Estimation of the Vocal Tract Length of Vowel Sounds based on the Frequency of the Significant Spectral Valley

T V Ananthapadmanabha[1] *and A G Ramakrishnan*[2]

[1]Voice and Speech Systems, Malleswaram, Bangalore 560003, India
[2]Department of Electrical Engineering, Indian Institute of Science, Bangalore 560012, India

tva.blr@gmail.com, agr@iisc.ac.in

Abstract

Estimating the vocal tract length (VTL), given the acoustic signal of a vowel sound, is an important problem, which is useful in speaker normalization for vowel recognition, in the inversion problem and in acoustic-phonetic studies. The common approach of using the formant data to estimate VTL works for a neutral vowel approximating a uniform tube. However, for natural vowels, formant data shift considerably away from the resonant frequencies of a uniform tube. The proposed method is motivated from these observations: (a) the frequency of a spectral valley, F_v, depends inversely on VTL; (b) there is much smaller shift in F_v, across vowels, from the corresponding valley frequency of a uniform tube; (c) F_v can be estimated from the spectral envelope itself. VTL has been estimated for the Peterson and Barney (33 male and 28 female speakers) and the TIMIT (326 male and 136 female speakers) databases. When the estimated F_v is used for normalization, the spread in the formant data due to gender differences is considerably reduced. The normalization procedure is vowel and speaker intrinsic. Additionally, we report applications such as front/back classification, gender recognition and phonetic feature mapping.

Index Terms: vocal tract length, speaker normalization, spectral valley, front/back classification, gender recognition.

1. Introduction

Many of the existing techniques for the estimation of vocal tract length (VTL) are based on formant data. For natural vowels, the formant frequencies shift substantially away from those of a uniform tube (500, 1500, 2500 Hz etc for a VTL of 17 cm). For example, for vowel /u/, F_2 shifts downwards by about 60% (from 1500 to 600 Hz) and for vowel /i/, F_2 shifts upwards by about 50% (from 1500 to 2250 Hz). The motivation for the proposed method is the fact that the frequency of a spectral valley or the valley frequency (a) varies inversely as VTL; (b) shifts only marginally (by less than 10%) from the corresponding value of a uniform tube (1000 or 2000 Hz etc); and (c) can be estimated from the spectral envelope itself, without the need to estimate the formant frequencies. We present some applications based on the estimated valley frequency. Spectral valley has been shown to play a significant role in a couple of our previous studies as well [1, 2].

The formant values of vowel sounds are strongly influenced by the VTL, which in turn depends on the speaker's gender and age [3, 4]. Anatomically, the average VTL of female (child) speakers is considerably shorter than that of male speakers. Also, VTL is different for different vowels [5]. Using an x-ray based study of an adult male speaker, Fant reported a variation of VTL (16.5 to 19.5 cm) for different vowels.

Given the acoustic signal of a vowel, the estimation of VTL is a problem that has received considerable attention. For a uniform tube (neutral vowel), formant frequencies are inversely proportional to VTL and are regularly spaced (odd integer multiples). For example, for a uniform tube of VTL = 17 cm, formant frequencies are at 500, 1500, 2500 Hz, etc. with a regular spacing of 1000 Hz [6]. For natural vowels, the formant data, especially the first two formant frequencies, F_1 and F_2, shift away from the values of a neutral vowel. The challenge is to estimate VTL, given the formant data of natural vowels. An early method proposed the use of both the poles and zeros of lip impedance [7]. This requires the measurement of lip impedance and is not suitable for the estimation of VTL from a speech signal. Another approach is to use higher formants, F_3, F_4, F_5 etc. [8, 9]. This approach is known to give errors of the order of 5 to 15% [10]. Wakita [10] proposes an analysis-by-synthesis procedure of constructing log area function of different VTLs using the formant frequencies and bandwidths estimated based on linear prediction technique, so as to minimize the mean squared log area under the constraint that the mean log area be zero. An accuracy of estimation of VTL of 1.6 to 8% has been reported on a small set of vowels spoken by a few adult male speakers. This method relies on the accuracy of estimation of formant frequencies and bandwidths, which is a challenging task especially for female speakers. Surprisingly, the method has a dependence on the bandwidths, which in turn depend on the acoustic losses [6] that have nothing to do with the VTL.

