International Journal of Engineering

Journal Homepage: www.ije.ir

Voice-based Age and Gender Recognition Based on Learning Generative Sparse Models

S. Mavaddati*

Department of Engineering and Technology, University of Mazandaran, Babolsar, Iran

P A P E R I N F O

Paper history:
Received 10 March 2018
Received in revised form 16May 2018
Accepted 30May 2018

Keywords:
Gender Recognition
Sparse Non-negative Matrix Factorization
Incoherence
Mel-frequency Cepstral Coefficient
Voice Processing

A B S T R A C T

Voiced-based age detection and gender recognition are important problems in the telephone speech processing to investigate the identity of an individual. In this paper, a new gender and age recognition system is introduced based on the generative incoherent models learned using sparse non-negative matrix factorization and the atom correction step as a post-processing method. The proposed classification algorithm includes training step to provide the appropriate trained atoms for each data class and also the test phase to assess the classification performance. Since the classification accuracy depends highly on the selected features, the Mel-frequency cepstral coefficients are employed to train basis for the better representation of the voice structure. These bases are learned over the data of male and female speakers using non-negative matrix factorization with the sparsity constraint. Then, atom correction is carried out using an energy-based algorithm to decrease the coherence between different categories of the trained dictionaries. In the sparse representation of each data class, the atoms related to other sets with the highest energy are replaced with the lowest energy bases if the reconstruction error does not exceed from a specified limit. The experimental results showed that the proposed algorithm performs better than the earlier methods in this context especially in the presence of background noise.

***doi**: 10.5829/ije.2018.31.09c.08*

1. INTRODUCTION

Individual identification is an important area in the voice processing that can be employed in different signal processing fields such as criminal cases, threatening calls, electronic banking and kidnapping [1-2]. Different biometric characteristics are used in identification process for example fingerprint, forensic verification, iris scanning, signature, face geometry and voice components [2-3]. Selection of these characteristics depends on the application and what kind of data is available. In this paper, an age/gender recognition system using features extracted from the voice signal is proposed. Several voice-based gender recognition algorithms have been presented in the recent years. Bocklet et al. [4] have trained supervectors for each speaker based on Gaussian mixture model (GMM).

They have employed as universal background models to classify the speakers based on support vector machine (SVM). Also, different kernels for SVM were investigated to determine the classification accuracy. A comparative study of different gender recognition methods on the telephone speech is investigated [5]. These identification systems are based on a parallel phone recognizer, Bayesian networks-based, linear prediction-based analyzer and GMM-based model. These mentioned classifiers employ Mel-frequency cepstral coefficients (MFCCs) as feature vectors. The first designed recognizer system results in a better performance than other methods when the observed utterance is not short. A combination model based on GMM and SVM classifier is presented that calculates the Gaussian weight supervectors for each speaker and employ them as features for the SVM classifier [6]. A gender detection algorithm using weighted supervised non-negative matrix factorization (WSNMF) and general regression neural network (GRNN) is introduced in literature [7]. This hybrid model is used to earn supervectors of SVM for training data and recognize the gender of the test signals. A gender recognition method

Corresponding Author's Email: *s.mavaddati@umz.ac.ir* (S. Mavaddati)

Please cite this article as: S. Mavaddati, Voice-based Age and Gender Recognition Based on Learning Generative Sparse Models, International Journal of Engineering (IJE), IJE TRANSACTIONS C: Aspects Vol. 31, No. 9, (September 2018) 1529-1535

