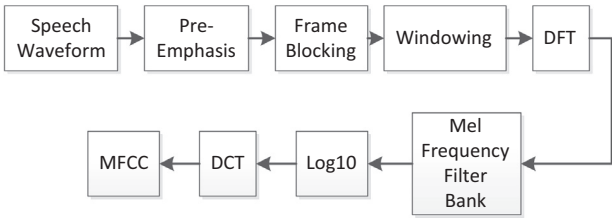

Fig. 3. Mel frequency cepstral analysis.

are mapped onto the 'mel' scale. The 'mel' scale is based on mapping actual frequency and detected pitch, since the human acoustic system does not detect pitch in a linear way. The next step is to take the logarithm of the powers at each of the 'mel' frequencies. The 'mel-spectral' vector components are highly correlated in this calculation. Therefore, in order to reduce the number of parameters in the system, the last step of MFCC feature calculation is to take the Discrete Cosine Transform (DCT). By using this transform, one energy and 12 cepstral features are obtained [18]. The main disadvantage of MFCCs is that it is sensitive to the environment and noise. From a statistical point of view, the traditional MFCC calculation based on windowed DFT is suboptimal because of high variance of the spectrum estimate [44–46].

MFCC (12 features) and energy features of C, YF, YM, MF, MM, SF, and SM are calculated in this work. The high-order moment analysis of MFCC set is given in Fig. 4.

Skewness is used as a measure of symmetry. A perfectly symmetric distribution is represented by the skewness of $S = 0$. A negative value indicates a skew to the left. The further from 0, the more skewed the data. Kurtosis is used to measure the flatness or peakedness of features. Positive kurtosis indicates a peaked distribution and negative kurtosis indicates a flat distribution [40]. The standard error of skewness can be calculated as $2\mathrm{sqrt}(6/12) = 1.414$ for MFCCs. Skewness, less than this, indicates the distribution of the features is symmetric. All seven classes show a very similar symmetric distribution. The flatness of the MFCC features and energy feature are also very similar for all classes. Therefore, we can say that the MFCC and energy feature set does not represent each age and gender class uniquely.

3.2.2. RASTA_PLP

This is an abbreviation for Relative Spectral Transform–Perceptual Linear Prediction. PLP was initially proposed by Hermansky to reduce the differences between speakers while retaining the main speech information [19]. However, this approach is vulnerable when the short-term spectral values are changed. The PLP provides limited functionality for dealing with these distortions by using a RASTA filter which makes PLP more powerful to handle linear spectral distortions. RASTA uses a band-pass filter in each frequency sub-band to smooth over short-term noise deviations [20]. The steps for calculating PLP features are shown in Fig. 5.

3.2.3. Fundamental frequency

F0 is known as voice pitch. Although there are conflicting results about the effects of age and gender on F0, it is a significant parameter for classifying male and female speech. For the voiced speech of an adult male, F0 can vary from 85 to 180 Hz, and for

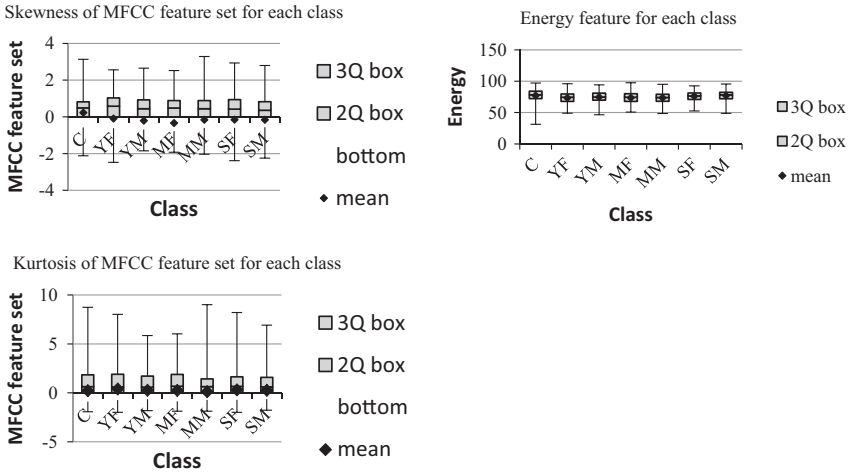

Fig. 4. Skewness and kurtosis of MFCC features and energy feature distribution for each class.

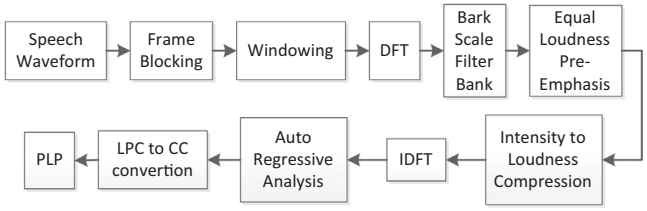

Fig. 5. Perceptual linear prediction.

