ROBUST FEATURES FOR AUTOMATIC ESTIMATION OF PHYSICAL PARAMETERS FROM SPEECH

Kalluri Shareef Babu, Deepu Vijayasenan

Department of E & C Engineering
National Institute of Technology Karnataka - Surathkal
Srinivasnagar, Mangalore, KA, India - 575025

ABSTRACT

Estimating speaker's physical parameters like height, weight and shoulder size can assist in voice forensics by providing additional knowledge about the speaker. In this work, statistics of the components of background GMM are employed as features in estimating the physical parameters.These features improved the performance of height and shoulder size estimation as compared to our earlier attempt based on a Bag of Word representation. The robustness of the features is validated using two different training subsets containing different languages.

Index Terms— Physical parameters, Speech forensics, height, weight, shoulder size, MFCC, first order statistics, GMM-UBM, SVR.

I. INTRODUCTION

Human speech contains information about the textual message as well as various characteristics of a speaker such as accent (the region where speaker belongs to), age (child, adult, late adult), gender identity (male/female), emotions (anger, happiness, sadness etc.). The speech also has a major part in investigating physical parameters of the speaker. In perspective of forensic analysis using speech data, the most studied physical parameter is height.

Speech processing researchers have reported that locations of formant frequencies are decreasing with an increment in length of the vocal tract of a person [1]. This study has shown that the length of vocal tract directly affects the structure of speech. Height and vocal tract length were found to be strongly correlated (0.855 for men and 0.832 for women) [2].

Previous Work: Different feature extraction methods have been proposed by researchers. Some of the well-known ones include the statistics of features extracted using *OpenSmile* toolkit [3], [13], GMMs models [4].

Recently, a novel method for predicting the height of a speaker using estimated sub glottal resonance frequencies and attained $6.2cm$ RMSE [5],. An alternate approach fuses short and long temporal windows to achieve an RMSE of $6.7cm$ for male and $6.1cm$ for female [6]. These papers report the state of the art results in height estimation. Researchers have attempted to predict other physical characteristics such as age and weight from speech signal. Weight is predicted using similar methods of approach and obtained a correlation coefficient ~ 0.52 [7]. The classification of speakers into different age groups were attempted by [8], [9] using GMM and SVM.

Authors in their previous work have explored prediction of shoulder size and waist size in addition to height and weight. This used Short-Time Fourier Transform (STFT) features along with its static and dynamic components. These STFT features are trained by using 512 components GMM-UBM and are represented using Bag of Words (BoW). These BoW represented features are used for support vector regression training and predicted the physical parameters like height, shoulder width, waist size and weight and achieved RMSE of $6.6cm$, $2.6cm$, $7.1cm$ and $8.9kg$ respectively [10] with the collected data.

Overview : The key aspects of our work are:

1) We try to improve our Bag of Word approach using first order statistics of trained background GMM-UBM.
2) The robustness of the proposed features with respect to spoken language are analyzed.

The rest of the paper is organized as follows, Section II discuss dataset used for this work. In Section III the baseline approach is described. Thereafter feature extraction method in Section IV. Section V explains about the experiments and results. Finally, we summarize our contributions and conclusions in Section VI.

II. DATA COLLECTION

For this study we are using the same dataset used as in [10]. This dataset is having the details of the physical characteristics such as height, shoulder width, waist size and weight. It has speech samples collected from 207 speakers (includes 161 male and 46 female speakers) fall in 18 – 35 years age group. We call this dataset as AFDS (Audio Forensics Dataset). The recordings were at 16 kHz sampling

978-1-5090-1134-6/17/$31.00 ©2017 IEEE

rate. In this dataset, each speaker has contributed roughly for around 2 minutes of data spanning across two or three sessions. Each session is around 40 seconds. Recordings consists of sentences read from Indian newspapers. Users have read text in their mother tongue as well as in English. The mother tongue could differ between speakers. There are 12 different mother tongues. height, shoulder width are measured in centimeters (cm) and weight is measured in kilograms (Kg).