in combination with acoustic and prosodic levels using different baseline subsystems is introduced [8]. SVM and GMM classifiers are employed to detect the gender of speakers. The discrete cosine transform (DCT) is employed to provide cepstral trajectories corresponding to the specified modulation frequencies [9]. Using these features in combination with the prosodic features such as energy, pitch resonance, formants frequencies and length of the vocal tract, a proper detection rate in age/gender estimation problem is obtained. In this paper, it is shown how to integrate sparse modeling using sparse non-negative matrix factorization (SNMF)with the incoherency concepts in order to represent the male and female voice components and recognize the gender/age of speakers. these factorization parameters are set in such a way to reduce the approximation error in the learning process and increase the classification rate. The MFCC features are used to model the structure of voice frames with different speakers. Also, an atom correction algorithm is employed to decrease incoherency measure between the learned atoms of male and female speakers and yield a factorization process over independent atoms. The novelties of the proposed method are as follows. Firstly, the proposed method uses the generative incoherent models learned based on the sparse non-negative matrix factorization approach. Secondly, the atom correction step is introduced to reduce the reconstruction error in the sparse factorization of each data frame over the related dictionary. Also, for the first time, the incoherent atoms are learned for each dictionary related to different categories based on employing IPR technique. Finally, a robust age/gender recognition algorithm is presented that works with an appropriate classification rate in the presence of white noise. The outline of this paper is organized as follows: In section 2, we describe age/gender recognition problem. Section 3 has a detailed overview of the proposed factorization method based on SNMF. In section 4, the experimental results are expressed. Finally, we provide the experimental results in section 5 and conclude the paper in section 6.

2. PROBLEM DESCRIPTION

The input signal should be transferred into a feature domain to simplify the signal analysis with maintaining the features. The MFCC feature domain is a good selection to present the most important content of the voice signals in a compact form [10-11]. In this paper, this feature domain is selected for better representation of the input data in the learning process. The voice signal $S \in \mathbb{R}^{M}$ can be represented linearly by SNMF algorithm as S=WH, where $W \in \mathbb{R}^{M\times P}, P > M$ is a dictionary matrix with P atoms shown by $\{w_p\}_{p=1}^{P}$ with unit norm $\|w_{(:,p)}\|_2 \ \forall p = 1, \dots, P$ and the activity matrix $H \in \mathbb{R}^{P}, P \gg K$ involves the coefficients of factorization process over W [12-13]. K, is the cardinality parameter or the non-zero coefficients in the sparse representation of a data frame. Non-negativity constraints should be considered for all coefficients of W and H. This factorization with reconstruction error and sparsity parameter is defined in the following problem [12-13]

$$F(W,H) = \|S - WH\|_2^2 + \alpha \sum_p h_p \quad (1)$$

where α is the weighted factor. The first term considers the approximation error and the second term calculates sparsity for each row of H. The approximation error can be measured based on the generalized Kullback-Leibler divergence. Therefore, SNMF is formulated using the following optimization problem [12-13].

$$F(W,H) = \sum_{m,p} \left(S_{m,p}.log(S_{m,p} / (WH)_{m,p}) - S_{m,p} \right) + \alpha \|h_p\|_0 \quad (2)$$

where $\|H\|_0$is sparsity constraint or the number of non-zero coefficients in each row of H.

3. OVERVIEW OF THE PROPOSED RECOGNITION ALGORITHM

Figure 1 indicates the block diagram of the proposed age/gender classification algorithm. At first, the MFCC features of the input data from each defined classes in the training and test steps are calculated. The detail of each block has been illustrated in the following sections.

3. 1. Incoherent Atom Learning Using SNMF A new solution for age/gender recognition problem using dictionary learning concepts is presented in this paper. The observed data in this learning process can be approximately modeled as a weighted linear combination of a small number of basis vectors to reduce dimensionality. In SNMF, the sparsity parameter is imposed to H to control the speed of learning process. In the training step, the generative models for training data related to the male and female speakers are learned using SNMF method. The training data in this paper is divided into seven sets that each of them involves the telephone speech of male and female speakers in different age and gender ranges. These categories for different age and gender are shown in Table 1. These age/gender sets denote C, YM, YF, MM, MF, SM, SF for children, young male, young female, middle male, middle female, senior male, senior female, respectively. The atoms and activity matrix related to each train set are yielded from the following equation [12-13].

