International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

A Review of Automatic Speaker Age Classification, Recognition and Identifying Speaker Emotion Using Voice Signal

Shivaji Chaudhari[1], Ramesh Kagalkar[2]

[1]Department of Computer Engineering, Pune University, Maharashtra, India
Dr. D Y School of Engineering and Technology, Lohegaon, Pune, Maharashtra, India

[2]Assistant Professor Department of Computer Engineering, Pune University, Maharashtra, India
Dr. D Y School of Engineering and Technology, Lohegaon, Pune, Maharashtra, India

Abstract: ***Accurate gender classification is mostly convenient in case of speech and speaker recognition and also in speech emotion classification; since a superior performance has been stated when separate acoustic models are employed for males and females. Gender classification is also specious into face recognition, particular video summarization, human or robot interaction (HCI), etc. In various criminal cases, an evidence either in the form of as phone conversations or in the form of as tape recordings. Thus, act of law enforcement agencies have been concerned which help the identification of a criminal about accurate approaches to profile dissimilar characteristics of a speaker from recorded patterns of voice. The importance of automatically recognizing expressed emotions from human speech has grown with the increasing role of spoken language interfaces in human-computer interaction (HCI) applications. This explores the detection of domain-specific emotions using language and discourse information in conjunction with acoustic correlates of emotion in speech signals. The main motivation is on a case study of detecting negative and non-negative emotions using spoken language data obtained from a call center application. Many previous surveys in emotion identification have used only the acoustic information contained in speech.***

Keywords: age estimation, gender detection, Gaussian Mixture Model (GMM), Hidden Markov Model (HMM), Mel Frequency Cepstral Coefficients (MFFCs), dimension reduction, speaker emotion recognition, weighted supervised non-negative matrix factorization.

1. Introduction

1.1. Speaker Age Classification and Recognition

Law enforcement agencies have been apprehensive about dissimilar biometric techniques to confirm the identity of an individual [2]. For forensic identification like fingerprint patterns, face characteristics, hand/palm geometry, signature dynamics and voice patterns different biometric characteristics can be used [2]. Choosing relevant a technique depends on its reliability in a specific application and the available information.

In certain criminal examples, existing evidence might be in the recorded conversations form. Thus speech patterns can consist of significant unique data for law enforcement persons [3]. For example, a person's speech pattern can make available information about his/her age, gender, idiom, emotive or mental state and involvement of a precise public or regional group. Hence, the speech can be used for speaker identification and recognition which is extremely demanded in numerous cases for example intimidating calls, kidnapping and wrong alarms [3].

In paper [4-5] research, author focus on speaker femininity detection and speaker age estimation. The perceptions of gender and age have an important shared impact on each other; these two fold characteristics are calculated together in several publications. Computerized speech age appraisal is problematic from different points of view. First, typically there be existent a difference between the speaker age as perceived, viz. the perceptual age, and the actual age of speaker, namely the consecutive age. Second, developing a robust age recognition system which requires a labeled, wide-ranging age and stable database. Third, the voice patterns are affected by several constraints, like weight, height and emotive condition, i.e. there is a significant intra-speaker inconsistency that is not correlated to or only related with age.

The various problem of age group recognition has been defined previously into [3-5]. For example, Bocklet and colleagues presented a technique which based on a GMM mean supervector as well as a Support Vector Machine (SVM) to classify speakers in seven age-gender classifications [3]. Authors in [3] used Mel Frequency Cepstral Coefficients (MFFCs) as features in their recognizer.

Though this system was attractive from some aspects, it demands working with very high dimensions if the amount of Gaussians in GMM be high. In [7], the GMM worldwide background model is combined with the SVM classifier and the issue of high dimensional supervectors is attacked by using Gaussian mixture weight supervectors. This technique has a worse dimension compared to mean or inconsistence supervectors. Zhang et al. conveyed age and gender recognition outcomes with the use of an unsupervised Non-negative Matrix Factorization (NMF) over Gaussian mixture weight supervectors in [8]. In their approach stated in paper, the acoustic features comprise Mel Spectra with mean normalization and Vocal Tract Length Normalization (VTLN) [9], increased with authors first and second order

time derivatives. But the drawback of this system is, although their technique could recognize the speaker's gender with high accuracy and high efficiency, but it is not that much positive for age estimation. In paper authors also determine that adding VTLN reduces the accuracy of gender detection but helps into age recognition.