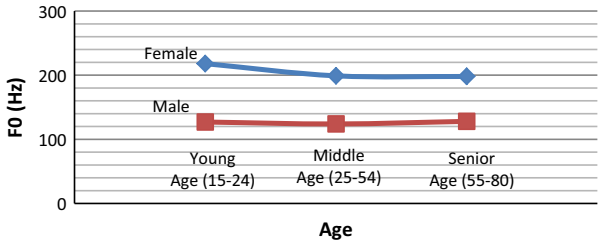

Fig. 6. Correlation between F0 and age for male and female speakers.

an adult female, it may vary from 165 to 255 Hz. For children, F0 is reported above 200 Hz. Table 1 and Fig. 6 present the mean F0, standard deviations, and changes in F0 for men and women versus age by using our database.

We have observed very similar values of mean F0 and standard deviation for young-aged and senior males and a significant decrease in F0 between young and middle-aged women. Little or no changes are observed in F0 between middle-aged and senior women. Mean F0 does not change linearly throughout life [27]. The highest mean F0 and standard deviation has found for children and female speakers, i.e. these groups have a larger F0 variation than male groups. An increase in F0 standard deviation with age has been observed for both sexes, similar to results from previous studies [28–30]. This increase implies variations and instability in speech. It is reported by many studies that both sexes show changes to laryngeal structure during the aging process. Refs. [28,29] stated that these changes are greater in women reported similar findings about the changes in F0 with age for Japanese male and female speakers.

3.2.4. Pitch-range (PR) feature set

Pitch analysis of speech signals is useful for many purposes. It has many applications such as speech separation, structured audio coding, automatic music transcription and music information retrieval. Barkana et al. introduced PR-based feature set for surveillance and environmental noise classification applications [5,21]. Their work showed that a PR-based feature set is promising in non-speech sound classification.

Pitch is the perceptual property of a sound wave detectable by the human ear. The pitch may be quantified as a frequency, but it is not a purely objective physical property. It is a subjective psycho-acoustical attribute of sound. In the real world, people are able to identify the pitch of several real-time sounds and separate each sound from a mixture of these. Pitch tracking in real-time situations usually involves additional steps beyond frame-by-frame pitch detection to enhance the quality of the measured pitch [21]. The autocorrelation function (ACF) technique creates the transient pitch of the input signal which will always contain some tracking error. Since speech signals of different age and gender speakers may change their acoustical characteristics in time, we focus on the "pitch range" instead of the pitch itself. Age and gender detection can be done according to how rapidly the pitch changes over time. The ACF is written as Eq. (3), where $x(n)$ is the speech signal and N represents the window size. The ACF measures the extent to which a signal correlates with a time offset (τ) version of itself [5].

$$\varnothing(\tau) = \sum_{n=0}^{N-1} x(n)x(n+\tau) \tag{3}$$

Pitch values are calculated using the short-time ACF method. Fig. 7 shows this process. The peaks simplify the similarity between the signal and a shifted copy of it. The time delay, T, between the first and the second positive peak values of the ACF for each window is calculated. Pitch, P, is defined as the reciprocal

Table 1
Mean fundamental frequency (F0) and standard deviations (in parentheses) for each class.

C Age: 7–14	YF Age: 15–24	YM Age: 15–24	MF Age: 25–54	MM Age: 25–54	SF Age: 55–80	SM Age: 55–80
236 (53)	218 (37)	127 (26)	199 (43)	124 (34)	198 (44)	128 (36)

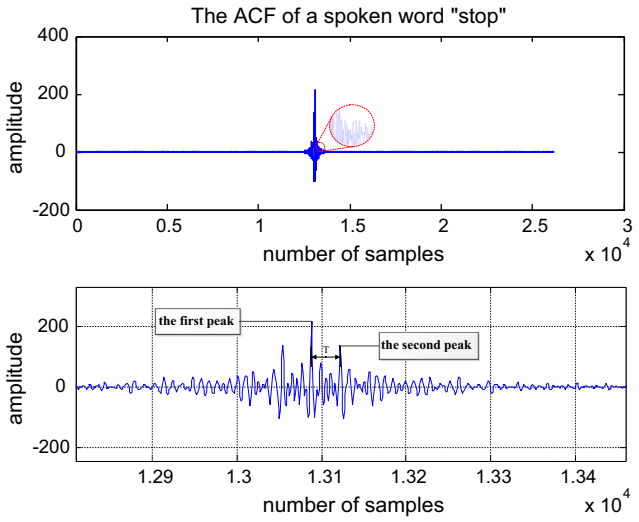

Fig. 7. ACF of a speech utterance "stop".

of the time delay, T, in Eq. (5), where M is the total number of the frames for any sound event.

$$T_i,\ 1 < i < M \tag{4}$$

$$p_i = \frac{1}{T_i},\ 1 < i < M \tag{5}$$

A PR feature set includes parameters which are calculated by using the maximum, minimum, mean, and standard deviation values for the pitch range. The first parameter, Par_1 in Eq. (6), is the ratio of the maximum and minimum of the pitch range, p_i where $1 < i < M$. The second parameter, Par_2 in Eq. (7), is the ratio of the standard deviation and the mean value of the pitch range, where $\bar{p}$ is mean value and $std\{p_i\}$ is the standard deviation of p_i.

$$\text{Par\_1} = max\{p_i\}/min\{p_i\} \tag{6}$$

$$\text{Par\_2} = std\{p_i\}/\bar{p} \tag{7}$$

$$\bar{p} = \frac{1}{N}\sum_{i=1}^{M} p_i \tag{8}$$

$$\text{std}\{p_i\} = \left(\frac{1}{N-1}\sum_{i=1}^{M}(p_i - \bar{p})^2\right)^{1/2} \tag{9}$$

In this work, PR feature sets are calculated by using two different designs that consist of low-pass filters (LPF) and high-pass (HPF) filters. Both designs are given in Figs. 8 and 9. The first design has one LPF (0–840 Hz) and two HPFs (1200–4000 Hz and 1600–4000 Hz). The second design has eighteen LPFs and two HPFs. The frequency bands of the filters are selected, after trial and error, to be in the range 0–4 kHz. Par_1 and Par_2 are extracted from each filter.