III. BASELINE APPROACH

The authors used Short Time Fourier Transform features for AFDS dataset. In this work, a set of representative cluster centroid is obtained using k-means algorithm. Each speech utterance is represented using a Bag of Words representation using these cluster centers. Support Vector Regression is employed to predict the physical parameters. Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) were reported.

Support Vector Regression:
Different regression models like linear and non-linear models have been experimented with in the context of physical parameter prediction [5], [11], [3]. In this work, we use support vector regression (SVR) [12], as the model for predicting the target physical parameter value for a given set of input features. We denote the set of input features as $\{y_1, y_2, ..., y_m\}$ and the respective target physical parameter values are $f(\mathbf{y}) = \mathbf{w}^T\ \mathbf{y} + b$. The linear SVR tries to learn a mapping which perform the following optimization:

$$\begin{aligned} min\ \frac{1}{2}\mathbf{w}^T\mathbf{w}\ \ subject\ \ to \\ |\mathbf{w}^T\mathbf{y}_i + b - t_i| < \epsilon \end{aligned} \tag{1}$$

The parameter ϵ controls the "fit"of the function. The optimization aims to find a solution that deviates from the target value by at most ϵ provided such a function exists.

IV. FEATURE EXTRACTION

In the past, several features had been explored for height estimation such as statistics of pitch and short term spectrum [13], [3], sub-glottal resonances [14] formants [15], STFT features [10].

The short-term mel spectrogram captures the gross level spectral characteristics. However, the short term spectrum is affected by spoken phoneme characteristics. In order to normalize the linguistic effect, we adopt a framework similar to the super-vector approaches in speaker verification [16]. The super-vectors are the concatenation of first order mean statistics of each mixture component of a Gaussian Mixture Model - Universal Background Model (GMM-UBM).

For this purpose, a Universal background model (GMM-UBM) is trained on short term mel frequency cepstral coefficients. We use 20 MFCCs estimated from $25ms$ windows along with frame level log energy. The velocity and acceleration coefficients are spliced with the MFCC coefficients yielding 60 dimensional features [17]. A Gaussian Mixture Model with diagonal covariance features is learned using training data. The GMM density function is given by,

$$f_{UBM}(\mathbf{x}) = \sum_{k=1}^{M} w_k \mathcal{N}(\mathbf{x}, \mu_k, C_k) \tag{2}$$

where $\mathbf{x}$ denotes the random vector. μ_k and C_k denote the mean vector and diagonal covariance matrix of the k^{th} GMM component with weight w_k. Given the sequence of feature vectors $\{x_1, x_2, ..., x_T\}$, the first order statistics are computed as:

$$\hat{\mu}_j = \frac{\sum_n \mathbf{x}_n p(j|\mathbf{x}_n)}{\sum_n p(j|\mathbf{x}_n)} \tag{3}$$

where *a-posterior* probabilities are computed as,

$$p(j|\mathbf{x}_n) = \frac{\mathbf{w}_j \mathcal{N}(\mathbf{x}_n, \mu_j, C_j)}{\sum_{k=1}^{M} \mathbf{w}_k \mathcal{N}(\mathbf{x}_n, \mu_k, C_k)} \tag{4}$$

The speech utterance is then represented by concatenating all the $\hat{\mu}_j$ for all GMM components. Intuitively, if each GMM component j represent a different sound class, then each of the $\hat{\mu}_j$ would account for the mean of features that belong to that sound class. We finally applied a dimensionality reduction using Principal Component Analysis (PCA) to reduce the computation time.

V. EXPERIMENTS & RESULTS

Data. We used the same dataset as described in Section II. The AFDS dataset is divided into training dataset containing 137 speakers (includes 104 male + 33 female) and test data containing 70 speakers (includes 57 male + 13 female). For training data we have 951 utterances and for testing 538 utterances. There is no overlap of speakers in training and testing samples of recorded utterances. The height values range from 147cm to 188cm, shoulder width from 30cm to 53cm, and weight from 39kg to 107kg. The algorithms are benchmarked using either Mean Absolute Error or Root Mean Square Error.

$$\begin{aligned} MAE &= \frac{1}{N}\sum_i |x_{tar,i} - x_{pred,i}| \\ RMSE &= \sqrt{\frac{1}{N}\sum_i (x_{tar,i} - x_{pred,i})^2} \end{aligned} \tag{5}$$

where x_{tar} and x_{pred} are the target and predicted values of each utterance i respectively.

