2016 22nd Annual International Conference on Advanced Computing and Communication

# STUDY OF WIRELESS CHANNEL EFFECTS ON AUDIO FORENSICS

Deepu Vijayasenan, Shareef Babu Kalluri, Sreekanth k, Ansal Issac
Department of Electronics and Communication Engineering
National Institute of Technology Karnataka
Surathkal, Mangalore
India-575025

*Abstract*— **In this work, we try to study the effect of a wireless channel on physical parameter prediction based on speech data. Speech data from 207 speakers along with corresponding speaker's height and weight is collected. A three path Rayleigh fading channel with typical values of Doppler shift, path gain and path delay is utilized to create the mobile channel output audio. A Bag of Words (BoW) representation based on log magnitude spectrum is used as features. Support Vector Regression (SVR) predicts the physical parameter of the speaker from the BoW representation. The proposed system is able to achieve a Root Mean Square Error (RMSE) of 6.6 cm for height estimation and 8.9 Kg for weight estimation for clean speech. The effect of Rayleigh channel increase the RMSE values to 8.17 cm and 11.84 Kg respectively for height and weight.**

**Keywords— Physical parameter, Speech forensics, Harmonic model, Height, Weight.**

## I. INTRODUCTION

Over the last decades, many research work happened in the field of person authentication using biometric characteristics, such as face recognition, iris and retina recognition, fingerprint recognition, DNA analysis and recognition of various speech-related biometric characteristics etc. In particular, speech based person authentication offers advantages over the other types of biometric processes in terms of intrusiveness, cost and easy of deployment. Speech data contains both textual message and speaker information. Many works related to speaker authentication using speech was reported, which utilized the dependency of speech signal with the speech producing system.

The primary source of human voice is from the vibrations of vocal cords, which are the main source of sound production mechanism in the human body. Morphological studies show that human height is correlated with vocal tract length [1]. This dependency is utilized for speaker height estimation. Previous work related to audio feature representation includes I-vector framework using Mel Frequency Cepstral Coefficients (MFCC), or based on statistics of frame based features [2]. Recently, researchers estimated height using the sub glottal resonance frequencies [4]. Through the literature it's proven that there are no particular features as the standard one for this process. To estimate the height from the extracted features, common prediction analysis methods like linear regression, support vector regression, artificial neural networks are used. These systems are able to achieve a Root Mean Square Error (RMSE) around 6cm in predicting the height of a person. Such system is helpful in audio forensic applications for narrow down to a person, who made the threating or blackmailing calls.

In most of the forensic applications speech data will be in the form of a mobile call recording. Therefore in this work, we focus on the effect of wireless communication channel (Three path Rayleigh fading channel) on speaker parameter estimation from recorded speech data and tried to optimize the efficiency of estimation by reducing the error. Speech signal that passed through the Rayleigh communication channel will vary randomly and fade according to Rayleigh distribution. This will be eventually reducing the estimation efficiency. To compensate for the effect of wireless communication channel we did a two-step de-noising techniques which includes adaptive Wiener filtering and harmonic speech synthesise.

Organisation of paper as follows, in section 2 detail explanation of data collected which is used for experiments. Section 3 explains about the compensation techniques employed for improving the efficiency channel. Method for parameter estimation is explained in section 4. The experiments and evaluations are explained in section 5 and finally the conclusion of the work.

## II. DATA COLLECTION

For conducting the experiments, we have recorded the speech from 207 volunteers across different regions of India. Along with that we collected physical parameters like height and weight from speakers. The data set comprising of 162 males and 45 females, fall under the age group of 18-35 years. The recordings were performed in living room, conference room and class room. The recordings were at sampling rate of 16 kHz using simple head-phone microphone. Each speaker contributed 2minutes of speech data in 3sessions of each 40 seconds long. The Indian Newspaper articles are given to the speakers to read for the recording the speech. Each volunteer is asked to read both the mother tongue and English scripts for recording. Each individual height and weight are measured


DOI 10.1109/ADCOM.2016.15

in centimeters and kilograms respectively. Table 1 gives the statistics of collected data.

