SPEAKER HEIGHT ESTIMATION COMBINING GMM AND LINEAR REGRESSION SUBSYSTEMS

Keri A. Williams, John H.L. Hansen

Center for Robust Speech Systems
University of Texas at Dallas, Richardson, Tx, USA

Kas067000@utdallas.edu, john.hansen@utdallas.edu

ABSTRACT

There are both scientific and technology based motivations for establishing effective speech processing algorithms that estimate speaker traits. Estimating speaker height can assist in voice forensic analysis [1], as well as provide additional side knowledge to improve speaker ID systems, or acoustic model selection for improved speech recognition. In this study, two distinct approaches for height estimation are explored. The first approach is statistical based and incorporates acoustic models within a GMM structure, while the second is a direct speech analysis approach that employs linear regression to obtain the height directly. The accuracy and trade-offs of these systems are explored as well a fusion of the two systems using data from the TIMIT corpus (which includes ground truth on speaker height).

Index Terms – height estimation, GMM, formants

1. BACKGROUND

Speaker identification systems can be very effective, however their performance is limited by the available training data needed for each speaker. In open-set speaker recognition, most systems are focused on recognizing the in-set group and rejecting all out-of-set speakers. However, in many applications, it is desirable to extract some information regarding the out-of-set speakers (as well as in-set speakers if that knowledge is not known a priori). Also, extracting supplementary physical characteristics could help improve speaker recognition systems as additional items to aid in identification. Extracting a speaker trait such as height from speech is one physical aspect that would be helpful to know about a speaker. Age, weight, gender, ethnicity, or health are other potential traits. The relationship between height and speech has been explored before, and lends itself to the feasibility of extracting height from speech. In speech, it has been well proven that an increase in the vocal tract length of a person leads to a decrease in formant frequency locations [2]. However, this simply shows that the vocal tract length directly affects the speech structure. Another study was conducted that examined the correlation between vocal tract length and height in men and women. The correlation was strong for both men and women with coefficients of .855 and .832 respectively, showing that height and vocal tract length are related [3]. Since the relationship between height and speech is easily shown, some approaches have extended this idea for automatic speaker height estimation. One technique used was linear regression which proved to be relatively successful, however one of those studies considered a single sustained vowel which is not useful in practical scenarios[4] [5]. A different regression study used the second subglottal resonance to determine the height of the speaker. This was based on the fact that if the vocal tract length is related to the height of the speaker, than the length below the vocal folds should be related as well [6]. Other studies have considered classification approaches using MFCCs and GMMs to enhance text independent speaker height estimation for voice forensic analysis[1]. Using such an approach has the advantage of being text independent which is ideal, but the result is only a height class and not an actual height, which can be achieved with regression techniques. The approach taken in this study is to develop two systems based on the general approaches taken in the past (i.e., GMMs and regression), and then combine them to achieve improved accuracy. The first system, Modified Formant Track Regression, is based on linear regression and uses smoothed formant tracks as the feature. The second system, Height Distribution Based Classification, is a classification approach that uses 19 static MFCCs within a dynamic height bin width GMM structure for different height classes. A confidence measure is included with the result of the second system.

2. CORPUS

Little if any formal major data collection has been undertaken specifically for height estimation. All data used here for training and test was taken from the TIMIT corpus since it contains height information for every speaker [7]. The distribution of the heights for males and females proved to be similar to the general US population [8]. This would allow for testing to better represent the a priori population of the USA. The heights for the TIMIT corpus however, were self-reported which is assumed to introduce some subject error. Studies have shown that individuals often overestimate their height, but the overestimation was small by a majority of the subjects [9]. Therefore, we expect this self-reported bias will introduce some error, but is expected to be minimal (i.e., since an IRB protocol was followed in collecting TIMIT, it is not possible to identify actual speaker names with ID labels, and therefore there is less of an issue subjects would intentionally inflate heights if they are short to average, or underestimate if they are average to tall).

This Project was funded by AFRL under contract FA8750-12-1-0188 and partially by the University of Texas at Dallas from the Distinguished University Chair in Telecommunications Engineering held by J.H.L. Hansen.