1.2 Speaker Emotion Identification

There is an increasing requirement to be familiar with not only what information a user bears but as well how it is being conveyed. Various researches by psychologists and neuroscientists has exposed that emotion is strictly related to decision-making [10] and therefore, emotion plays an important role in the rational activities of human beings. The significance of emotion recognition from human speech has improved significantly with the requirement to progress both the naturalness and efficiency of spoken language human-machine interfaces [11]. Emotion recognition in spoken sentences not simply desires signal processing and analysis methods, on the other hand incorporates psychological and morphological examines of emotion. Whereas, generally, cognitive philosophy in psychology claims in contradiction of categorical labeling from only physiological human voice features [12-13], it delivers a practical starting point, specifically from a Voice processing of engineering perspective. The purpose for this is, the recognition of negative emotions can be used as a policy to increase the quality of the service in mechanical call center applications. Maximum preceding efforts involving emotion recognition from speech have been restricted to acoustic information [14, 15]. Discourse data of emotion recognition has been shared with acoustic correlates to advance the whole performance of emotion classification [16], [17].

2. Literature Review

In this section survey on literature is addressed in table 1.

Table 1: Literature survey

Author	*Paper*	*Description*	*Drawback*
M. Feld, F. Burkhardt, and C. Müller, 2010	Automatic speaker age and gender recognition in the car for tailoring dialog and mobile services	In this paper, M. Feld et al. addressed the first question how any one find out which group the present user belongs to. Author presented a Gaussian Mixture Model-GMM/SVM-supervector system for the speaker age and gender recognition, a system that is accepted from state-of-the-art speaker recognition examination.	In this problem is of high dimensional data. The High dimensional Data affect accuracy for age and gender recognition.
R. Porat, D. Lange, and Y. Zigel 2010	Age recognition based on speech signals using weights supervector	This paper proposes a new age-recognition system methodology - building a Gaussian mixture model which based weights supervector features for support vector machine (SVM). This methodology uses the hypothesis that it is probable to find unique Gaussians for every age-group model in the universal background model (UBM). The weights of those Gaussians can prime to a discriminant mode to separate the age sets.	In this only Age estimation is done instead of age group recognition. Using supervectors of GMM means and variances and combining these features
X. Zhang, K. Demuynck, and H. Van hamme, 2011	Rapid Speaker Adaptation in Latent Speaker Space with Non-negative Matrix Factorization	This paper described a novel model space wild speaker adaptation system that modifies the Gaussian mixture weights. The targeted speaker weights are conveyed as a linear combination of latent speaker vectors. The latent speaker vectors encrypt systematic patterns of distinction in Gaussian usage in between speakers. The vectors are learned by means of NMF on statistics composed for all speakers that made up the training data.	Authors intend to apply hierarchical weight decomposition as to adjust the degrees of liberty in the NMF-adaptation to the amount of accessible adaptation data. And Latent vector quality should be enriched
T. Bocklet, A. Maier, J. G. Bauer, F. Burkhardt, and E. Noth, 2008	Age and gender recognition for telephone applications based on GMM supervectors and support vector machines	This paper compares two different approaches of automatic age and gender classification with seven classes. The first approach is Gaussian Mixture Models (GMMs) with Universal Background Models (UBMs), which is well known for the job of speaker identification/verification. The training performed by the Expectation Maximization (EM) algorithm or MAP adaptation correspondingly. The second approach for each speaker of the testing set and training set a GMM model is trained.	In this only spectral features are investigated
L. .T Bosch, J. Driesen, H. Van hamme, and L. Boves, 2009	On a computational model for language acquisition: modeling cross-speaker generalization	This paper describes experimental outcomes achieved by a computational model that simulates these two procedures. This model is capable to form word like illustrations on the basis of multimodal input information (stimuli) without the support of an a priori stated lexicon. In this paper author investigate how internal illustrations generalize across speakers.	There was not been able to describe a reliable distance quantity for the fit between a word and the internal illustrations.
O. Rasanen, and J. Driesen, 2009	A comparison and combination of segmental and fixed-frame signal representations in NMF-	Segmental and fixed-frame signal demonstrations were compared in diverse noise situations in a weakly supervised word recognition job using a non-negative matrix factorization (NMF) framework. From this	Drawback occurred here, information obtain from larger temporal scales appears to become more and more vital as