In this paper, we present the detailed analysis and evaluation of the 3PR set. The statistical analysis of it is performed by finding the mean, standard deviation, median, maximum, and minimum values for each class (see Table 2). Increased pitch range variations are observed with age for men and women in feature 1, although the highest variation is calculated for children. It is found that men have higher pitch range variations than women. Feature 1 is also fairly similar for both sexes. From observations of mean and median values, features 3 and 5 show slightly increasing uneven distribution in pitch range for both sexes with age. It is also observed that young-aged and senior male classes have similar characteristics regarding features 2 and 6.

Low order moments for PR (mean and standard deviation) do not vary enough to distinguish each class, and so higher orders (skewness and kurtosis) for PR feature set are investigated. Fig. 10(a) shows the skewness of the 3PR feature set for each age and gender class. The 3PR feature set of children is distributed more symmetrically, compared to men and women. Male speakers have less symmetric distributed 3PR features compared to female speakers. Women's 3PR feature data shows decreasing symmetry with age while there is no change for men. This information implies that C, YF, MF, SF classes can be classified with higher accuracies than YM, MM, and SM by using a 3PR set.

We also examined the kurtosis of the 3PR set as a measure of flatness or peakedness. Fig. 10(b) shows the kurtosis of 3PR

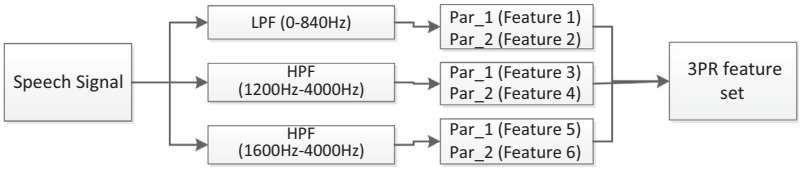

Fig. 8. Design 1: 3PR feature set; one LPF and two HPFs.

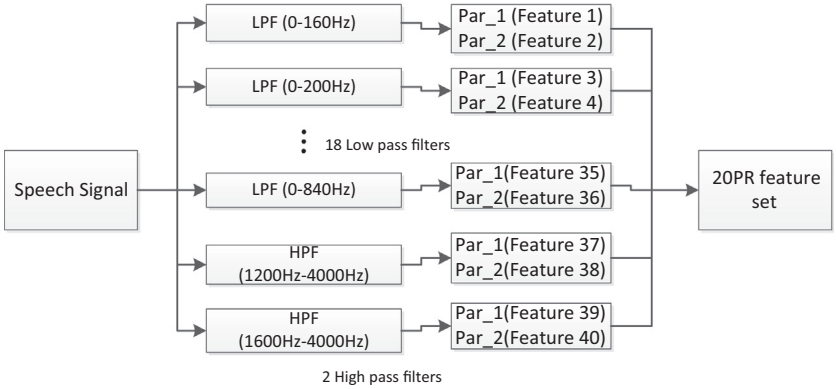

Fig. 9. Design 2: 20PR feature set; eighteen LPFs and two HPFs.

Table 2
Mean, standard deviation, median, maximum, and minimum values of 3PR features for each class. (mean; **std**; median; *max*; min).

	3PR feature set					
	Feature 1	Feature 2	Feature 3	Feature 4	Feature 5	Feature 6
C	16; **9.6**; 13.3; *66.3*; 1.6	1.1; **0.1**; 1.1; *1.9*; 1	11; **4.7**; 11.3; *23*; 1.8	1.2; **0.1**; 1.2; *1.6*; 1.1	11.6; **5.1**; 12; *24.6*; 1.8	1.2; **0.1**; 1.2; *1.4*; 1.1
YF	14.1; **6.5**; 13; *49.8*; 1.5	1.1; **0.1**; 1.1; *1.8*; 1	8.8; **4.8**; 9; *25.8*; 1.5	1.3; **0.1**; 1.3; *1.5*; 1.1	8.8; **5.2**; 8.5; *25.6*; 1.5	1.3; **0.1**; 1.3; *1.6*; 1.1
YM	18.4; **7.5**; 17.8; *49.6*; 3.1	1.2; **0.1**; 1.1; *1.8*; 1	6.6; **5.4**; 4; *30.4*; 1.5	1.3; **0.1**; 1.3; *2.3*; 1.2	6.3; **5.3**; 4; *27.7*; 2.5	1.3; **0.1**; 1.3; *1.6*; 1.2
MF	12.7; **6.8**; 10.9; *49.6*; 1.9	1.4; **0.3**; 1.4; *2.5*; 1	7.3; **4.7**; 6; *22.3*; 2.3	1.7; **0.4**; 1.8; *2.5*; 1.2	8.1; **4.7**; 6.6; *22.3*; 2.3	1.7; **0.4**; 1.6; *2.5*; 1.2
MM	21.8; **8.5**; 22.3; *66.1*; 3.6	1.2; **0.1**; 1.2; *2*; 1	7.8; **6.2**; 4.8; *30*; 1.9	1.3; **0.1**; 1.3; *1.7*; 1.2	5.9; **5.2**; 3.7; *34*; 1.7	1.4; **0.2**; 1.4; *2.3*; 1.2
SF	18.8; **7.9**; 17.6; *66.1*; 2.2	1.2; **0.1**; 1.2; *1.5*; 1.1	8.6; **5.6**; 6.7; *31.1*; 3	1.4; **0.1**; 1.4; *1.5*; 1.2	8.5; **5.6**; 6.6; *25.4*; 5	1.4; **0.1**; 1.4; *1.6*; 1.2
SM	19.7; **9.5**; 19.1; *64*; 1.4	1.2; **0.2**; 1.1; *2.1*; 1	6.8; **5.6**; 4.2; *36*; 1.4	1.4; **0.2**; 1.3; *2.1*; 1.1	6.2; **5.5**; 3.9; *34.4*; 1.4	1.4; **0.2**; 1.3; *2.2*; 1.2