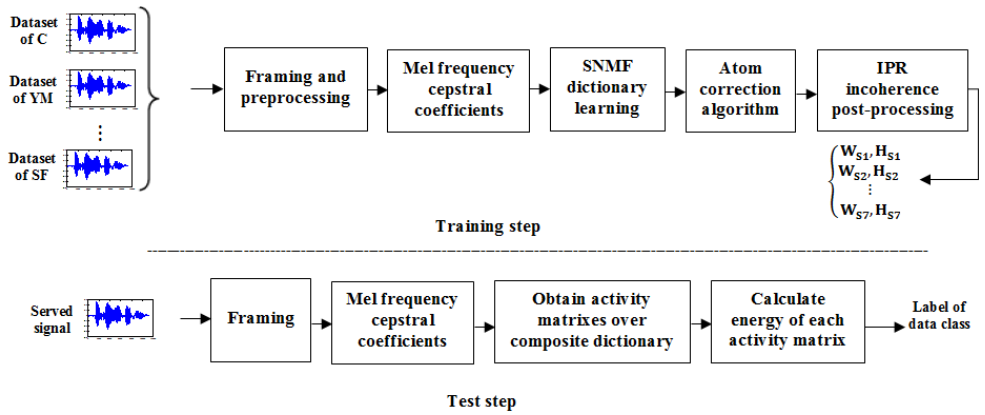

Figure 1. Block diagram of the proposed age/gender recognition algorithm with learning and test steps

TABLE 1. Age/gender categories in the proposed classification problem

Age/ gender class	Age (years)/ gender	Number of speakers
C	≤13 (male or female)	15
YM	14-19 (male)	25
YF	14-19 (female)	25
AM	20-64 (male)	25
AF	20-64 (female)	25
SM	≥65 (male)	25
SF	≥65 (female)	15

$$W_S^*, H_S^* = arg \min_{W_S, H_S} \| S - W_S H_S \|_2^2 \; s.t. \; \| H_S \|_0 \le K \quad (3)$$

where K is cardinality value in sparsity constraint.

In the next step of the training process, the atoms correction step is performed to ensure that each data class is only matched to the related atoms. Thus, the atoms with the highest energy values in the sparse factorization of each data class over composite dictionary (a dictionary consists of all seven dictionaries) are replaced with atoms with the lowest energy values in each row of H. These replacements are done under this condition that the selected atoms with the lowest energy must have the minimum reconstruction error in the sparse factorization for the related data class. The same operation is carried out for each data class to replace ten percent of the atoms with more desired atoms. If an atom does not have a change in the consecutive iterations, the correction algorithm stops. This atom correction was first introduced for the speech enhancement [14] and in this paper is employed to recognize the age and gender ranges of different speakers. The main issue that should be considered when more than one dictionary participates in the factorization process is mutual coherence parameter in the learning procedure. This parameter determines the dependency between the learned atoms related to the one data class. Lower mutual value obtains a dictionary with high independency between its basis vectors. In this paper, a post-processing method is employed to achieve the incoherent dictionaries. This parameter is set in such a way that the reconstruction error is reduced in sparse representation and the factorization process is done with more accuracy. This purpose will be achieved if the atoms have low coherence value similar to an equiangular tight frame (ETF) [15]. The coherence measure is defined as the maximum absolute value calculated from the correlation between different atoms or maximum absolute value of off-diagonal elementsin Gram matrix $G = D^T D$ with normalized atoms [15].

So, an iterative projection and rotation (IPR) post-processing technique is employed to result in an incoherent dictionary [16]. This algorithm adjusts the structural and spectral constraints on the number of non-zero eigenvalues of the gram matrix as a projection step. Then, the residual norm of the approximation error is minimized by rotation of atoms with an orthogonal matrix [15]. The importance of training incoherent atoms and the prominent rule of this procedure in obtaining better results have been proved in the earlier speech enhancement methods [17-18]. In our previous works [14], a modified version of K-SVD algorithm for incoherent dictionary learning was proposed.

The least angle regression with coherence criterion (LARC) sparse coding is used to increase coherence between the atoms and training data and IPR method for decreasing the mutual coherence between the dictionary atoms [14]. The spectrum coefficients are considered as the training data in this learning process. A low-rank sparse decomposition model is introduced using LARC sparse coding and IPR post-processing to train incoherent