A statistical, data-driven approach was proposed by Kirlin [11]. Lammert and Narayanan [12] proposed a method to estimate VTL based on the deviations of formant data from those of a uniform tube. For a uniform tube, the n-th formant frequency, F_{nu}, is an odd integer (2n-1) multiple of the first formant frequency, F_{1u}. Thus $F_{nu}/(2n-1)$ of all the formants would be the same, as an ideal case. Alternately, the mean of $F_n/(2n-1)$, $n = 1$ to m is an estimate of F_{1u}. For natural vowels, the formant frequencies (F_n) differ significantly from F_{nu}, the formant frequencies of a uniform tube of the same length as that of speaker's VTL. Hence, Lammert and Narayanan [12] proposed a weighting function β_n in addition to the factor 1/(2n-1), i.e., $\beta_n F_n$/(2n-1), whose mean value is an estimate of F_{1u} and detetmined the optimum weights β_n using a data-driven approach. The higher formants get a greater weightage. This method has been validated for simulated vocal tracts and against the anatomical data of five speakers. To our knowledge, the estimation of VTL for a large number of speakers has not been previously reported.

The estimation of VTL is useful in minimizing the differences in acoustic data arising due to gender and age of speakers, especially for the task of vowel recognition. An alternative to using VTL is to warp the spectral envelope [13, 14], either linearly or nonlinearly, to that of a reference template so as to minimize the overall error in automatic speech recognition. In addition, the estimation of VTL is required in solving the in-

version problem, namely, the estimation of VT shape from an acoustic signal [15, 16] and in acoustic-phonetic studies [17].

The paper is organized as follows. In Sec. II., we define the significant spectral valley and a procedure for determining its frequency from the short-time spectral envelope of a vowel. We relate this valley frequency to VTL. Applications such as speaker normalization, front/back classification of vowels and gender recognition are presented in Sec.III. In Sec. IV, we summarize the findings and muse over some future directions.

2. The Frequency of Significant Spectral Valley and the Vocal Tract Length

2.1. Significant Spectral Valley, SSV

Between every pair of formants in the short-time spectral envelope of a vowel sound, there exists a spectral valley. The frequency (F_v) of the deepest or the most significant spectral valley (SSV) is used to estimate VTL (See Fig.1). Occasionally, when two adjacent formants have large bandwidths, the two peaks may merge to show a single broad peak and the valley in-between may not be discernible. However, F_v can be unambiguously identified from the spectral envelope itself, without an explicit knowledge of the formant frequencies. It is expected that F_v lies in the range of 800 to 2600 Hz and the level of the SSV is below the mean spectral level. In case there are two contenders for the SSV, the valley possessing the largest spectral level difference with its immediate neighbouring (either to the left or right) spectral peaks is taken as the SSV.

2.2. Estimation of VTL for the P and B Formant Data

We use the first three formant frequencies published by Peterson and Barney [3], henceforth abbreviated as P&B data [18]. Histograms of F_1 and F_2 for all the vowels (except the retroflex vowel /ɝ/) for both repetitions of 33 male (594 samples) and 28 female speakers (504 samples) are shown in Fig.2. The mean and the standard deviation (SD) of F_1 for male (female) speakers are 500 Hz and 32% of the mean (586 Hz and 35% of the mean) and those of F_2 are 1430 Hz and 36% of the mean (1700 Hz and 40% of the mean), respectively (see Table 1). In the histograms, we note a multi-modal distribution with a very wide spread of samples.

In order to estimate the valley frequency, F_v, given only the formant data, we compute the frequency response. For male (female) speakers, we use a constant fourth formant at 3500 Hz (4200 Hz) and compute the frequency response using a sampling frequency of 8000 Hz (10000 Hz). The valley frequency is determined from the computed frequency response. Histograms of F_v for male and female speakers are shown in Fig. 3a. Histograms for both genders show a bi-modal distribution with sharp peaks. It has been ascertained that the two modes correspond to the front and back vowels, respectively.

For front vowels of male (female) speakers, F_v happens to be the first spectral valley, F_{1v}. The mean and S.D. of F_{1v} for male (female) speakers are about 1153 Hz and 7% of the mean (1427 Hz and 11% of the mean), respectively. For back vowels of male (female) speakers, F_v happens to be the second spectral valley, F_{2v}. The mean and S.D. of F_{2v} for male (female) speakers are about 1795 Hz and 6% of the mean, (2112 Hz and 5.7% of the mean), respectively (see Table 1). This shows that the spread in the valley frequencies is much lower than that in the formant frequencies.