	based word recognition	combination it is shown that a both fixed-frame and segmental representations vintages the best recognition rates in diverse noise conditions.	the signal-to-noise ratio becomes worse.

3. Dimension Reduction Approaches

3.1. Principal Components Analysis

Principal components analysis (i.e. PCA) [36] is defined as an orthogonal linear transformation which projects a set of vectors to newly basis whose constituents are linearly uncorrelated and arrangement of constituents is in a decreasing order of variance. PCA methodology is assumed to be most of the relevant data is search in the first coordinates of the projected space, as stated above it contain most of the variance.

3.2. Supervised PCA

Supervised PCA (SPCA) is a PCA variant where the feature vectors are preprocessed before applying preprocessing technique PCA on them. The preprocessing on feature vector contains screening out lowest correlation labeled value of coordinates. The feature vector consist the correlation coefficient in between each and every coordinate of the feature vectors and the labels. This technique is usually used in regression issues for preprocessing [37], in which the label is continuous.

3.3. Anchor Modeling

Anchor modeling is dimension reduction technique which usually used for speaker verification [38] to project a given session in a low-dimensional scores space. This anchor modeling technique uses anchor models trained on a predefined set of speech sessions.

In [39] it was shown that using standardized GMM supervector, the log-likelihood standards achieved by the anchor models. The anchor-supervectors requirement to be diversified and characterize speakers from all class labels to ensure minimal information loss in the projected space.

3.4. Weighted Pairwise PCA

In PCA a dimension reduction is achieved that preserves most of the vectors variance without bearing in mind the class labels. There is no assurance that the idea of maximum variance will offer good features for discrimination.

The Nuisance Attribute Projection (NAP) projection framework addressed in [40] was found useful to disregard inter-session speaker inconsistency for speaker verification. The aim of applying this method is its ability to disregard the unwanted variability common to speakers of the similar age.

4. Classification Techniques

Speaker classification can be understood as speaker identification within which each class nothing but a speaker. The gender classification task can be assumed as identifying whether a test sound is from which speaker i.e. a male or female speaker. An automatic speaker classification technique consists of two phases: training phase and testing phase. In the training phase, the data used for training which is of the digital input signal of voice is administered and also feature vectors are extracted. And thereafter these feature vectors of altogether classes are used to train the speaker class models of a classifier. In the test phase, the input voice signal feature vectors are yet again extracted. Then feature vectors are scored in the classifier to every model and classified in the model assumed the best score.

4.1. SVM technique

Models to identify how frequently speakers use particular words or phrases (idiolect) have been proposed by [18] and, though these poor models described, were set up to increase speaker recognition techniques in the combination with other knowledge resources. [19] Employed a version regarding n-gram based features in Support Vector Machines (SVM), which authors reported gave better output than language-model-based approaches. A number of iterative training algorithms have been proposed [20], [21] in the purpose of to solve the quadratic optimization included in the SVM training, but algorithm complexity is depend on the nature of data used and also the resulting number of support vectors.

4.2. GMM Technique

In [7], the GMM universal background model is merged with the SVM classifier and thus the problem of high dimensional supervectors is tackled by using Gaussian mixture weight supervectors, which have a lower dimension as compared to mean or variance supervectors.