dictionaries in the time-frequency domain [17]. These dictionaries are employed in the factorization process based on the robust principal component analysis. An incoherent dictionary learning process in wavelet packet transform domain is proposed in literature [18]. In this process, Limited-memory BFGS algorithm was employed to solve two resulted sub-problems in approximation and detail subbands with an iterative procedure [18]. The presented scheme for solving this classification problem using trained generative model is detailed in Algorithm 1. The dictionary atoms that have more coherence with the training data frames are obtained using the Algorithm 1 to represent precisely different data classes. If any error occurs in SNMF step for training different atoms, it can be compensated due to the existence of a generative model fitted to each data category. The MFCC is the most widely used feature in speech processing algorithms particularly for speaker modeling [4]. This feature is applied for representation of each input frame to detect age and gender of different speakers. These feature vectors involve 13 standard features with the energy coefficient and their first and second order derivatives. Finally, 39 MFCC coefficients are extracted for each training frame [11, 19]. In the test step, a composite dictionary is employed to factorize the observed signal over the related dictionaries.

The composite dictionary will be consisting of all seven dictionaries corresponding to each data class. Then, the activity matrices of this sparse representation over all learned dictionaries are calculated. Each dictionary that obtains the highest energy for its sparse coefficients, determines the input data label. In fact, a dictionary with more energy for its coefficient matrix obtained using SNMF of the test data over the composite dictionary reflects the better representation of the test data on the learned atoms. This means that there is a greater similarity between the test data and the trained atoms. So the data label is set according to these atoms.

4. IMPLEMENTATION DETAILS

The proposed recognition method is assessed using different evaluation measures using SpeechDat II corpus that is a telephone speech databases with gender and age labels [20]. This large multi-talker data base has been used for research about speech recognition and speaker verification problem that includes words or sentences read by 5000 speakers of both genders. The training and test sets include 100 speakers, 70 and 30 speakers in the training and test steps, respectively. The sampling rate of the input data is 8 kHz and feature vectors are yielded using 39 MFCC coefficients. The number of Mel filters is set to 12. The signals are segmented with 37.5 ms frame length and 40% overlap with a Hamming window. All framing, pre-processing steps are the same in the training and test steps of different speakers. The defined parameters in different steps of Algorithm 1 are assigned according to the experimental evaluations. The value of ε parameters is set to 0.01 and 0.015 for male and female speakers. Also, the value of ε_1 is adjusted to 0.01 for both male and female speakers. N, μ_0, K are set to 10, 0.2 and 8 for all training sets, respectively.

5. EVALUATIONS

The objective of this paper is to present a novel age/gender detection method which leads to superior classification rate over other methods and also in the presence of white noise.

Algorithm 1: Proposed sparse modeltraining procedure for age/gender detection

Input: $W_{S1}, \dots, W_{S7}$, $S_1,\dots,S_7$, K (cardinality value) , T (number of iteration), μ_0(specified coherence value), N (number of selected eigenvalue)

Output:$H_{S1}, \dots, H_{S7}$, $W_{S1}, \dots, W_{S7}$

Initialization: $H_{S1}, \dots, H_{S7} = [0]$

% Train dictionary atoms using SNMF for each data class

For*i=1:7*

$W_{Si}^{*}, H_{Si}^{*} = arg \min_{W_{Si},H_{Si}} \|S_i - W_{Si}H_{Si}\|_2^2 \; s.t. \; \|H_{Si}\|_0 \le K$

End

% Using atom correction step to provide atoms with maximum coherence to train data

For*t*=1→ *T*

% For each row of a activity matrix H_{Si}^{*}

$l^{max} = arg \max_{1\le l\le L} norm(H_{Si}^{*}) \; or \; arg \min_{1\le l\le L} sparsity(H_{Si}^{*})$

$l^{min} = arg \min_{1\le l\le L} norm(H_{Si}^{*}) \; or \; arg \max_{1\le l\le L} sparsity(H_{Si}^{*})$

% Using SNMF

$$S_i = [W_{S1}^{*} W_{S2}^{*} \dots . W_{S7}^{*}] \begin{bmatrix} H_{S1}^{*} \\ H_{S2}^{*} \\ \vdots \\ H_{S7}^{*} \end{bmatrix}$$

% Column number l^{max} *of* W_{Si}^{*} *is replaced by colun numberl*^{min} *and yields* $\widehat{W}_{Si}^{*}$, *if*$\|S_i - W_{Si}^{*} H_{Si}^{*}\|_F^2 < \varepsilon_1$