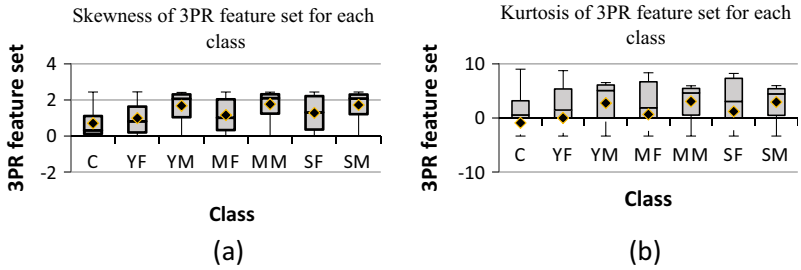

Fig. 10. Skewness and kurtosis of 3PR feature set for each age and gender group.

features for each class. Children have a flat 3PR feature distribution. Women's feature distribution changes from flat to peaked with age. Men have, relatively speaking, the same peaked distribution throughout the ages, and this will result in low classification accuracies for YM, MM, and SM classes.

4. Classification

To compare the performance of the presented feature sets, two well-known classification algorithms are used: k-Nearest-Neighbor (kNN) and Support Vector Machine (SVM). Both classifiers have been identified as the top 10 classification algorithms [22].

4.1. kNN: k-nearest neighbor classification

In machine learning, kNN is one of the simplest classifiers and is based on closest training examples in the feature space. The main idea of the kNN classification is that it finds a group of k objects in the training set that are closest to the test object. The class labels of the k-nearest neighbors are used to determine the class label of an unlabeled object. Euclidean distance or similarity metric is computed between objects. An unlabeled object is classified based on the distance of this object to the labeled objects [22]. Although the performance of a kNN classifier is affected by the choice of *k*, the distance measure, and training objects, kNN classifiers are widely used in many classification problems because of their easy implementation with good performance. In this work, leave-one-out cross-validated kNN classifier is used.

4.2. SVM: Support vector machines classification

SVM was developed by Vapnik [23] and has become a popular classifier algorithm recently because of its promising performance in different type of studies. SVM is based on structural risk minimization where the aim is to find a classifier that minimizes the boundary of the expected error [24]. In other words, it seeks a maximum margin separating the hyper-plane and the closest point of the training set between two classes of data [18]. SVM is one of the most robust and accurate methods among classification algorithms because of its sound theoretical foundation, less number of training set, and insensitivity to the number of dimensions. [22]. In our experiments we used the publicly available implementation LibSVM [25] with radial basis function (RBF) kernel since it yielded higher accuracies in the cross-validation compared to other kernels. The parameters are optimized by using a 3-fold cross-validation over the training dataset.

5. Experimental results and evaluation

We chose a random two thirds portion of the aGender database for training and the remaining one third for testing. The overall classification results of the feature sets for age + gender, age, and gender classification are given in Table 3. We have included 14 selected combinations of the feature sets in the table by using SVM and kNN classifiers. One should note that our purpose is not to compare kNN and SVM classifiers but to analyze and compare the performance of the feature sets for age and gender classification problems. Results show an accuracy of 35% in age + gender, 39.9% in age, and 64.6% in gender classification by using MFCC + Energy + F0 feature set (14 features) by using SVM. MFCC + Energy + F0 set delivered lower accuracies for age + gender and age classification compared to 3PR. We showed and discussed in the Feature Extraction section that MFCC and Energy features do not represent different age and gender classes uniquely (see Fig. 4). F0 is very effective to differentiate speakers' gender. On the other hand, it carries very little information about the age of a speaker, especially for male speakers. The F0 of the young-aged male and the senior male is almost the same, while the F0 of female speakers decreases with age (see Table 1 and Fig. 6). The proposed feature set, 3PR (6 features), achieved overall accuracy of 63.7% in age + gender, 40.9% in age, and 62.7% in gender classification by using an SVM classifier. For several cases of only age (four classes) or only gender (two classes) classification, kNN performed better

Table 3
Overall accuracies in age + gender, age, and gender classification.

