AN ESTIMATE OF PHYSICAL SCALE FROM SPEECH

Lawrence H. Smith

National Institute of Health
Bathesda, Md. 20892

Douglas J. Nelson

U.S. Department of Defense
Fort George G. Meade, Md. 20755-6514

ABSTRACT

We present an algorithm, based on the EM algorithm, which simultaneously estimates both physical scale and vowel identities from a segment of speech. The validity of the algorithm depends on the scale hypothesis that the variation of the formant frequencies for a given vowel is mainly influenced by the physical size of the speaker. This is both a new application and a new confirmation of a hypothesis that is often accepted without proof.

1. INTRODUCTION

We present an algorithm, based on the EM algorithm, which simultaneously estimates both physical scale and vowel identities from a segment of speech. The validity of the algorithm depends on the scale hypothesis that the variation of the formant frequencies for a given vowel is mainly influenced by the physical size of the speaker. This is both a new application and a new confirmation of a hypothesis that is often accepted without proof.

The formant frequencies of a phoneme are determined by the length of the speaker's vocal tract during vocalization [4], and the vocal tract length is correlated with the size of the speaker. This connection has been established for rhesus macaques [5] and for children [15], and it has been shown that phonetically similar vowels spoken by different speakers satisfy a scale relationship related to the MEL scale of hearing [16, 17]. However, size is not necessarily the direct cause of formant frequency variation in human speech, since the frequencies are varied intentionally in the production of phonemes. But the reason behind the correlation of physical size and formant frequency is immaterial. It need only be accepted that there is a real mechanism connecting them.

It is our hope that this algroithm may provide some advantage in speaker identification and speaker adaptation in speech recognition. In addition, it provides information about the speakers gender, independent of fundamental frequency, F0.

2. DATA ANALYSIS

The scale hypothesis was developed and tested on Hillenbrand's phonetically marked and annotated western Michigan vowel data [6]. These data consist of 12 vowels spoken in the "hVd" context by 45 men and 48 women. The child data were not used and to equalize the sample sizes, only 33 of the female speakers were used.

By assuming the identities and formant frequencies of vowels, a statistical model of the data was tested based on a single unknown scale parameter. The formant frequencies and covariances observed in these data were then used to define the parameters of the algorithm. The algorithm was then applied to telephone quality voice recordings [14] to estimate the physical scale of the speaker and the identities of the spoken vowels.

The published time-markings of the phonemes were used in the analysis. However, the vowel identities were not used by the algorithm, although the inputs were limited to phonemes that were tagged as being vowels. The sampled speech corresponding to each phoneme was then analyzed using a 12 point linear prediction model using the correlation method [12]. The formant frequencies were then estimated by root solving and taking the frequencies greater than 200 Hz. The 12 coefficient correlation method was chosen over all other methods because it was found to provide the best correlation to the quality controlled formant frequencies of Hillenbrand [6] (correlation of 0.9155 and 0.7947 for formants 1 and 2 respectively).

3. SCALE HYPOTHESIS

Analysis of variance confirms that gender can be inferred by comparing the formant frequencies of a known vowel spoken by an unknown speaker to averaged male and female formant frequencies [18]. When the same analysis is applied to the current data, one is led to the same conclusion (results not shown). The average frequencies are shown in Table 1.

	Men		**Women**		**All**	
Vowel	$\mathbf{F_1}$	$\mathbf{F_2}$	$\mathbf{F_1}$	$\mathbf{F_2}$	$\mathbf{F_1}$	$\mathbf{F_2}$
ae	595	1922	690	2320	642	2121
ah	759	1317	930	1523	845	1420
aw	657	1033	816	1190	737	1112
eh	586	1799	733	2063	659	1931
ei	478	2081	540	2521	509	2301
er	475	1370	528	1583	502	1476
ih	432	2024	488	2367	460	2196
iy	343	2316	441	2740	392	2528
oa	495	909	569	1035	532	972
oo	469	1129	524	1226	496	1178
uh	622	1194	771	1435	697	1314
uw	378	990	464	1101	421	1045

Table 1. Average Formant Frequencies, men, women and men and women combined.