% Convergence condition

If $\left\|W_{Si}^{*} - \widehat{W}_{Si}^{*}\right\|_F^2 > \varepsilon$ *Then t=t+1*

End

%Using IPR technique to earn atoms with minimum coherence value to each other

For *i=1:7*

% Structural projection of G to ETF matrix K_{μ_0} *with specified coherence value* μ_0 *using IPR*

%Projection step:$G_i = \widehat{W}_{Si}^{*} \widehat{W}_{Si}^{*T}$

$$W_{Si}^{**} = Thresh\left(G_{i,j}|_{i\neq j}, \mu_0\right) = \begin{cases} g_i \; if \; |g_i| \le \mu_0 \\ \mu_0 \quad if \; g_i > \mu_0 \\ -\mu_0 \; if \; g_i < -\mu_0 \end{cases}$$

% Spectral projection by limiting the eigenvalue numbers (λ) *(obtained from singular value decomposition (SVD)) of G to N largest values*

$SVD(G_i) = Q_i \Lambda_i {Q_i}^T$

$\widehat{\Lambda_i} = Thresh(\Lambda_i, N) = \begin{cases} \lambda_i & if\ i \le N, \lambda_i > 0 \\ 0 & if\ i > N\ or\ \lambda_i < 0 \end{cases}$

$\widehat{G_i} = Q_i \widehat{\Lambda} {Q_i}^T$

% Rotation process to reduce approximation error in factorization procedure using orthogonal rotation matrix D

$W_{Si}^{***} = arg\min_{w} \|Y - D{W_{Si}}^{**}{H_{Si}}^{**}\|_F^2$

End

The assessment of the proposed classification method is done using recognition measure defined as the number of correct classifications divided by the total number of test data. The average values of recognition accuracy results for all test data related to different age/gender sets are reported in Table 2. These experimental results show that the proposed detection algorithm obtains better classification precision than other mentioned algorithms for different defined categories. Also, the average results over age sets as young (Y), middle (M) and senior (S) are summarized in Table 3 that indicates the superiority of the proposed scheme. Table 4 shows the relative confusion matrix of the proposed method for different age/gender categories in percentage.

TABLE 2. Recognition rate of different methods for age/gender categories

	C	YM	YF	MM	MF	SM	SF
[7]	54.2	61.4	60.3	59.6	60.7	59.2	53
[8]	61.1	66	68.2	66.4	62.1	63.3	59.1
Proposed	**65.2**	**74.3**	**73.4**	**72.3**	**68.1**	**70.4**	**66.8**

TABLE 3. Average value of recognition rate of different methods for age/gender categories

	C	Y	M	S
[7]	54.23	60.89	60.15	56.14
[8]	61.18	67.13	62.70	61.23
Proposed	**65.21**	**73.89**	**69.24**	**68.59**

TABLE 4. Relative confusion matrix of the proposed method for age/gender categories

Detected label / Data label	C	YM	YF	MM	MF	SM	SF
C	**74.7**	6.14	5.21	4.10	4.47	3.27	2.05
YM	5.87	**71.9**	5.13	6.25	4.32	3.54	2.96
YF	5.42	4.25	**75.4**	4.67	4.01	3.56	2.65
MM	4.21	6.51	4.25	**75.5**	4.61	3.25	1.68
MF	4.01	3.52	4.27	4.06	**79**	3.51	1.62
SM	2.19	2.76	2.49	4.25	3.76	**81.3**	3.28
SF	3.98	3.03	2.42	3.01	3.15	4.11	**80.30**

This evaluation matrix shows that each data set is assigned to the specific class with high accuracy value. As mentioned, the performance of the presented approach is assessed in the presence of white noise at different signal to noise ratios (SNRs). In fact, the dictionaries are trained with clean data and in the test step, white Gaussian noise is added to the input signal. The monaural recording of the speech signal in a real environment can be linearly defined as

$$Y(t) = S(t) + N(t) \quad (4)$$

where Y(t), S(t) and N(t) are the observed, voice and noise signal. The white noise signal is selected from Noisex92 [21]. The results of this recognition scenario for test data mixed with white noise signal at SNR= +5dB are expressed in Table 5. The average results for different age classes are shown in Table 6. Moreover, the confusion matrix is reported for the proposed method in Table 7.