V-A. Baseline Results

The authors computed log-magnitude STFT and its dynamic components using a window length $25ms$ and frame shift is $10ms$. 512 clusters centroids are obtained from k-means algorithm. BoW feature were extracted. Support Vector Regression with a normalized polynomial kernel

(for details refer [10]) is employed for physical parameter prediction. Table I lists the baseline results along with the population mean (Predicting the train mean for every test sample). This approach is better in terms of both MAE and RMSE as compared to the population means (i.e. by blindly predicting the mean of training data without looking at the speech samples).

Table I. Baseline results with Support Vector Regression (SVR) using Bag of Words (BoW) features.

Physical	Population mean		BoW + SVR	
parameter	MAE	RMSE	MAE	RMSE
Height(cm)	6.81	8.22	5.20	6.58
Shoulder width (cm)	2.76	3.43	2.12	2.57
Weight(Kg)	8.29	10.57	6.72	8.91

V-B. First order Statistics

We perform speech activity detection [18] before feature extraction for every speech sample. 20 Mel frequency Cepstral features along with its dynamic components are extracted. The feature dimension is 60 with δ & $\delta\delta$ features. A 256 component diagonal covariance GMM-UBM is learned from the training data. The first order statistics are computed according to equation 3 for each of the 256 mixture components. Thus, the first order statistics has a dimension of $60 \times 256 = 15360$. As this feature dimension is high we perform dimensionality reduction using principal component analysis (PCA). We reduce to a smaller dimension of size K. A Support Vector Regression is trained on the dimensionality reduced features. Figure 1 shows the variation of MAE for various dimensions. The algorithm was found to be robust to value of K. We choose $K = 256$ for our further experiments. Table II reports the results for the same. The MAE is $5.1cm$ for height, $2cm$ for shoulder size and $6.9kg$ for weight. There is an improvement of RMSE $0.21cm$ in height and $0.1cm$ in shoulder size when compared with our baseline results.

V-C. Effect of UBM on Language

Here, we study the robustness of the system to spoken language. The system is separately trained and evaluated with two different subsets of the data – native language utterances and English utterances. Both the GMM-UBM and SVR are trained using one of the subsets at a time. The system is then evaluated using the matched as well as mismatched subset. We perform the following experiments.

1) First, the system is trained using English utterances only, and is evaluated separately on the English utterances (matched condition) and the native languages (mis-matched condition). The MAE of each physical parameter is shown in Fig. 2.

Table II. MAE, RMSE and Correlation of first order statistics of MFCC $+\delta + \delta\delta$ with 256 dimension.

Physical Characteristic	MAE	RMSE	Correlation
Height (cm)	5.10	6.37	0.63
Shoulder width (cm)	2.00	2.47	0.70
Weight (Kg)	6.85	9.00	0.53

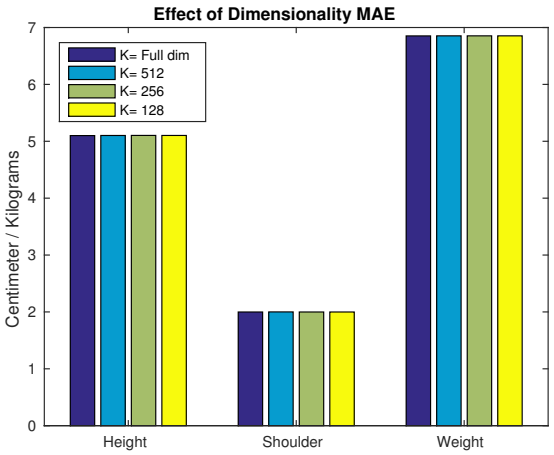

Fig. 1. MAE of each physical parameter with different dimensionality reduction on first order statistics.