Rather than making a comparison to gender normal frequencies, formant frequencies were modelled as scaled versions of normal frequencies. This model is suggested by the data shown in figure 1.

For each of the 12 vowels and for each formant frequency the average value of the 33 men is plotted as the ordinate and the average value for the 33 women is plotted as the abscissa. The striking feature of this plot is that the data obeys a linear relationship.

The linear relationship in the plot is the quantitative consequence of the scale hypothesis. The hypothesis implicitly assumes that, for each vowel, there is a set of normal formant frequencies, and the formant frequencies for a specific speaker are all a fixed multiple of the normal frequencies. For speaker s and phoneme v, let $\mathbf{F}_{sv}$ denote the vector of expected formant frequencies. Then

$$\mathbf{F}_{sv} = a_s \mathbf{F}_{0v} + \epsilon_{sv} \tag{1}$$

where a_s is the scale factor and ϵ_{sv} is a mean zero random variable. For the sake of modelling, we assume that ϵ_{sv} are independent identically distributed multivariate normal random variables with mean $\mathbf{0}$ and a given covariance matrix Σ. We will assume the covariance estimated from the data set shown in Table 2.

$$\Sigma = \begin{pmatrix} 5.48 & 7.01 \\ 7.01 & 33.74 \end{pmatrix} \times 10^3 \qquad \Sigma^{-1} = \begin{pmatrix} 24.84 & -5.16 \\ -5.16 & 4.04 \end{pmatrix} \times 10^{-5}$$

Table 2. Σ and Σ^{-1}

The assumption of equality of variances is contrary to evidence that the variance of formant frequencies is greater for women than for men, and generally increases with frequency.

As a_s and $\mathbf{F}_{0v}$ are not well defined, the model is unchanged if we multiply all a_s by a constant while multiplying all $\mathbf{F}_{0v}$ by its inverse. Therefore it can be assumed that the mean of a_s for the combined population of male and female speakers is 1. Then taking the mean over s we find

$$\operatorname*{mean}_s \mathbf{F}_{sv} = \mathbf{F}_{0v},$$

and this is how we estimate $\mathbf{F}_{0v}$. The normal formant frequencies, i.e. the average of the 90 speakers of the formant frequencies for each of the 12 vowels, is shown in Table 1.

4. TEST OF SCALE HYPOTHESIS

The significance of the scale hypothesis defined by equation (1) is found by testing the statistical hypothesis $H_0 : a_s = 1$ against $H_1 : a_s \neq 1$. This test of hypothesis can be converted to an equivalent test of hypothesis in which the statistical variations have multivariate standard distributions. The maximum likelihood estimate of a_s based on (1) minimizes the Malanobis distance $\sum_v \mathrm{d}_\Sigma(\mathbf{F}_{sv}, a_s \mathbf{F}_{0v})$, and it is equivalent to a maximum likelihood estimate that minimizes a Euclidean distance. Here d_V denotes the metric

$$\mathrm{d}_V(\mathbf{y}, \mathbf{x}) = (\mathbf{y} - \mathbf{x})' V^{-1} (\mathbf{y} - \mathbf{x}).$$

Furthermore, the F statistic for testing H_0 is the same statistic as in the Euclidean version with Σ replacing I, that is

$$F = \frac{(SS_0 - SS_1)/1}{SS_1/23}$$

$$SS_0 = \sum_{v=1}^{12} \mathrm{d}_\Sigma(\mathbf{F}_{sv}, \mathbf{F}_{0v}) \qquad SS_1 = \sum_{v=1}^{12} \mathrm{d}_\Sigma(\mathbf{F}_{sv}, \hat{a}_s \mathbf{F}_{0v})$$

where $\hat{a}_s$ is the MLE a_s,

$$\hat{a}_s = \frac{\sum_{v=1}^{12} \mathbf{F}'_{sv} \Sigma^{-1} \mathbf{F}_{0v}}{\sum_{v=1}^{12} \mathbf{F}'_{0v} \Sigma^{-1} \mathbf{F}_{0v}}. \tag{3}$$

The 0.05-critical level for rejection of the null hypothesis is $F_{0.05,1,23} = 1.71945$. The F statistic for 33 men and 29 women out of the total of 33 selected from the Western Michigan database exceed this critical value, whereas under the null hypothesis we would expect only 5% or 4.5 speakers. Clearly H_0 is rejected in favor of the scale hypothesis.