TABLE 5. Recognition rate of different methods for age/gender categories in the presence of white noise at SNR=+5 dB

	C	YM	YF	MM	MF	SM	SF
[7]	50.1	57.2	56.9	56.7	58.5	56.8	51
[8]	57.5	62.3	65	63	57.6	59.2	56.3
Proposed	**62.7**	**73.2**	**70.4**	**69.4**	**66.1**	**68.9**	**64**

TABLE 6. Average value of recognition rate of different methods for age/gender categories in the presence of white noise at SNR=+5 dB

	C	Y	M	S
[7]	50.12	57.10	57.63	53.94
[8]	57.53	63.66	60.31	57.78
Proposed	**62.76**	**71.81**	**67.77**	**66.47**

TABLE 7. Relative confusion matrix of the proposed method for age/gender categories in the presence of white noise at SNR=+5dB

Detected label / Data label	C	YM	YF	MM	MF	SM	SF
C	**65.4**	7.78	6.89	5.47	5.67	4.59	4.21
YM	6.9	**66.9**	5.34	6.38	5.41	4.87	4.11
YF	6	6.24	**67.1**	5.01	5.92	5.69	4.03
MM	5.78	7.03	6.44	**66.1**	6.21	4.88	3.52
MF	6.22	5.14	5.13	5.94	**68.9**	4.77	3.82
SM	4.56	5.06	4.75	4.10	5.42	**70.3**	5.81
SF	3.01	4.45	5.86	4.51	5.89	6.71	**69.5**

As can be seen, the performance of all methods have slightly decreased when the noisy test signals have seen but the proposed method has also been able to get better results in comparison with other algorithms. These results when the test signal is polluted with the white noise signal at SNR= +5dB are expressed in Tables 8-10. The results are the averaged value over all observed data of each age/gender category. Since, the energy of noise signal increases in this test scenario, the recognition accuracy values are prominently reduced. The experimental results showed that the proposed algorithm with pre-trained incoherence dictionaries achieved the best accuracy rates in this test condition and outperform other methods in this context. The white noise signal is unstructured and cannot be sparsely modeled over a trained dictionary W_s. The components of white noise are disregard in factorization over a fixed dictionary and represented in the reconstruction error term of Equation (1).

TABLE 8. Recognition rate of different methods for age/gender categories in the presence of white noise at SNR=0dB

	C	YM	YF	MM	MF	SM	SF
[7]	28.3	31	29.5	30.5	31.6	33.7	29.9
[8]	35.6	37.8	36.2	35.6	33.2	36	32.8
Proposed	**50.2**	**51.2**	**52.3**	**53.2**	**54.4**	**53.6**	**51.4**

TABLE 9. Average value of recognition rate of different methods for age/gender categories in the presence of white noise at SNR=0dB

	C	Y	M	S
[7]	28.31	30.28	31.07	31.82
[8]	35.63	37.01	34.42	34.41
Proposed	**50.23**	**51.77**	**53.83**	**52.51**

TABLE 10. Relative confusion matrix of the proposed method for age/gender categories in the presence of white noise at SNR=0dB

Detected label / Data label	C	YM	YF	MM	MF	SM	SF
C	**47.7**	8.62	8.69	8.03	8.33	9.89	8.74
YM	7.51	**47.3**	9.88	8.59	9.39	8.54	8.75
YF	8.58	7.91	**52**	7.43	8.14	7.52	8.43
MM	7.28	8.53	7.88	**51.3**	8.05	7.96	9.02
MF	8.59	7.58	8.27	7.93	**50.7**	8.24	8.65
SM	7.54	8.33	7.59	8.46	8.54	**50.8**	8.57
SF	8.63	8.05	8.47	9.21	7.92	8.83	**48.9**

Therefore, the reason of superiority of the proposed method in the presence of white noise is that this unstructured noise is incoherent to the voice dictionary and can be neglected properly during factorization process.This superiority originates from two issues, the first is using the atom correction technique to ignore the atoms with unimportant rules in the representation of the input frames. The second issue is training incoherent models for the input data related to each set based on IPR to reduce the coherence between dictionary atoms.