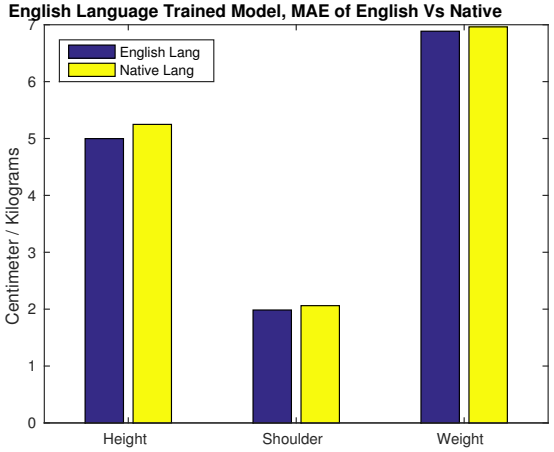

Fig. 2. MAE of each physical parameter with SVR & GMM-UBM trained on English samples and tested on both Native and English samples.

Even though the model is trained on English utterances and tested on Non-English (native) languages, the effect of language in predicting the physical parameters error is minimal. There is only 5.2% change in height, 4% change in shoulder and 1.2% change in weight MAE.

2) Next, the system is trained on multiple native languages and tested separately on both English (mismatched) and native language subsets. MAE corresponding to each physical parameter is shown in Fig. 3.

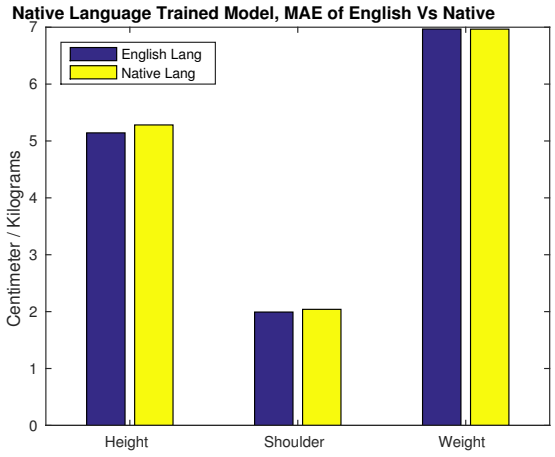

Fig. 3. MAE of each physical parameter with SVR & GMM-UBM trained on Native language samples and tested on both Native and English samples.

Similar to the above case, when the model is trained on native languages and tested on native languages and English there is minimal change in the predicting error. The MAE change in between native and English language while predicting height is 2.65%, 2.45% in shoulder size and no change in weight.

Table III summarizes the results of these experiments with matched and mismatched conditions of training models.

Table III. MAE comparison of matched and mismatched conditions of trained model using English Language (E L) and Native Languages (N L).

Physical Parameter	E L Train		N L Train	
	E L Test	NL Test	E L Test	N L Test
Height (cm)	4.99	5.25	5.14	5.28
Shoulder width (cm)	1.98	2.06	1.99	2.04
Weight (Kg)	6.88	6.96	6.97	6.97

The matched condition results in very similar performance as earlier (Table II), eventhough the training data has been reduced to almost half. Also it is interesting to note that the percentage degradation is more when the training subset contains only one language (English). The degradation is less while multiple languages (12 Native languages) are used in training even though the tested language (English) is unseen in the training data.

VI. SUMMARY AND CONCLUSION

In this work we explored first order statistics of a GMM-UBM as features for physical parameter estimation. The same set of features are used to predict multiple physical parameters of the speaker. A dimensionality reduction is performed as the input feature dimension is high. These set of features improved our earlier results using BoW based features for height and shoulder size estimation by $0.23cm$ and $0.13cm$ respectively in RMSE.

We also studied the effect of language on the system. While only English is used for training the system there is $\sim 5.2\%$ performance degradation in the mismatched condition. While training with native languages the performance degradation is even smaller $\sim 2.5\%$.

In future the authors would like to explore robust features like i-vectors [19] that could potentially address channel variabilities in predicting the physical parameters.

VII. REFERENCES

[1] David RR Smith, Roy D Patterson, Richard Turner, Hideki Kawahara, and Toshio Irino, "The processing and perception of size information in speech sounds," *The Journal of the Acoustical Society of America*, vol. 117, no. 1, pp. 305–318, 2005.