Features		Case#1	Case#2	Case#3	Case#4	Case#5	Case#6	Case#7	Case#8	Case#9	Case#10	Case#11	Case#12	Case#13	Case#14
MFCC + Energy		√		√		√	√	√			√		√		√
RASTA_PLP			√		√	√	√	√				√		√	√
F0					√		√	√		√	√	√	√	√	√
3PR(1LPF + 2HPF)				√					√	√	√	√			√
20PR(18LPF + 2HPF)								√					√	√	
Dimension of feature vector		13	13	14	14	26	27	67	6	7	20	20	54	54	33
Age + Gender	kNN_ACC (%)	22.7	18.9	25.3	22.7	26.1	26.2	47.3	57.9	62.5	40.9	36.6	50.0	47.7	38.4
	SVM_ACC (%)	34.2	21.1	35.0	26.9	35.5	35.0	48.6	63.7	58.2	54.8	58.2	48.8	46.9	56.9
Age	kNN_ACC (%)	31.8	28.4	32.8	30.0	34.0	34.2	52.2	62.9	66.2	46.7	41.6	54.6	52.3	43.8
	SVM_ACC (%)	40.2	32.6	39.9	33.9	37.6	40.7	38.1	40.9	43.0	44.0	40.9	40.2	37.8	44.6
Gender	kNN_ACC (%)	67.6	62.7	72.4	70.3	71.5	73.1	76.9	77.0	82.5	75.2	74.3	74.8	73.8	77.2
	SVM_ACC (%)	75.4	64.6	84.7	79.9	75.4	84.7	83.5	62.7	77.2	84.4	79.9	83.5	78.1	82.9

than SVM. SVM and kNN handle data differently. kNN performs better when data has low level of sparsity Its performance is reduced in the case of high level sparsity since it fails to form reliable neighborhoods. SVM outperforms kNN when data has high level of sparsity. Regardless the classification algorithms, PR feature sets show stable performance in age + gender, age, and gender classification. Moreover, PR sets improve the performance of the other feature sets when they are used together. The confusion matrices of selected cases are given in Tables 4–9 by using LIBSVM classifier with 3-fold cross validation and an RBF kernel.

5.1. Age + gender classification

Children are classified with 83% accuracy by using 3PR set (Table 4). This class is misclassified as Senior Male (11.32%) and Young-aged Female (5.66%). The distribution of 3PR features for Children is observed to be more symmetric and flat compared to the other classes. Young-aged female speakers show a similar 3PR feature distribution to Children. This explains the misclassifications between children and young-aged female speakers. Further analysis of a 3PR set needs to be performed to explain the misclassification between children and senior male speech. YF, MF, and SF speech are correctly classified with ∼67%, ∼93%, and ∼84% accuracies by using a 3PR set, and its efficiency to separate YF, MF, and SF is shown in Fig. 10. For female speech, the skewness and kurtosis of the features decrease with age. Each age group is represented as uniquely as possible, affected by the classifier's performance. The lowest classification accuracy has been calculated for the young-aged male class as ∼5%. The majority of the YM class is misclassified as Senior Male and Middle-aged Male. This is not surprising since these classes' 3PR features have very similar symmetrical and flat distribution.

The addition of fundamental frequency to the 3PR feature set (Table 5) reduced the overall accuracy rates about 5% for C, YF, MF, MM, and SF while increasing the accuracies of YM and SM by around 3%.

Although F0 is one of the most effective features for gender classification, we observed a slight decrease of F0 for females and an insignificant increase of F0 for male speakers with age. The F0 distribution of our database is depicted in Fig. 6. Our results regarding

Table 4
Confusion matrix of age + gender classification by 3PR feature set (by SVM classifier).

(%)	C	YF	YM	MF	MM	SF	SM
C	**83.02**	5.66	0	0	0	0	11.32
YF	4.08	**67.35**	0	0	0	2.04	26.53
YM	0	4.55	**4.55**	2.27	25	9.09	54.54
MF	0	0	0	**92.86**	0	7.14	0
MM	0	0	0	7.55	**73.58**	15.09	3.77
SF	3.28	0	0	6.56	3.28	**83.61**	3.28
SM	7.46	0	1.49	10.45	5.97	40.30	**34.33**

Table 5
Confusion matrix of age + gender classification with 3PR + F0 feature set (by SVM classifier).

(%)	C	YF	YM	MF	MM	SF	SM
C	**77.35**	11.32	0	0	0	0	11.32
YF	8.16	**71.43**	0	4.08	0	0	16.33
YM	0	0	**9.09**	0	25	13.64	52.27
MF	0	0	0	**80.36**	0	19.64	0
MM	3.77	1.89	3.77	0	**54.72**	13.21	22.64
SF	1.64	0	1.64	11.48	8.20	**72.13**	4.92
SM	2.99	2.99	1.49	5.97	13.43	35.82	**37.31**

Table 6
Confusion matrix of age + gender classification using MFCC + Energy + F0 set (by SVM classifier).