3. MODIFIED FORMANT TRACK REGRESSION (MFTR)

3.1 Feature Estimation - Height

It is well known from speech analysis using acoustic tubes that vocal tract length, which is correlated to a speaker's height, is related to formant locations. However formant estimation can be erroneous, so the raw formant tracks are modified to eliminate spurious peaks.

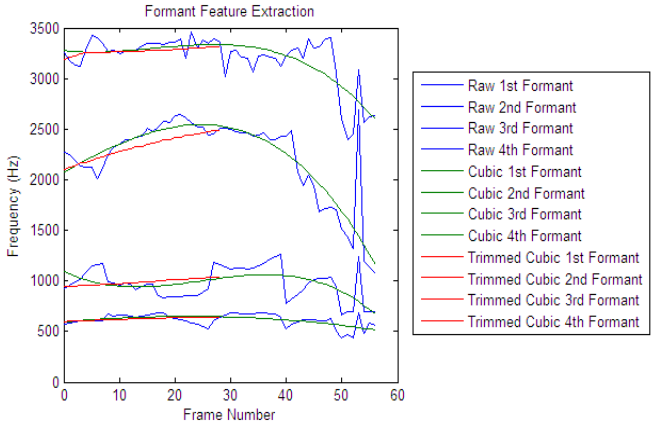

Figure 1: Example of Feature Extraction Steps

The first step in creating this feature is to extract the first 4 formant tracks for voiced speech from the particular speaker (see Figure 1a). This is accomplished by finding the poles of an all pole model. In order to find the poles, the number of coefficients is determined as,

$$n = \frac{fs}{1000} + 2 \quad . \tag{1}$$

The next step is to find the LPC coefficients and determine the roots of the equation. Once the roots are found, the formant location estimates are calculated by;

$$F_i = \frac{fs}{2\pi} * \arctan\left(\frac{im(r)}{re(r)}\right) \quad . \tag{2}$$

After the raw formant tracks are estimated, the next step is to fit the result to a cubic equation and find these coefficients. A cubic function is used since it has been determined to be sufficient in representing a formant track [10]. Once the coefficients are determined for each formant track, the raw formant tracks are replaced with the result of the cubic function (see Figure 1b). The cubic formant track is then sorted to prepare for trimming. The lowest 25% and highest 25% are then eliminated, leaving only the middle 50% (see Figure 1c). After processing the formant tracks, the output tracks are much smoother with less wide dynamic variations. This should help reduce the error caused by formant estimation.

3.2 Algorithm: MFTR

The modified formant track regression algorithm for height estimation is based on solving an equation that represents the height of a speaker in terms of the first four formants, and then cleaning up the height estimates through post-processing (see Figure 2). The first step of the algorithm is to recognize four distinct vowels from a given sentence. The four vowels are /AA/, /AE/, /AO/, and /IY/. They were chosen due to the quantity of the speakers that uttered these vowels. Formants for vowels are steady and are different for each vowel. As a result, the feature will be more reliable for a vowel, but it must be calculated for each of the 4 phonemes.

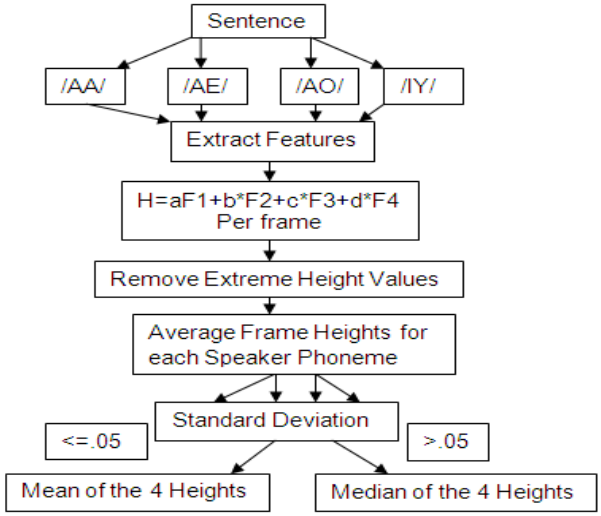

Figure 2: Modified Formant Track Regression Algorithm

Once the 4 sets of features are calculated, they are incorporated into Equation 3 which relates height as a linear combination of the four formants.

$$H_r = a * F_1 + b * F_2 + c * F_3 + d * F_4 \tag{3}$$

This equation produces a height estimate for each particular frame for each vowel. The next step is to combine the heights across the different frames to achieve a height for the speaker for each vowel. First, the extreme values of the heights are removed and the resulting frames are averaged to result in 4 different heights for one speaker, one for each vowel. The standard deviation of these 4 heights is then calculated and if it is above a threshold the median of the 4 heights is calculated, otherwise they are averaged. This happens because for a high standard deviation, the heights are more spread out so there is less confidence in the result. In this way, only the two middle values are considered in the calculation. With a low standard deviation, the heights are more tightly clustered so an average would represent the height quite well. At the end of the algorithm there is only one height estimate for each speaker.