[2] W Tecumseh Fitch and Jay Giedd, "Morphology and development of the human vocal tract: A study using magnetic resonance imaging," *The Journal of the Acoustical Society of America*, vol. 106, no. 3, pp. 1511–1522, 1999.

[3] Iosif Mporas Todor Ganchev and Nikos Fakotakis, "Automatic height estimation from speech in real-world setup," in *Proceedings of 18th European Signal Processing Conference (EUSIPCO)*, 2010.

[4] Bryan L Pellom and John HL Hansen, "Voice analysis in adverse conditions: the centennial olympic park bombing 911 call," in *Circuits and Systems, 1997. Proceedings of the 40th Midwest Symposium on*. IEEE, 1997, vol. 2, pp. 873–876.

[5] Harish Arsikere, Steven M Lulich, and Abeer Alwan, "Estimating speaker height and subglottal resonances using mfccs and gmms," *Signal Processing Letters, IEEE*, vol. 21, no. 2, pp. 159–162, 2014.

[6] Rita Singh, Bhiksha Raj, and James Baker, "Short-term analysis for estimating physical parameters of speakers," in *Biometrics and Forensics (IWBF), 2016 4th International Workshop on*. IEEE, 2016, pp. 1–6.

[7] Amir Hossein Poorjam, Mohamad Hasan Bahari, et al., "Multitask speaker profiling for estimating age, height, weight and smoking habits from spontaneous telephone speech signals," in *Computer and Knowledge Engineering (ICCKE), 2014 4th International eConference on*. IEEE, 2014, pp. 7–12.

[8] Jitendra Ajmera and Felix Burkhardt, "Age and gender classification using modulation cepstrum.," in *Odyssey*, 2008, p. 25.

[9] Ming Li, Chi-Sang Jung, and Kyu Jeong Han, "Combining five acoustic level modeling methods for automatic speaker age and gender recognition.," in *INTERSPEECH*, 2010, pp. 2826–2829.

[10] Shareef Babu Kalluri, Ashwin Vijayakumar, Deepu Vijayasenan, and Rita Singh, "Estimating multiple physical parameters from speech data," in *Machine Learning for Signal Processing (MLSP), 2016 IEEE 26th International Workshop on*. IEEE, 2016, pp. 1–5.

[11] Sorin Dusan, "Estimation of speaker's height and vocal tract length from speech signal," in *Ninth European Conference on Speech Communication and Technology*, 2005.

[12] Alex J Smola and Bernhard Schölkopf, "A tutorial on support vector regression," *Statistics and computing*, vol. 14, no. 3, pp. 199–222, 2004.

[13] Todor Ganchev, Iosif Mporas, and Nikos Fakotakis, "Audio features selection for automatic height estimation from speech," in *Artificial Intelligence: Theories, Models and Applications*, pp. 81–90. Springer, 2010.

[14] Harish Arsikere, Steven M Lulich, and Abeer Alwan, "Estimating speaker height and subglottal resonances using mfccs and gmms," *Signal Processing Letters, IEEE*, vol. 21, no. 2, pp. 159–162, 2014.

[15] John HL Hansen, Keri Williams, and Hynek Bořil, "Speaker height estimation from speech: Fusing spectral regression and statistical acoustic models," *The Journal of the Acoustical Society of America*, vol. 138, no. 2, pp. 1052–1067, 2015.

[16] Douglas A Reynolds, "An overview of automatic speaker recognition technology," in *Acoustics, speech, and signal processing (ICASSP), 2002 IEEE international conference on*. IEEE, 2002, vol. 4, pp. IV–4072.

[17] Xinhui Zhou, Daniel Garcia-Romero, Ramani Duraiswami, Carol Espy-Wilson, and Shihab Shamma, "Linear versus mel frequency cepstral coefficients for speaker recognition," in *Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on*. IEEE, 2011, pp. 559–564.

[18] Zheng-Hua Tan and Børge Lindberg, "Low-complexity variable frame rate analysis for speech recognition and voice activity detection," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 5, pp. 798–807, 2010.

[19] Najim Dehak, Patrick J Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.