6. CONCLUSION

In this paper, the incoherent modeling and sparse non-negative matrix factorization are applied to age and gender recognition problem in the telephone speech processing. The proposed method employs atom correction algorithm and iteration projection-rotation technique as post-processing methods to earn dictionaries with high data-atom coherence and low mutual coherence for a fixed dictionary. Also, Mel-frequency cepstral coefficients are used as training data to represent precisely the structure of each input component. The dictionary atoms are trained over data of male and female speakers in different age ranges using the sparsity constraint imposed to non-negative matrix factorization method. The experimental results show that the better recognition rates are yielded using the proposed algorithm in comparison with other algorithms in this processing filed. In order to have more investigation about the performance of the proposed method, evaluations done in the presence of background white Gaussian noise and similar results earned for the presented recognition scheme.

7. REFERENCES

1. Tanner, D. C., Tanner, M. E., "Forensic aspects of speech patterns: voice prints, speaker profiling, lie and intoxication detection", Lawyers & Judges Publishing Company, (2004).
2. Reddy, T. R., Vardhanb. B. V., Reddy P. V., " A document weighted approach for gender and age prediction based on term weight measure", ***International Journal of Engineering-Transactions B: Applications***, Vol. 30, No. 5, (2017), 643-651.
3. Jain, A. K., Flynn, P., Ross, A. A., Handbook of biometrics, Springer, (2008).
4. Bocklet, T., Maier, A., Bauer, J. G., Burkhardt, F., Noth, E., "Age and gender recognition for telephone applications based on GMM supervectors and support vector machines" , ***In proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)***, USA, (2008), 1605-1608.
5. Metze, F., "Comparison of four approaches to age and gender recognition for telephone applications", ***In proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)***, USA, (2007), 1089-1092.
6. Porat, R., Lange, D., Zigel, Y., "Age recognition based on speech signals using weights supervector", ***In proc. Interspeech, Japan, (2010)***, 2814-2817.

7. Bahari, M. H., Van hamme, H., "Speaker age estimation and gender detection based on supervised non-negative matrix factorization", ***In proc. of the IEEE Workshop on Biometric Measurements and Systems for Security and Medical Applications (BIOMS)***, (2008), 1-6.
8. Li, M., Han, K.J., Narayanan, S., "Automatic speaker age and gender recognition using acoustic and prosodic level information fusion", ***Comput. Speech Lang***., Vol. 27, (2013), 151-167.
9. Ajmera, J., Burkhardt, F., "Age and gender classification using modulation cepstrum", ***The Speaker and Language Recognition Workshop Stellenbosch***, South Africa, Speaker Odyssey, (2008).
10. Kumar, V. M., Thipesh, D.S.H, "Robot arm performing writing through speech recognition using dynamic time warping algorithm", ***International Journal of Engineering, Transactions B: Applications***, Vol. 30, No. 8, (2017), 1238-1245.
11. Loizou, P. C., Speech Enhancement: Theory and Practice, Taylor and Francis., 2007.
12. Kim, H., Park, H., "Non-negative matrix factorization based on alternating non-negativity constrained least squares and active set method", ***Technical report, Technical Report GT-CSE-07-01, College of Computing, Georgia Institute of Technology***, (2007).
13. Kim, H., Park, H., "Sparse non-negative matrix factorizations via alternating non-negativity constrained least squares for microarray data analysis", ***Bioinformatics***, Vol. 23, No. 12, (2007), 1495-1502.
14. Mavaddaty, S., Ahadi, S.M., Seyedin, S., "Modified coherence-based dictionary learning method for speech enhancement", ***Signal Processing, IET***, Vol. 9, No. 7, (2015), 1-9.
15. Barchiesi, D., Plumbley, M.D., "Learning incoherent dictionaries for sparse approximation using iterative projections and rotations", ***IEEE Transactions on Signal Processing***, Vol. 61, (2013), 2055-2065.
16. Sustik, M., Tropp, J., Dhillon, I., Heath, R., "On the existence of equiangular tight frames",***Linear Algebra and Its Applications***, Vol. 26, (2007), 619-635.
17. Mavaddaty, S., Ahadi, S.M., Seyedin, S., "A novel speech enhancement method by learnable sparse and low-rank decomposition and domain adaptation", ***Speech Communication***, Vol. 76, (2016), 42-60.
18. Mavaddati, S., Ahadi, S.M., Seyedin, S., "Speech enhancement based on voice activity detection using dictionary learning in wavelet packet transform domain," ***Computer Speech & Language***, (2017), Vol. 44, 22-47.
19. Shrawankar, U., Thakare, V. M., "Feature Extraction for a speech recognition system in a noisy environment:a study",***International Journal of Engineering Science and Technology (IJEST)***, Indonesia , (2010), Vol. 3, No. 2, 1764-1769.
20. https://www.phonetik.uni-muenchen.de/forschung/BITS/TP1 /Cookbook/node187.html.
21. Varga, A., Steeneken, H. J. M., Tomlinson, M., Jones, D., "The Noisex-92 study on the effect of additive noise on automatic speech recognition",***Technical Report. Malvern, U.K.: DRA Speech Res. Unit, (***1992).