Assuming that the hypothesis is established, the distribution of the MLE a_s are shown in Figure 2.

As one would expect, the scale values are significantly higher for women (mean value 1.0856) than for men (mean value 0.9144). The rank sum test for the two populations of scale values gives the highly significant statistic of NA, NA standard deviations (77.98) above the mean 1105.5.

5. GENDER PREDICTION

The fundamental frequency of a speaker provides the most obvious and accurate clue to the gender of the speaker (c.f.[13]). Yet within gender groups fundamental frequency is not correlated with physical size [10, 8]. Other approaches to gender identification also do not rely explicitly on a model of the physical size of the speaker. In ref.[18, 3] feature vectors are computed for an utterance (LPC, for example) and these are associated with a nearest gender-associated template vector. In ref. [1] vocal tract length is used similarly as one feature in gender identification, but since this is defined as an explicit function of formant frequency, it cannot be said that the physical size is being "modelled."

From the above observations, scale value is clearly a strong indicator of gender. As designed, a scale value greater than 1 is indicative of a smaller than average speaker, and hence associated more with female speakers. Scale values less than 1 are similarly associated with male speakers. Since F_0 is known to indicate gender, we might ask whether this scale value is related to F_0. The mean value of F_0 (as given by Hillenbrand [6] is plotted as the ordinate against the computed scale for each of the 90 speakers. Since F_0 and scale both correlate with gender, there is an overall correlation (value o.89339). However, within gender, the correlation is less (for men 0.14217 and for women 0.46589). This suggests that within gender, and more so for men, the scale estimate is independent of fundamental frequency.

6. VOWEL AND SCALE ESTIMATION

If formant frequencies are given for some known vowels spoken by an unknown speaker, then the MLE a_s can be obtained from (3). But if formant frequencies are given for some unknown vowels spoken by an unknown speaker, then we can no longer estimate a_s directly. Instead, both vowels and a_s must be estimated. In the intended application, this problem is to be solved given only estimates of F_1 and F_2.

The method of estimating vowels and scale is an iterative EM algorithm [11]. We assume that the first two formant frequencies are given for some number of unknown vowels spoken by the same unknown speaker. Begin the algorithm by setting $a_s = 1$. Scale the normal frequencies for each of the 12 vowels by a_s and for each unknown vowel, choose the scaled vowel that is closest to it (in Mahalanobis distance). Then update a_s with the MLE a_s using (3) assuming that the chosen vowels are correct. Continue rescaling, classifying vowels, and updating the value of a_s until a_s "converges."

In practice we may not know that a particular sampled phoneme is one of the known vowels. In this case the program must also decide whether the observed formants are consistent with the set of known vowels, and if not then reject the phoneme as an 'outlier.' Outliers are determined by the minimum distance to a known vowel. If the minimum distance is greater than a critical distance for that vowel, then the sample is rejected and excluded from the reestimation. This outlier rejection process is performed on each iteration since the rejection at any one step is at best an approximate rejection. A sample that is rejected early in the calculation could theoretically be readmitted if it is rescaled within a valid range.

A conservative outlier rejection was used in this study. Based on the Western Michigan formant data, a Mahalanobis distance of 10 to the F_1, F_2 vowel centers covered more than 95% of the data. Using this cutoff value resulted in rejection of obvious outliers in the TIMIT database and improved the scale and vowel estimates in cases where there was an obvious non-vowel.

7. RESULTS

The EM algorithm described above, based solely on the mean frequencies derived from Hillenbrand [6], was applied to the vowels of the TIMIT and NTIMIT recordings in both 2 and 3 formant versions. The results were found to be better for the TIMIT than the NTIMIT database, as would be expected since the NTIMIT database is degraded by telephone exchange filtering. The results were also better when 2 formants were used instead of 3, and this may be the natural consequence of the LPC algorithm used to estimate the formant frequencies, which has an increased error of identifying the formant number of the higher formants. We report the NTIMIT results based on 2 formants.