3.3 Training: MFTR

All data used for evaluation is from TIMIT with a 16 kHz sampling frequency, however not all of the data could be used due to the phoneme dependence of this method. The four vowels were chosen since a large number of the speakers uttered them in the sa1 sentence. In total, there were 268 males and 127 females. For the other 9 sentences produced by each of these speakers, all were examined to see if they had any of the 4 vowels, if so the sentence

was included. None of the speakers in the training set were in the test set, and half of the speakers were used for training.

3.4 Results: MFTR

The results for the modified formant track regression are illustrated in Table 1. The metric used to examine the performance was mean absolute error (MAE (cm)), which has been used in previous studies for height estimation [4,5,6]. It was calculated on a per speaker basis since there is only one height per speaker.

	MAE (cm)				
	/AA/	/AE/	/AO/	/IY/	All
Male	7.23	7.52	7.43	6.58	6.36
Female	7.21	8.43	6.65	8.37	6.8

Table 1: MAE results for females and males for MFTR Method

The best result is obtained when combining the heights from the 4 vowels which is expected. When combined, there is more information available. Also, if one phoneme performs poorly, the other three can help counteract the error. Using the different phonemes, there is built-in backup system available. For females, the /AO/ phoneme performed best but it could be due to the smaller data set. All phonemes performed differently because formant estimation errors can differ, and depending on neighboring phonemes the formants can change towards the edges. The extreme formant estimation errors are addressed with the smoothing and trimming performed in the feature processing, but coarticulation effects and minor estimation errors are not necessarily eliminated.

4. HEIGHT DISTRIBUTION BASED CLASSIFICATION (GMM-HDBC)

4.1 Feature Estimation - Height

The feature used for this method is 19 static MFCC coefficients along with normalized energy. MFCCs have been shown in a previous study to be effective in representing a speaker's height [1]. This is possible since the static MFCC coefficients tend to be related to a person's vocal tract configuration [1]. The normalized energy is included in order to use a threshold to eliminate silence, since silence would not add any useful information.

4.2 Algorithm: GMM-HDBC

This method is focused on a sentence level analysis and extracts 19 static MFCC coefficients as described in Section 4.1. From there, the features are processed into different traditional GMMs. In order for the GMM structure to work, the heights need to be grouped within height ranges. Instead of employing an equally spaced scale where heights are distributed along uniform marks (as was performed in [1]), the groups were partitioned based on how much data was available for each height (see Figure 3). In this manner, the intrinsic a priori probability of the height distribution of the population under train/test would be incorporated, which also allows for data balancing of the models. Some heights have significantly more data than others, especially around the centroid of the height distribution scale. Using a linear partitioned scale, the tails of the height models do not have as much training data, so the height GMMs become more speaker dependent versus central height models that are more speaker independent.