(%)	C	YF	YM	MF	MM	SF	SM
C	**39.62**	20.75	0	7.55	3.77	22.64	5.66
YF	14.29	**32.65**	4.08	22.45	0	20.41	6.12
YM	2.27	0	**2.27**	0	36.36	2.27	56.82
MF	8.93	10.71	1.79	**26.79**	7.14	28.57	16.07
MM	0	0	9.43	3.77	**43.40**	0	43.40
SF	19.67	16.39	0	14.75	3.28	**26.23**	19.67
SM	7.46	1.49	0	4.48	17.91	5.97	**62.69**

Table 7
Confusion matrix of age + gender classification with MFCC + Energy + F0 + 3PR feature set (%) (by SVM classifier).

(%)	C	YF	YM	MF	MM	SF	SM
C	**60.38**	30.19	1.89	0	1.89	0	5.66
YF	4.08	**67.34**	0	14.29	0	2.04	12.25
YM	0	0	**18.18**	2.27	27.27	2.27	50
MF	0	0	0	**76.79**	5.36	17.85	0
MM	0	0	24.52	0	**52.83**	5.66	16.98
SF	1.64	1.64	1.64	4.92	8.20	**75.41**	6.56
SM	8.96	0	2.99	4.48	23.88	29.85	**29.85**

age-related changes of F0 are consistent with some previous results [29,31–34]. There are diverse and varying results in the literature about age-related changes in F0 for both sexes. While some of the previous studies observed a significant changes in F0 with age for male speakers [30,34–36], some observed a decrease in F0 with age for both sexes [37]. Therefore, we believe F0 is not a suitable feature in age classification.

The classification performance of the MFCC + Energy + F0 feature set is also examined (Table 6). Low classification accuracies are observed for all classes. All classes are misclassified as Senior Male to a certain degree. MFCC and Energy features did not represent the different age and gender groups uniquely. The statistical analysis of MFCC and Energy was discussed earlier (Fig. 4). Similar changes in MFCCs were found with age for both sexes. The combination of MFCC, Energy, F0, and 3PR sets did also not perform well (Table 7). In this work, we witnessed that Mel frequency cepstral coefficients are not suitable for age + gender classification.

5.2. Age classification

The proposed PR-based feature set has been tested for the age classification problem. The overall accuracies are given in Table 3. The highest average accuracy of 66.2% has been achieved by 3PR + F0 features with kNN classifier, while the MFCC + Energy + F0 set achieved an accuracy of 32.8% with the same classifier. The confusion matrices of MFCC + Energy, 3PR, MFCC + Energy + 3PR + F0, and 3PR + F0 are given in Tables 8a and 8b for SVM and kNN classifiers.

For all feature sets, 55–80 year old age group (Class#4) is misclassified most by the other age groups since senior male and female speech show more inconsistencies and variations resulting in overlapping with the others. As seen in Table 8b, the addition of F0 to 3PR feature set improved the accuracies of young-aged and senior classes. Young-aged class remains to have lower classification accuracies comparing to other classes since young-aged male speakers in this class are misclassified as senior male as a result of similar characteristics of their feature sets.

5.3. Gender classification

The highest overall accuracy of 84.7% has been achieved by using MFCC + Energy + F0 feature set in gender classification

Table 8a
Confusion matrix (%) of age classification of selected features sets. Class#1 = Children (7–14); Class#2 = YF + YM (15–24); Class#3 = MF + MM (25–54); Class#4 = SF + SM (55–80) (by SVM classifier).

Class (%)	MFCC + Energy				3PR				MFCC + Energy + 3PR + F0			
	#1	#2	#3	#4	#1	#2	#3	#4	#1	#2	#3	#4
#1	**22.64**	28.30	9.43	39.62	**54.72**	24.53	1.89	18.87	**50.94**	45.28	0	3.77
#2	1.06	**20.21**	15.95	62.76	0	**10.64**	10.64	78.72	0	**19.14**	11.7	69.1
#3	1.81	21.81	**20**	56.36	0.91	0.91	**35.45**	62.72	0	9.09	**40.9**	50
#4	4.65	9.30	6.97	**79.07**	1.55	6.20	30.23	**62.01**	3.10	9.30	25.6	**62.0**

Table 8b
Confusion matrix (%) of age classification of selected features sets. Class#1 = Children (7–14); Class#2 = YF + YM (15–24); Class#3 = MF + MM (25–54); Class#4 = SF + SM (55–80) (by kNN classifier).

Class (%)	MFCC + Energy				3PR				3PR + F0			
	#1	#2	#3	#4	#1	#2	#3	#4	#1	#2	#3	#4
#1	**26.89**	26.89	17.5	28.77	**68.4**	9.9	8.96	12.74	**68.87**	11.79	8.02	11.3
#2	11.5	**26.2**	28.1	34.22	7.49	**50.8**	14.44	27.27	7.75	**56.42**	15.24	20.59
#3	10.23	27.04	**31.2**	31.59	3.18	11.36	**70.9**	14.55	4.1	12.5	**69.77**	13.64
#4	10.7	21.79	29.2	**38.33**	5.06	18.48	14	**62.45**	2.53	14.59	13.8	**69.1**

Table 9
Confusion matrix (%) of gender classification of selected feature sets. (Female = YF + MF + SF and Male = YM + MM + SM) (by SVM classifier).