**Table 1**. Minimum, maximum, mean and standard deviation values of height and weight of the dataset

| Parameter | Minimum | Maximum | Mean | Standard deviation |
|---|---|---|---|---|
| Height | 147 | 188 | 167.99 | 8.45 |
| Weight | 39 | 107 | 64.47 | 12.09 |

## III. COMPENSATION TECHNIQUES

The wireless environment is highly unstable and fading is due to multipath propagation. Multipath propagation leads to rapid fluctuations of the phase and amplitude of the signal. As a result, the receiver sees the superposition of multiple copies of the transmitted signal, each traversing a different path with different SNR. After passing through the wireless communication channel the average SNR value of training and test data set reduces. Eventually this will reduce the parameter estimation accuracy. The compensation techniques used for cancelling the effect of multipath fading and Doppler shift includes Wiener filtering and Harmonic synthesize of speech data.

### *A. Wiener filter*

Wiener filter design requires a priori information about the statistics of the data to be processed. Optimum filter can be designed only when these characteristics match with the statistical characteristics of the input data. However, availability of such information cannot be guaranteed always and if that is the case, we may use estimate and plug procedure. The filter first estimates the statistical parameters of the relevant signals and plugs the obtained results into a non-recursive formula for computing the filter parameters. This procedure is not well suited for real time applications as it will require complex and costly software. To avoid this disadvantage, we can use an adaptive filter which is self-designing and relies for its operation on a recursive algorithm. Adaptive Wiener filter modifies its transfer function from frame to frame based on the speech signal statistics. This filter provides better results compared to traditional Wiener filter as well as spectral subtraction methods.

The Wiener filtering is a linear estimation of the original speech and it is optimal in terms of mean square error. In other words, it minimizes the overall mean square error in the process of inverse filtering and noise smoothing. The Wiener filter transfer function is given as, $w = \frac{S_{yy}-S_{nn}}{S_{yy}}$ where $S_{yy}$ is the Power Spectral Density (PSD) of the recorded signal along with noise and $S_{nn}$ is the Power Spectral Density of noise. The idea is that in frequency ranges where noise is high, $S_{nn}$ will be high and subsequently the Wiener filter coefficients of that frequency range will be low. The effect of noisy frames will be less after filtering owing to this. This works reasonably well for the stationary noise. However it fails to eliminate non-stationary random noises.

When the noise become non-stationary the local statistics of the speech signal may change. In such conditions, using an adaptive Wiener filter which is self-modifying in each signal frame can give much better signal enhancement. The adaptive filter transfer function of $i^{th}$ frame is, $H_i(\omega) = \frac{P_{iy}(\omega)-P_n(\omega)}{P_y(\omega)}$. And the output speech signal estimate in each frame is $\acute{P}_{is}(\omega) = P_{iy}(\omega)H_{i-1}(\omega)$, where, $i$ = 1, 2, 3… Frame size. Since the Wiener filter transfer function gets modified in each voiced frame, the filter identifies the non-stationary nature of the noise and gives much better performance when compared to classical Wiener filter.

### *B. Harmonic model*

Speech signal composed of two parts, deterministic signal and random signal. Harmonic model aims to represent the deterministic part of speech signal with a set of parameters such as frequency, amplitude and phase. For a 20 millisecond time duration we assume that the pitch of the speaker is constant. So by finding pitch we can model the speech as H harmonic components at the fundamental frequency *f0* and integer multiples of it, harmonic frequencies $f_h = (h + 1) f_0$. Periodic property of autocorrelation is used for finding the pitch.

$$s_l(n) = q(n) \sum_{h=0}^{H-1} 2A_{h,l} \cos(2\pi Fn + \phi_{h,l}) \quad (1)$$

$F = \frac{f_h}{f_s}$ and $f_s$ is the sampling rate and $\phi_{h,l}$ is the initial phase of component $h$ at the beginning of segment $l$. As mentioned earlier the pitch and harmonic amplitude $A_{h,l}$ for each segment is constant. The synthesise part generates each harmonics successively for the whole signal.