In both commercial and research systems, GMMs have become the dominant approach. This approach has been used to generate partition of spectral information from short time frames of speech. It can reflect information about a speaker's vocal physiology, and is text-independent because it does not rely on phonetic content [22]. GMMs were effectively used for robust text-independent speaker identification and verification [23, 24].

4.3. HMM Technique

Dimensionality reduction of the acoustic features for de-correlation or enhancement is not a new concept. There are various mechanisms found in the literature that perform this task, together with DCT, Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), Heteroscedastic LDA (HLDA), [25, 26, 27]. The main purpose for this process able to model the features via diagonal covariance matrix GMM/HMM for either speech or speaker recognition. These speech/speaker detection techniques can be classified mainly into two groups by their way of operation: 1) Signal processing domain, and 2) Model domain. The very common mechanism is the DCT application for the log-filter bank energies [28] promoted by

the MFCC illustration. Speaker classification is nothing but speaker identification in which every class is a speaker class. For example, the gender classification task can be identifying by checking whether a test utterance is from a male or female speaker. As stated before an automatic speaker classification technique includes both training phase and testing phase. The training data of the various digital input signal of speaker voice is processed and the feature vectors of each digital input signal are extracted and used to train the speaker class models of a classifier. Also test data input voice signal feature vectors are extracted. Then these testing feature vectors are scored in the classifier to each training model and classified into the best score of a given model.

5. Matrix Factorization

5.1. WSNMF

Non-negative Matrix Factorization (NMF) is a prevalent machine learning algorithm [29], which is effectively applied to word recognition in [30], separation of sound source in [31] and various spam filtering [32]. Different postponements of NMF such as a Supervised Non-negative Matrix Factorization (SNMF) addressed in [33] and Weighted Non-negative Matrix Factorization (WNMF) [34] have been developed to resolve real world issues, during the last few years. In paper [1], the idea is to merge WNMF with SNMF and results in WSNMF i.e. Weighted Supervised Non-negative Matrix Factorization to highlight on non-negative matrix in the factorization process. In [34], N. Ho introduced a technique called weighted NMF to adjust the value of district elements of non-negative Matrix.

5.2. GRNN

A General Regression Neural Network (GRNN) is defined as a universal function approximator that introduced in [35]. A GRNN has various advantages over other neural networks (NNs) which are pointed out as follows:

- A GRNN does not need to use iterative learning algorithms. As an alternative, GRNN has a one pass and fast learning. The standard supervised neural network architectures such that multilayer perceptrons and radial basis functions infer a parameterized model from the existing training data. These neural networks use the back-propagation algorithm for training, which may take a high number of iterations to converge, however global convergence cannot be guaranteed.
- A GRNN have need of only a fraction of the training samples which a back propagation based neural network would need. In other disputes, a GRNN can be successfully applied in the case of sparse data.

6. Conclusion

In this paper there represent reviews of Automatic Speaker Age Classification, Recognition and Identifying Speaker Emotion based on gender Using Voice Signal. The techniques implemented on different kinds of systems, an age-group classifier and a precise age estimator by regression. The results taken on multiple dataset's of speech and different languages and different and the number of voice files for greater accuracy and efficiency of system performance.