Class (%)	MFCC + Energy		3PR		F0		MFCC + Energy + 3PR + F0	
	Female	Male	Female	Male	Female	Male	Female	Male
Female	**76.79**	23.21	**51.19**	48.81	**78.57**	21.43	**82.73**	17.26
Male	26.06	**73.94**	25.45	**74.54**	18.79	**81.21**	13.94	**86.06**

problem (Table 3). Table 9 presents the confusion matrices for MFCC + Energy, 3PR, and F0 sets. The class of Children is not included in the gender classification.

F0 and MFCCs are frequency–domain analyses. They are well-studied and their effectiveness in speakers' gender classification is shown in the literature. PR features represent how rapidly the pitch changes over time. Different pitch values may have similar changes over time by resulting misclassifications between genders. The performances of different combinations of feature sets can be seen in Table 3. Based on our experimental results, we have observed that PR-based features do not provide any advantage in gender classification. Frequency domain features have showed higher performance in this work.

6. Conclusions

In this work, a pitch-range based feature set is proposed for text-independent age + gender classification applications. The performance of the PR features is compared with that of MFCC, Energy, RASTA_PLP, and F0 sets by kNN and SVM classifiers with the aGender corpus that contains telephone speech. Detailed statistical analysis is performed for the PR features regarding low-order and high-order moments. Both classifiers have achieved higher accuracies by using PR sets in age-related classification. In age + gender and age classifications, the highest accuracy rates are achieved by the proposed 3PR feature set, which contains only six coefficients. Middle-aged female speakers are classified with the accuracy of 92.86% and are followed by senior-female with 83.61%, children with 83.02%, middle-aged male with 73.58%, young-aged female with 67.35%, and senior male with 34.33% by the SVM classifier. Based on statistical analysis and experimental results, PR features are promising in speakers' age + gender classification. The classification accuracies may increase with a text-dependent corpus. The overlap between the classes of senior female/male and young male speakers will be investigated further in our future work.

It is observed that PR-based features, which are calculated by time-domain analysis and low-order moments, provide higher accuracies in age + gender and age classification compared to MFCC, F0, and RASTA_PLP. Frequency–domain features, MFCC and F0, have shown better performance for gender classification. At all times in age classification, kNN classifier has achieved higher accuracies with PR-based features comparing to SVM classifier. For example, kNN has classified four age groups with 31.8%, 62.9%, and 66.2% overall accuracies while SVM has achieved 40.2%, 40.9%, and 43% overall accuracies by using MFCC, 3PR, and 3PR + F0 features, respectively. In general, PR-based features and the addition of F0 improved the classification accuracies notably for age + gender and age classification regardless the type of classifier. Frequency–domain features are found to be more effective for classifying speaker's gender.

In future work, we plan to design a multi-level classifier. The first level will classify the gender of a speaker by using F0 and MFCCs while the second level classifies the speaker's age by using the PR features.