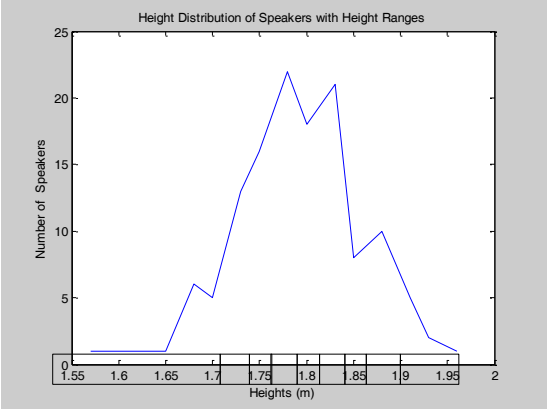

Figure 3: Male Height Ranges for GMMs with Training Set Speakers

To address this problem, a minimum threshold was set for the number of speakers needed to construct each height range GMM. From this strategy, the groups were formed based on the distribution of how many speakers are present for each height, and if insufficient, that group was added to the neighboring group. The minimum number of speakers for males was set to 20, and for females it was set to 12. This configuration will result in a height class being determined for each speaker. The centroids in meters for the males are 1.635, 1.73, 1.75, 1.78, 1.8, 1.83, 1.85, 1.88, and 1.935, while for females they are 1.51, 1.6, 1.63, 1.65, 1.68, 1.7, 1.73, and 1.79. It would be useful to also include a confidence measure to show how likely that height class is. The confidence measure used is the probability closeness measure, whose formula is shown in Equation 4 [11].

$$confidence\ m1 = \frac{\frac{1}{p1}}{\frac{1}{p1}+\frac{1}{p2}+\frac{1}{p3}} \quad (4)$$

This confidence measure will state how separable the top 3 height models probabilities are, which reflects confidence in the model choice. The higher the top result's probability is compared to the second and third, the closer the measure approaches one. Now each speaker will have a height class associated with it, as well as a confidence measure.

4.3 Training: GMM-HDBC

Even though all of the TIMIT data could be used, since this method is text independent, the same data used in the MFTR method was used in assessing this method. This was done to provide consistency and allow the two methods to be easily combined. Each GMM has 64 mixtures to cover all of the given speaker independent data in the specified height range.

4.4 Results: GMM-HDBC

In order to examine the accuracy of this method, the classification accuracy within 5 cm was examined as the confidence measure increased. As the confidence measure increases, there is less data used to calculate the accuracy so the 25% and 50% data elimination points were plotted for reference with a vertical line (see Figure 4).

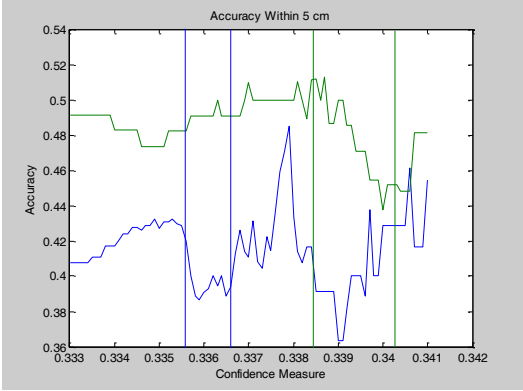

Figure 4: Accuracy of GMM-HDBC Method (blue=males, green=females)

The accuracy is dependent upon the limited amount of data used for training and testing, as well as the text independence nature of the method. The confidence measure is shown to be helpful in judging how well a result can be relied on since as it increases the accuracy by as much as 8%.

5. FUSION OF THE TWO METHODS

5.1 Algorithm: MFTR & GMM-HDBC

The MFTR algorithm results in a height for each speaker while the GMM-HDBC method results in a height class along with a confidence score. The fusion system will take both outputs and combine them to result in one height for each speaker (see Figure 5).

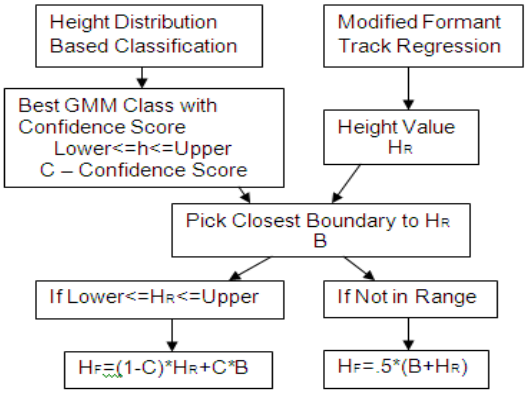

Figure 5: Algorithm for Fusion

When combining the two systems, the first step was to find which boundary from the classification system was closest to the result from the regression method. Once the upper or lower boundary of the height class is chosen, the next step is to see if the two systems agree. This means that the result from the regression system is within the height class range. If the two results agree, the final height is determined from the equation in Figure 5 which relates the closest boundary, B, the result from the regression system h_R, and the confidence measure, C. For higher confidence measures, more emphasis is placed on the boundary while for low confidence measures more emphasis is placed on the regression result. If the two systems disagree, which means that the regression result is not in the height class range, then the regression result and the closest boundary are averaged together. This will result in a compromised height estimate. With this method, there will be only one height result per speaker.