References

[1] Gil Dobry, Ron M. Hecht, Mireille Avigal, and Yaniv Zigel, "Supervector Dimension Reduction for Efficient Speaker Age Estimation Based on the Acoustic Speech Signal", Member of IEEE, VOL. 19, NO. 7, SEPTEMBER 2011.
[2] K. Jain, P. Flynn, and A. A. Ross, Handbook of biometrics. Springer, 2008.
[3] D. C. Tanner, and M. E. Tanner, Forensic aspects of speech patterns: voice prints, speaker profiling, lie and intoxication detection. Lawyers & Judges Publishing, 2004.
[4] T. Bocklet, A. Maier, J. G. Bauer, F. Burkhardt, and E. Noth, "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, pp. 1605– 1608, 2008.
[5] F. Metze, et al. "Comparison of four approaches to age and gender recognition for telephone applications," In proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP), USA, pp. 1089–1092, 2007.
[6] M. Feld, F. Burkhardt, and C. Müller, "Automatic speaker age and gender recognition in the car for tailoring dialog and mobile services," In proc. Interspeech, Japan, pp. 2834-2837, 2010.
[7] R. Porat, D. Lange, and Y. Zigel, "Age recognition based on speech signals using weights supervector," In proc. Interspeech, Japan, pp. 2814-2817, 2010.
[8] X. Zhang, K. Demuynck, and H. Van hamme, "Rapid speaker adaptation with speaker adaptive training and non-negative matrix factorization," In proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP), Czech republic, pp. 4456-4459, 2011.
[9] L. .T Bosch, J. Driesen, H. Van hamme, and L. Boves, "On a computational model for language acquisition: modeling cross-speaker generalization," In Proc. Int. Conf. Text, Speech and Dialogue, Czech Republic, pp. 315-322, 2009.
[10] Damasio, Descartes' Error: Emotion, Reason, and the Hitman Brain. London, U.K.: Putman, 1994.
[11] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, Fellenz, and J. Taylor, "Emotion recognition in human-computer interaction," IEEE Signal Process. Mag., vol. 18, no. 1, pp. 32–80, Jan. 2001.
[12] Ortony, G. Clore, and A. Collins, The Cognitive Structure of Emotions. Cambridge, U.K.: Cambridge Univ. Press, 1988.
[13] K. Scherer, "Toward a concept of 'modal emotions'," The Nature of Emotion: Fundamental Questions, pp. 25–31, 1994.
[14] Lee, S. Narayanan, and R. Pieraccini, "Recognition of negative emotions from the speech signal," in Proc. Automatic Speech Recognition Understanding, Dec. 2001.