Voice-based Age and Gender Recognition Based on Learning Generative Sparse Models

S. Mavaddati

Department of Engineering and Technology, University of Mazandaran, Babolsar, Iran

P A P E R I N F O

Paper history:
Received 10 March 2018
Received in revised form 16May 2018
Accepted 30May 2018

Keywords:
Gender Recognition
Sparse Non-negative Matrix Factorization
Incoherence
Mel-frequency Cepstral Coefficient
Voice Processing

چکیده

تشخیص جنسیت و سن گوینده به کمک مشخصات سیگنال گفتار، یکی از مسائل مهم در پردازش گفتار تلفنی به منظور تعیین هویت یک فرد می‌باشد. در این مقاله یک سیستم جدید تشخیص سن و جنسیت گوینده براساس مدل‌های ناهمدوس و جامع آموزش‌دیده به کمک الگوریتم تجزیه مقادیر غیرمنفی تنک و گام پس‌پرداش اصلاح اتم‌ها، ارائه شده است. مشابه سایر الگوریتم‌های طبقه‌بندی، روش پیشنهادی نیز شامل گام یادگیری و تست می‌باشد که در گام یادگیری، اتم‌های مربوط به هر کلاس داده آموزش داده شده و در گام تست، عملکرد طبقه‌بند مورد ارزیابی قرار می‌گیرد. از آنجاییکه دقت دسته‌بندی به میزان زیادی به ویژگی انتخابی از سیگنال داده وابسته است، در این مقاله از ضرایب کپسترال فرکانس مل به منظور یادگیری اتم‌های بازنمایی‌کننده پایه‌های فضایی و ساختار سیگنال گفتار استفاده شده است. این پایه‌ها بر روی مجموعه داده‌های گفتاری هر یک از گوینده‌ها با جنسیت‌های مختلف به کمک الگوریتم تجزیه مقادیر غیرمنفی تنک مدل می‌شوند. سپس اصلاح اتم‌ها مبتنی بر انرژی ضرایب فعال این تجزیه به منظور کاهش همدوسی میان گروه‌های داده‌ای مختلف بر روی واژه‌نامه معادل هر یک انجام می‌گیرد. در بازنمایی تنک هر کلاس داده، اتم‌های مربوط به دیگر کلاس‌ها که انرژی ضرایب فعال بیشتری بدست می‌دهند با اتم‌هایی که در بازنمایی این کلاس‌ها انرژی ضرایب فعال کمتری بدست می‌دهند، جایگزین می‌شوند و این جایگزینی به شرطی صورت می‌گیرد که خطای تقریب در بازنمایی داده اصلی از حد مشخصی بیشتر نشود. ارزیابی‌های مختلف در آزمایشات انجام شده نشان می‌دهد که دقت طبقه‌بندی الگوریتم پیشنهادی بیشتر از سایر الگوریتم‌های ارائه شده در این زمینه می‌باشد. این آزمایشات در حضور نویز سفید گوسی نیز انجام شده و نتایج مشابهی را بدست داده است.

***doi**: 10.5829/ije.2018.31.09c.08*