5.2 Results: MFTR & GMM-HDBC

The results for the fusion system are determined in terms of mean absolute error, which is the same measure used for the MFTR method. The results for males and female are summarized in Table 2.

	MAE (cm)
Male	5.37
Female	5.49

Table 2: MAE results for Fusion

These MAE for the fusion method is better than the MAE for the regression method. The classification method helped modify the original regression method results by using more information and the confidence score. Finally, it should be note that an upper bound on performance is not really known, since speech structure including vocal tract length are not perfectly correlated with height.

6. CONCLUSION

Two methods were developed for engaging an automatic speaker height estimation solution as well as a fusion of the two methods. The first method, MFTR, obtains a single exact height for each speaker but is dependent on 4 specific vowels to obtain the results. This can result in setting aside a portion of the speech data due to required vowel coverage. The GMM-HDBC method was text independent but did not result in an exact height. It resulted in a height class which included a range of heights. The classification method also resulted in a confidence measure to provide feedback on the result. Both methods have their strong and weak points, so a fusion system was developed to provide better accuracy. The fusion system resulted in a single height per speaker and helped the regression results by using the height class and confidence score. The results of these methods were very promising, but further work could be considered to improve the feature for the regression method and using i-Vectors instead of GMM models for the height ranges to improve robustness since only clean data is used in this study.

7. RELATION TO PRIOR WORK

Height estimation has been examined before but only a regression based technique [4,5,6] or a GMM based technique [1] was used. This paper formulated modified/improved ideas along with a combination of the two approaches. The selection of the GMM height classes was new as well as the modifications made to the formant tracks. The confidence measure was also a new addition from previous work.

8. REFERENCES

[1] B. Pellom, J.H.L. Hansen, "Voice Analysis in Adverse Conditions: The Centennial Olympic Park Bombing 911 Call," *IEEE Midwest Symposium on Circuits & Systems*, pp. 873-876, Aug., 1997.

[2] D. Smith, R. Patterson, R. Turner. "The Processing and Perception of Size Information in Speech Sounds". *Journal of the Acoustical Society of America*, Vol. 117, pp. 305-318, Jan 2005.
[3] J. Giedd, W. Fitch. "Morphology and Development of the Human Vocal Tract: A Study Using Magnetic Resonance Imaging". *Journal of the Acoustical Society of America*, Vol. 106, pp. 1511-1522, Sept 1999.
[4] R. Greisbach. "Estimation of Speaker Height From Formant Frequencies". *Forensic Linguistics*, Vol. 6, pp. 265-277, 1999.
[5] I. Mporas, T. Ganchev,"Estimation of Unknown Speaker's Height From Speech," *International Journal of Speech Technology*, pp. 149-160, Jan, 2010.
[6] A. Alwan, H. Arsikere, G. Leung, and S. Lulich, "Automatic Height Estimation Using the Second Subglottal Resonance," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Kyoto, Japan, 2012, pp. 3989-3992.
[7] National Institute of Standards and Technology (NIST), "Getting Started With The DARPA TIMIT CD-ROM: An Acoustic Phonetic Continuous Speech Database," NIST, 1988.
[8] "Cumulative Percent Distribution of Population by Height and Sex."http://www.allcountries.org/uscensus/230_cumulative _percent_distribution_of_population_by.html [28 Feb 2012].
[9] I. Perry, J. Brestoff, J. Van der Broeck, "Challenging the Role of Social Norms Regarding Body Weight as an Explanation for Weight, Height, and BMI Misreporting Biases: Development and Application of a New Approach to Examining Misreporting and Misclassification Bias in Surveys," *BMC Public Health*, pp. xx-xx, 2011.
[10] C.C. Goodyear and A.R. Greenword, "A Polynomial Approximation to the Acoustic-To-Articulatory Mapping," *IEEE Colleqium on Techniques for Speech Processing and their Application*, pp. 8/1-8/6, 1994.
[11] L. Rabiner and R. Schafer, "Algorithms for Estimating Speech Parameters," in *Theory and Applications of Digital Speech Processing*, 1st ed. Upper Saddle River, NJ: Pearson Higher Education Inc, 2011, ch. 10, sec. 7, pp. 650.