The overall distribution of scale values derived from 6300 sentences is shown in figure 3 for men and for women.

Comparing with figure 1, we see that the tendency of female speakers to have a higher scale is preserved, (mean value for women l.0282 and for men 0.9785). The differences are not nearly as pronounced, but nevertheless they

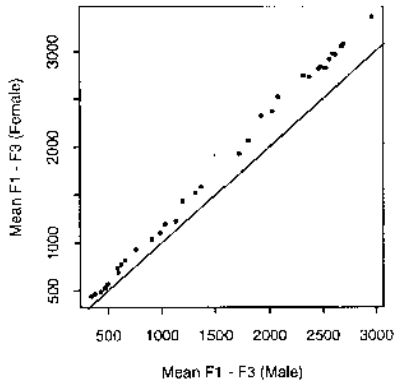

Figure 1: Male and Female Formant Frequencies

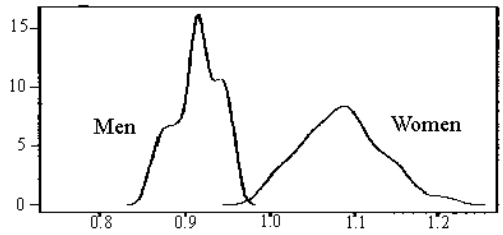

Figure 2: Male and Female Scale Values

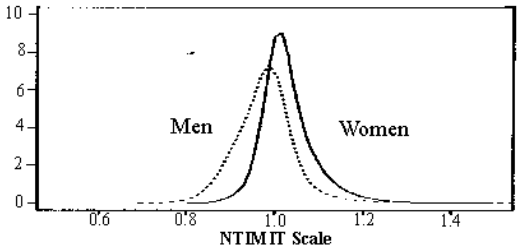

Figure 3: Male and Female Scale Values

are significant (rank sum test is 32.6 deviations). The distributions of scale within each of the 8 regions shows the same tendency of higher female scale. The high degree of overlap in the distributions indicate a degradation in the ability of the algorithm to predict scale.

The NTIMIT database also provides the height of each speaker. The overall correlation of scale ot height was found to be -0.27547 which is negative as expected, and significant ($p = Beta - distribution$). But this is not the whole story. Within each of the 8 regions, there is a diversity of correlations of height to scale, as shown in table 3.

Region	Men	Women
1	−0.06137	−0.26168
2	−0.03872	0.0466
3	0.00027	−0.18447
4	−0.01479	−0.16981
5	−0.0009	−0.27784
6	0.01163	−0.31775
7	−0.02061	0.16202
8	−0.11579	−0.20999

Table 3. Correlation of Scale to Height by Region and Sex

Generally the negative correlation is preserved, but in many regions the correlation is not significant. For men, the correlation is almost never significant, and it is generally less negative than for women. Similar kinds of sex differences were observed by Blandon [2] and it was suggested to be a consequence of variant cultural norms.

There are 15 vowels identified and annotated in the TIMIT database and 12 of these correspond to the vowels of the Western Michigan Vowel database. To compare the entire confusion matrix of 15×12 statistically would be complex and unintuitive. Instead a measure of the accuracy of the vowel identification can be assessed by focusing on the distinct vowels [i] (eve) and [I] (it). There were $l2563$ instances of these vowels observed in the TIMIT database. These were classified based on minimum Mahalanobis distance. Before scaling, 5973 were identified as being one of these two vowels, while after optimal scaling 6060 were correctly identified. This corresponds to an error rate of 52.46% before and 51.76% after. It is not possible to assess the statistical significance of this minor improvement without a detailed model of the random error of classification.

8. CONCLUSIONS

We have proposed and tested a scale model, and found that the data strongly support the model. The model implies an algorithm for scale estimation which has been successfully applied to real speech data.