## IV. PHYSICAL PARAMETER ESTIMATION

In this Section, authors explain the detail procedure of physical parameter estimation. In Training phase the speech signals are passed through the simulated wireless channel by applying the compensation techniques as explained in section 3. Here we extract the log magnitude Short Time Fourier Transform (STFT) features for both training and test data. In the training stage the algorithm performs a *k-means* clustering. There after these are used for deriving the Bag of Words (BoW) representation during both the training and testing stages. These are given to support vector regression (SVR) model for predicting the parameters of the test data. This is shown in figure 1.

### *A. Feature extraction*

Most of the previous works related to height estimation from speech data is based on features such as MFCC, linear prediction coefficients (LPC), fundamental frequency (F0) and formants [5]. As stated earlier, there is no agreement on any single best feature set. In this work the authors use the bag of words representation of features. Each speech data is framed into 25msec window with 10msec frame shift between successive frames. Feature set includes log magnitude STFT along with optional derivatives. These feature set undergoes

K-means clustering and a set of representative mean vectors $M$i, i = 1,2,...K is extracted. Each speech frame is compared with each of the mean vectors for getting the BoW representation. Let $n_i$ be the number of speech frame nearer to $M_i$ these counts are normalized with the number of feature vectors $N$ in the utterance to obtain the BoW representation vector $v = \frac{1}{N}[n_1, n_2...n_k]^t$. This BoW features are given to support vector regression model as input. After modeling with training dataset, this model is employed for predicting the physical parameter of test dataset.

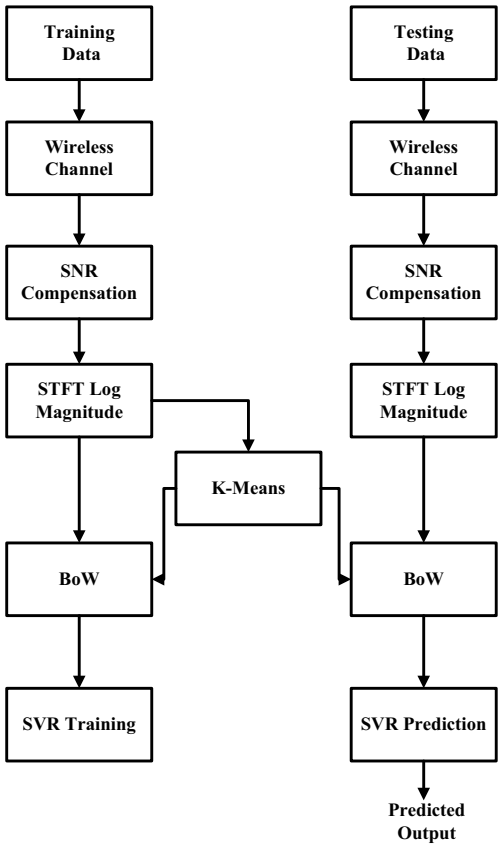



Fig. 1. Block diagram of speaker parameter estimation.

### B. Regression Model

Common approximation methods like linear and nonlinear regression models and artificial neural networks are used previously for predicting the physical parameters. Here in this work we are using support vector regression [6] to predict the physical parameters from BoW representation. Using a kernel support vector regression can be easily generalized to a non-linear regression model. Consider the training data $\{(v_1, y_1), (v_i, y_i), \ldots. (v_N, y_N)\}$, Where $v_i$ is the input BoW representation and $y_i$ represents the physical parameter to be predicted. Support vector regression model determines the maximum deviation of $\varepsilon$ from the target value. If the errors are larger than $\varepsilon$ then only they are penalized. The function is optimized to have as much flatness as possible.