[15] Roy and A. Pentland, "Automatic spoken affect analysis and classification," in Proc. Int. Conf. Automatic Face Gesture Recognition, Killington, VT, 1996, pp. 363–367.
[16] Batliner, K. Fischer, R. Huber, J. Spiker, and E. Noth, "Desperately seeking emotions: Actors, wizards, and human beings," in Proc. ISCA Workshop Speech Emotion, 2000, pp. 195–200.
[17] J. Ang, R. Dhillon, A. Krupski, E. Shriberg, and A. Stolcke, "Prosodybased automatic detection of annoyance and frustration in human-computer dialog," in Proc, ICSLP, Denver, CO, Sep. 2002, pp. 2037–2040.
[18] G. Doddington, "Speaker recognition based on idiolectal differences between speakers", in P. Dalsgaard, B. Lindberg, H. Benner, and Z. Tan, editors, Proc. EUROSPEECH, pp. 2521–2524, Aalborg, Denmark, Sep. 2001.
[19] Shriberg, L. Ferrer, S. Kajarekar, N. Scheffer, A. Stolcke, and M. Akbacak, "Detecting nonnative speech using speaker recognition approaches", in Proceedings IEEE Odyssey-08 Speaker and Language Recognition Workshop, Stellenbosch, South Africa, Jan. 2008.
[20] L. Kaufman, "Solving the quadratic programming problem arising in support vector classification," in Advances in Kernel Methods—Support Vector Learning, B. Scholkopf, C. J. C. Burges, and A. J. Smola, Eds. Cambridge, MA: MIT Press, 1998.
[21] Rong-En, C. Pai-Hsuen, and L. Chih-Jen, "Working set selection using second order information for training support vector machines," J. Mach. Learn. Res., vol. 6, pp. 1889–1918, 2005.
[22] Shriberg. Higher-level features in speaker recognition. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 4343 LNAI:241–259, 2007.
[23] D. Reynolds. Speaker identification and verification using Gaussian mixture speaker models. Speech Commun., 17(1-2):91–108, 1995.
[24] D. Reynolds and R. Rose. Robust text-independent speaker identification using Gaussian mixture speaker models. IEEE Transactions on Speech and Audio Processing, 3(1):72–83, 1995.
[25] L. Burget, P. Matejka, P. Schwarz, O. Glembek, and J. Cernocky, "Analysis of feature extraction and channel compensation in a GMM speaker recognition system," IEEE Trans. Audio, Speech, Lang. Process., vol. ASSP-15, no. 7, pp. 1979–1986, Sep. 2007.
[26] E. Batlle, C. Nadeu, and J. Fonollosa, "Feature decorrelation methods in speech recognition. A comparative study," in Proc. ICSLP, Sydney, Australia, 1998, vol. 7, pp. 2907–2910.
[27] T. Eisele, R. Haeb-Umbach, and D. Langmann, "A comparative study of linear feature transformation techniques for automatic speech recognition," in Proc. ICSLP, 1996, vol. 1, pp. 252–255.
[28] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," IEEE Trans. Acoust., Speech, Signal Process., vol. ASSP-28, no. 4, pp. 357–366, Aug. 1980.
[29] Lee, D. D., and H. S. Seung, "Algorithms for non-negative matrix factorization," Advances in neural information processing systems, pp. 556–562, 2001.
[30] O. Rasanen, and J. Driesen, "A comparison and combination of segmental and fixed-frame signal representations in NMF-based word recognition," Nordic Conf. Computational Linguistics, NEALT Proceedings Series, vol. 4, pp. 255-262, 2009.
[31] C. Yang, M. Ye, and J. Zhao, "Document clustering based on nonnegative sparse matrix factorization," In Advances in Natural Computation, Lecture Notes in Computer Science, vol. 3611, pp.557–563, 2005.
[32] T. Virtanen, "Monaural sound source separation by non-negative matrix factorization with temporal continuity and sparseness criteria," IEEE Trans. Audio, Speech, and Language Processing, vol. 15, no.3, pp. 1066-1074, 2007.
[33] Van hamme, "HAC-models: a novel approach to continuous speech recognition," In proc. Interspeech, Australia, pp. 2554-2557, 2008.
[34] N. Ho, "Nonnegative matrix factorization algorithms and applications," PhD thesis, Université. Catholique de Louvain, 2008
[35] D. F. Specht, "A general regression neural network," IEEE Trans. Neural Networks, vol. 2, no. 6, pp. 568- 576, 1991.
[36] S. Lindsay, "A tutorial on principal components analysis introduction," Statistics, 2002.
[37] E. Bair, T. Hastie, D. Paul, and R. Tibshirani, "Prediction by supervised principal components," J. Amer. Statist. Assoc., pp. 119–137, 2006.
[38] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted Gaussian mixture models," Digital Signal Process., vol. 10, pp. 19–41, 2000.
[39] E. Noor and H. Aronowitz, "Efficient language identification using anchor models and support vector machines," IEEE Odyssey, pp. 1–6, 2006.
[40] Solomonoff, W. Campbell, and C. Quillen, "Channel compensationfor SVM speaker recognition," in Proc. Odyssey, 2004, pp. 57–62.

Author Profile

Shivaji J Chaudhari Research Scholar Dr. D.Y.Patil School of Engineering and Technology, Charoli, B.K.Via –Lohegaon, Pune, Maharashtra, India. University of Pune. He received B.E. in Information Technology from SVPM COE Malegaon, Baramati, Pune University. Currently He is pursuing M.E. in Computer Network from Dr. D. Y. Patil School of Engineering & Technology, Pune, University of Pune.

Prof Ramesh. M. Kagalkar was born on Jun 1st, 1979 in Karnataka, India and presently working as a Assistant. Professor, Department of Computer Engineering, Dr. D.Y.Patil School of Engineering and Technology, Charoli, B.K.Via –Lohegaon, Pune, Maharashtra, India. He is a Research Scholar in Visveswaraiah Technological University, Belgaum, He had obtained M.Tech (CSE) Degree in 2005 from VTU Belgaum and He received BE (CSE) Degree in 2001 from Gulbarga University, Gulbarga. He is the author of text book Advance Computer Architecture which cover the syllabus of Visveswaraiah Technological University, Belgaum. He has published many research paper in International and international conference.