The inverse of the scale is implicitly assumed to be a measure proportional to a hypothetical natural vocal tract length.

The correlation between inverse scale and height is observed to be more significant in women than in men, just as the correlation of scale to fundamental frequency is more significant in women. This is also consistent with the previous finding that formant frequencies and height correlate oppositely for women than for men [9, 10], which may be explained by the hormonally induced divergence in the laryngeal growth in males [7]. It has also been suggested that social influences may exist that may encourage learned gender differences in formant frequencies.

But, these results cannot be interpreted reliably because of the regional variation that was found. The formant frequencies used in this algorithm were taken from Hillenbrand [6] which most likely consists of North Midland speakers. Modeling dialect region as an unknown speaker parameter is required, but was not incorporated into this algorithm because of the increased complexity, primarily the training complexity.

9. REFERENCES

[1] Bachorowski, J., and M.J. Owren, "Acoustic correlates of talker sex and individual talker identity are present in short vowel segment produced in running speech," J. Acoust. Soc. Am., 106 (2), pp. 1054-1063, 1999.

[2] Bladon, A., "Acoustic phonetics, auditory phonetics, speaker sex and speech recognition: A thread." In Computer Speech Processing, pp. 29-38, edited by F. Fallside and A. Woods, Prentice-Hall, Englewood Cliffs, NJ,

[3] Childers, D.G, and Ke Wu, "Gender recognition from speech. Part II: Fine analysis," J. Acoust. Soc. Am., 90 (4), pp. 1841-1856, 1991.

[4] Rabiner, L.R., and R.W. Schafer, Digital Processing of Speech Signals. Prentice-Hall, Englewood Cliffs, NUJ 1978.

[5] Fitch, W.T.,"Vocal tract length and formnat frequency dispersion correlate with body size in Rhesus Makaques,"J. Accoust. Soc. Am102(2), ,pp 1213-1222, Aug. 1997.

[6] Hillenbrand, J.; Getty,L.A.; Clark, J.M.; Wheeler, K., "Acoustic charactaristics of American English vowels," J. Acoust. Soc. Am., 97 (5) Pt. 1, pp. 3099-3111, 1995.

[7] Pressman and Keleman's physiology of the larynx. American Academy of Opthalmology and Otolaryngology.

[8] Kunzel, "How well does average fundamental frequency correlate with speaker height and weight?" Phonetica, 46, 117-125., 1989

[9] Lass, N.J., and M. Davis, "An investigation of speaker height and weight identification," J. Acoust. Soc. Am., 60 (3), pp. 700-703, 1976.

[10] Lass, N.J., and W.S. Brown, "Correlational study of speakers' heights, weights, body surface areas, and speaking fundamental frequencies," J. Acoust. Soc. Am., 63 (4), pp. 1218-1220, 1978.

[11] Dempster, A.P., Laird, N.M., and Rubin, D.B., "Maximum likelihood from incomplete data via the EM Algorithm," J. Roy. Stat. Soc., 39(1), 1-38, 1977.

[12] Markel, J.D., and H.A. Gray, Hr., Linear Prediction of Speech. Springer-Verlag, 1982.

[13] Nelson, D.J., Smith, D.C., Richman, D.J., Townsend, J.L.,"Biometric Speaker Classification," Proc. SPIE Conf. on Adv. Sig. Alg., San Diego, 2000.

[14] DARPA NTIMIT Acoustic Phonetic Database, available through Lingusitic Data Consortium, 1990.

[15] Perry,THis is the paper on child speech, get if from Doug. ,2001

[16] Umesh, S. Cohen, L. Marinovic, N. and Nelson, D. "Frequency-Warping in Speech," in Proc. ICSLP-96, Philadelphia, 1996

[17] Umesh, S. Cohen, and Nelson, D. "Improved Scale-Cepstral Analysis in Speech," in Proc. IEEE ICASSP-98, Seattle, 1998.

[18] Wu, Ke, and D.G. Childers, "Gender recognition from speech. Part I: Course analysis," J. Acoust. Soc. Am., 90 (4), pp. 1828-1840, 1991.