The regression function in the linear SVR is represented by :

$$f(x) = < w, x > + b \quad (2)$$

where $< . , . >$ denotes the inner product , $w$ and $b$ are the parameters of SVR. Optimizing the flatness in the linear case is equivalent to

$$\text{minimize } \frac{1}{2} ||w||^2 \quad (3)$$

$$\text{subject to } | < w, x_i > + b - y_i | < \varepsilon \quad (4)$$

This however assumes that such a $f(x)$ exists that results in an absolute error less than $\varepsilon$ for every sample. Often this assumption is invalid, and a set of slack variables are introduced in order to allow deviation from this constrain. This is similar to the " Soft Margin " approach in Support Vector Machines (SVM).

The SVR objective function is optimized using the dot products of the data points among themselves. Therefore, the linear SVR can easily be extended to a non-linear SVR by means of kernel trick. In this method valid kernel replaces the dot product in the algorithm. The authors chose a normalized polynomial kernel as proposed by [7] for the task of height estimation.

The kernel is given by:

$$K(x,y) = \frac{(<x,y>)^n}{\sqrt{(<x,x>)^n (<y,y>)^n}} \quad (5)$$

Where $<x,y>$ denotes the inner product of vectors $x$ and $y$

## V. EXPERIMENTS AND EVALUATION

We use the dataset described in section 2 for the evaluation purpose. The data set is divided in to training and testing, which includes 137 speakers (690 training speech files) and 70 speakers (316 test speech files) data respectively. It is ensured that both train and test dataset are mutually exclusive. Simulated a three path Rayleigh fading channel with typical values for channel parameters such as Doppler shift, path gain, path delay, signal to noise ratio etc. Generally mobile station speed varies from 5km/hr. to 100km/hr., therefore the maximum possible Doppler shift will be in the range of hundreds. Values taken for Doppler shift, path gain, path delay, and signal to noise ratio are 100 Hz, [0, -1, -1], [0, 0.15 msec, 0.32 msec], 10 dB respectively. GMSK modulation technique is used for modulating the carrier.

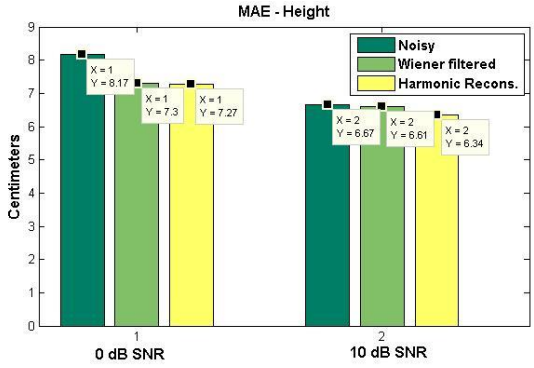
Each train and test speech data is passed through the simulated wireless communication channel. To compensate for the reduction in average SNR by the fading effect of channel, Wiener filtering and harmonic reconstruction is done as explained in section III. To study the effect of channel on audio forensics parameter estimation is done at every stage.

Mean Absolute Error (MAE) and RMSE for height and weight of clean speech data after SVR prediction is given in Table. 2. From the result we can infer that, the RMSE value for height that we got is almost equal to what is reported in literature (6.2cm is reported in [2] and 6.8cm in [3]).

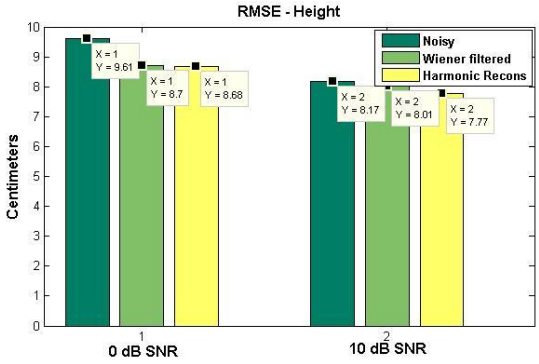
Table. 2. MAE and RMSE of clean speech after prediction.

| Parameter | MAE | RMSE |
|---|---|---|
| Height | 5.24 | 6.65 |
| Weight | 7.28 | 9.86 |

Wireless channel is simulated for 10dB and 0dB SNR and predicted physical parameters. MAE and RMSE values of height and weight predictions are shown in below figure 2 and 3 respectively.