References

[1] Shafran I, Riley M, Mohri M. Voice signatures. In: Proc. of IEEE workshop automatic speech recognition and understanding; 2003. p. 31–6.
[2] Wu K, Childers DG. Gender recognition from speech part I: coarse analysis. J Acoust Soc Am 1991;90(4):1828–40.
[3] Bugalho M, Portelo J, Trancoso I, Pellegrini T, Abad A. Detecting audio events for semantic video search. Proc Interspeech 2009:1151–4.
[4] Rabiner LR, Schafer RW. Theory and applications of digital speech processing. 1st ed. Prentice Hall; 2011.
[5] Barkana BD, Uzkent B. Environmental noise classifier using a new set of feature parameters based on pitch range. Appl Acoust 2011;72(11):841–8.
[6] Cemto L, Falcone M, Paoloni A. Subjective age estimation of telephonic voice. Speech Commun 2000;31:107–12.
[7] Braun A, Cerrato L. Estimating speaker age across languages. Proc ICPhS 1999;2:1369–72.
[8] Minematsu N, Sekiguchi M, Hirose K. Automatic estimation of one's age with his/her speech based upon acoustic modeling techniques of speakers. Proc ICASSP 2002:137–40.
[9] Mueller C, Wittig F, Baus J. Exploiting speech for recognizing elderly users to respond to their special needs. Proc Interspeech Eurospeech 2003:1305–8.
[10] Wilpon JG, Jacobsen CN. A study of speech recognition for children and the elderly. Proc ICASSP 1996;1:349–52.
[11] Davis SB, Mermelstein P. Comparison of parametric representations for monosyllabic word recognition. IEEE Trans Acoust Speech Signal Process 1980;28(4):357–66.
[12] Zhan Y, Leung H, Kwak KC, Yoon H. Automated speaker recognition for home service robots using genetic algorithm and dempster-shafer fusion technique. IEEE Trans Instrum Meas 2009;58(9):3058–68.
[13] Hu Y, Wu D, Nucci A. Pitch-based gender identification with two-stage classification. Secur Commun Netw 2012;5(2):211–25.
[14] Schuller B, Steidl S, Batliner A, Burkhardt F, Devillers L, Mueller C, et al. The interspeech 2010 paralinguistic challenge. Proc Interspeech 2010:2794–7.
[15] Ramírez J, Górriz JM, Segura JC. Voice activity detection. Fund Speech Recogn Syst Robust 2007:1–22.
[16] Haigh JA, Mason JS. Robust voice activity detection using cepstral features. Proc IEEE TENCON 1993:321–4.
[17] Huang X, Acero A, Hon H. Spoken language processing: a guide to theory, algorithm, and system development. 1st ed. Prentice Hall; 2000.
[18] Logan B. Mel frequency cepstral coefficients for music modeling. In: Proc of International Symposium on Music Information Retrieval; 2000.
[19] Hermansky H. Perceptual linear predictive (PLP) analysis of speech. J Acoust Soc Am 1990;87(4):1738–52.
[20] Hermansky H, Morgan N. RASTA processing of speech. IEEE Trans Speech Audio Process 1994;2(4):578–89.
[21] Uzkent B, Barkana BD. Pitch-range based feature extraction for audio surveillance systems. Proc. of 8th ITNG 2011;476–80.
[22] Wu X, Kumar V, Quinlan JR, Ghosh J, Yang Q, Motoda H, et al. Top 10 algorithms in data mining. Knowl Inf Syst 2008;14:1–37. http://dx.doi.org/10.1007/s10115-007-0114-2.
[23] Vapnik, Vladimir N. The nature of statistical learning theory. 1st ed. New York: Springer-Verlag; 1995.
[24] Shao C, Bouchard M. Efficient classification of noisy speech using neural networks. Proc ISSPA 2003;1:357–60.
[25] Chang CC, Lin CJ. LIBSVM: a library for support vector machines. ACM Trans Intell Syst Technol 2011;2. 27:1-27:27.
[26] Burkhardt F, Eckert M, Johannsen W, Stegmann J. A database of age and gender annotated telephone speech. In: Proc. of 7th international conference on language resources and evaluation; 2010.
[27] Harnsberger JD, Shrivastav R, Brown Jr WS, Rothman H, Hollien H. Speaking rate and fundamental frequency as speech cues to perceived age. J Voice 2008;22(1):58–69.
[28] Gorham-Rowan MM, Laures-Gore J. Acoustic-perceptual correlates of voice quality in elderly men and women. J Commun Disord 2006;39:171–84.
[29] Nishio M, Niimi S. Changes in speaking fundamental frequency characteristics with aging. Folia Phoniatrica et Logopaedica 2008;60:120–7.
[30] Ramig LO, Gray S, Baker K, Corbin-Lewis K, Buder EH, Luschei E, et al. The aging voice: a review, treatment data and familial and genetic perspectives. Folia Phoniatrica et Logopaedica 2001;53(5):252–65.
[31] Torre III P, Barlow JA. Age-related changes in acoustic characteristics of adult speech. J Commun Disord 2009;42(5):324–33.
[32] Mueller PB. The aging voice. Semin Speech Lang 1997;18:159–69.
[33] Russel A, Penny L, Pemberton C. Speaking fundamental frequency changes over time in women: a longitudinal study. J Speech Hear Res 1995;38:101–9.
[34] Mysak E. Pitch and duration characteristics of males' voices. J Speech Hear Res 1959;2:46–54.
[35] Hollien H, Shipp T. Speaking fundamental frequency and chronological age in males. J Speech Hear Res 1972;15:155–9.
[36] Benjamin BJ. Frequency variability in the aged voice. J Gerontol 1981;36:722–6.
[37] Harrington J, Palethorpe S, Watson CI. Age-related changes in fundamental frequency and formats: a longitudinal study of four speakers. In: Proc of interspeech; 2007.
[38] Ming Li, Chi-Sang Jung, Kyu J. Han, Combining five acoustic level modeling methods for automatic speaker age and gender recognition. In: Proc of Interspeech; 2010.
[39] Metze F, Ajmera J, Englert R, Bub U, Burkhardt F, Stegmann J, et al. Comparison of four approaches to age and gender recognition for telephone applications. In: Proc of Interspeech; 2007.
[40] Newell KM, Hancock PA. Forgotten moments: a note on skewness and kurtosis as influential factors in inferences extrapolated from response distributions. J Mot Behav 1984;16(3):320–35.
[41] Kinnunen T, Saeidi R, Sedlák F, Lee KA, Johan Sandberg, Hansson-Sandsten M, et al. Low-variance multitaper MFCC features: a case study in robust speaker verification. IEEE Trans Audio Speech Lang Process 2012;20(7).

[42] Sangwan A, R. Muralishankar R, O'Shaughnessy D, Performance analysis of the warped discrete cosine transform cepstrum with MFCC using different classifiers. In: IEEE workshop on machine learning for signal processing; 2005.
[43] Polzehl T. Personality in speech assessment and automatic classification. Springer International Publishing; 2015.
[44] Percival DB, Walden AT. Spectral analysis for physical applications. Cambridge, MA: Cambridge Univ. Press; 1993.
[45] Zhang WQ, Deng Y, He L, Liu J. Variant time-frequency cepstral features for speaker, recognition. Proc Interspeech 2010:2122–5.
[46] Zhang WQ, He L, Deng Y, Liu J, Johnson M. Time frequency cepstral features and heteroscedastic linear discriminant analysis for language recognition. IEEE Trans Audio Speech Lang Proc 2011;19(2):266–76.