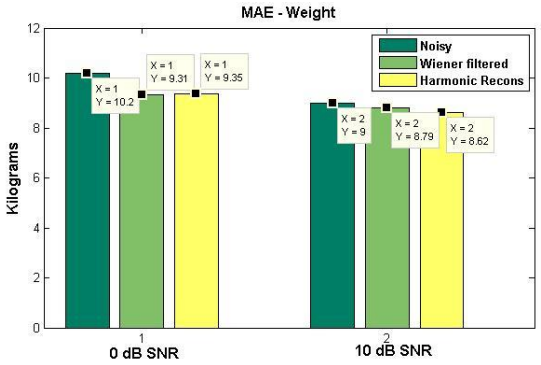

(a) MAE



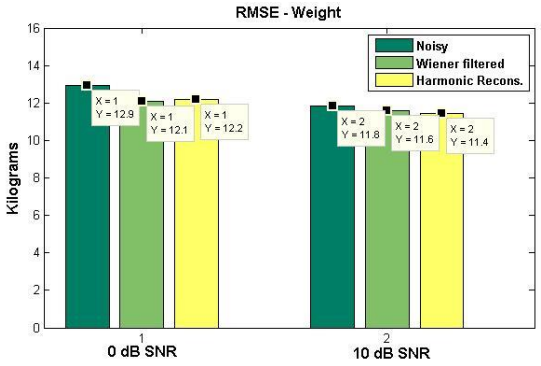

(b) RMSE

Fig 2. MAE and RMSE of parameter Height of Noisy. Wiener filtered and Harmonic Reconstructed speech after prediction.

Compared with the previous result for clean speech, the obtained RMSE values clearly showing the effect of channel. This is because of the multipath fading effects of Rayleigh channel. We can see a slight increment in RMSE for 0dB SNR compared with 10dB. Therefore to improve the efficiency of prediction adaptive Wiener filtering and Harmonic model speech reconstruction is implemented as described in section 3. Prediction results after Wiener filtering and Harmonic reconstruction is shown in figure 2 and figure 3. From the figure we can infer that both of the compensation techniques helped to improve the efficiency of prediction. Adaptive Wiener filter removed both stationary and non-stationary noise from the signal.




(a) MAE




(b) RMSE

Fig 3. MAE and RMSE of parameter weight of noisy, wiener filtered and Harmonic Reconstructed speech after prediction.

## VI. CONCLUSION

In this paper, we had studied the wireless channel effects on audio forensics and predicted the physical parameters like height and weight. For this work, we have collected the new dataset of 207 speakers.

From the recorded speech data set we extracted the log magnitude spectrum and its delta features. These features are represented in BoW used for support vector regression model to estimate the height and weight of speakers. The obtained RMSE and MAE for height are 6.65cm and 5.24cm respectively.

In forensics analysis, most of the cases the speech data will be in the form of mobile recording or telephonic call. Therefore authors decided to study the effect of wireless channel. Simulated three path Rayleigh fading channel and speech data is passed through the channel. This resulted the noisy speech and found there is a higher RMSE of 8.17cm for height and 11.84cm for weight. To compensate the effect of wireless channel fading, we implemented adaptive Wiener filter and harmonic model reconstruction. The compensation worked reasonably well and RMSE value reduced to 7.77 for height and for weight it reduced to 11.43.

## REFERENCES


[1] 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.

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

[3] Martin Krawczyk-Becker and Timo Gerkmann, "MMSE-optimal combination of wiener filtering and harmonic model based speech enhancement in a general framework", *Workshop on Applications of Signal Processing to Audio and Acoustics, IEEE,* 2015.

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

[5] Julio Gonzalez, "Estimation of speakers' weight and height from speech: A re-analysis of data from multiple studies by lass and colleagues," *Perceptual and motor skills,* vol. 96, no. 1, pp. 297-304, 2003.

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

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