The reported ratio of the average physical VTL of female

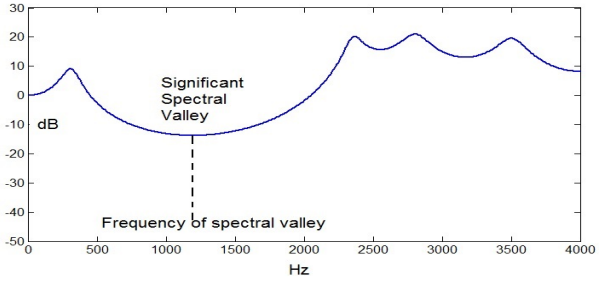

Figure 1: *A typical frequency response of vowel /i/ showing the locations of the formant frequencies and the significant spectral valley.*

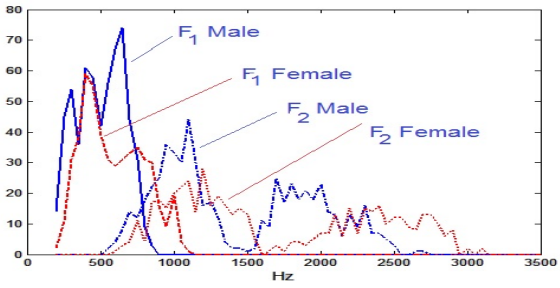

Figure 2: *Histograms of F_1 and F_2 of P&B data for male and female speakers.*

speakers to that of male speakers is about 0.8 [4]. Based on the mean of F_v, the estimated ratio of average VTLs of female to male speakers for P&B data is 1153/1427=0.8 for front vowels and 1795/2112=0.85 for back vowels, which match well with the published anatomical data.

The expected valley frequency is 1000 Hz (1200 Hz) for a uniform tube representative of an adult male (female) speaker with VTL of about 17 cm (14 cm). For front vowels of male (female) speakers, the mean value of F_{1v} is 1153 Hz (1427 Hz) implying VTL is shorter than 17 cm (14 cm). For back vowels of male (female) speakers, the mean value of $0.5F_{2v}$ is about 900 Hz (1050 Hz), implying that the VTL is longer than 17 cm (14 cm). This deviation from the expected VTL may be explained as follows.

We postulate that the effective VTL from the acoustic point of view may differ from the physical VTL from the anatomical point of view. As per the x-ray data of an adult male Russian speaker [5], the VTL of vowels /i/ and /e/ is about 16.5 cm. For front vowels, there is an abrupt increase in the VT area immediately in front of the constriction. It is known that for vowel /i/, the first three formant frequencies are insensitive to large changes of mouth area [19]. The radiation of acoustic waves into free-field may be assumed to begin at the exit of the constriction itself, instead of at the lips. Hence, for front vowels, the acoustic VTL may be assumed to be from the glottis to the exit of constriction, which is much shorter than the physical VTL. As per the x-ray data [5], the VTL of vowel /ɑ/ is about 17 cm, whereas that of /u/ is about 19.5 cm for an adult male speaker. Hence, the mean VTL of back vowels is greater than 17 cm (14 cm for female speakers). Further, due to end correction for the radiation, the acoustic VTL of a back rounded vowel would be longer than the physical VTL. Some researchers [12] have suggested an ad-hoc correction in order to match the estimated acoustic VTL to the physical VTL.

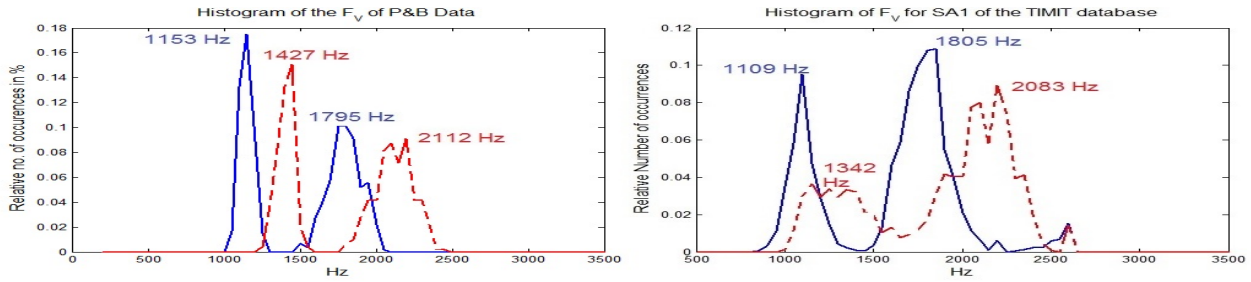

Figure 3: *Histogram of F_v for male (solid) and female (dashed) speakers computed using the (a) P&B data (594 male and 504 female samples); (b) SA1 utterances (978 male and 408 female vowel samples) of the TIMIT database.*

2.3. Estimation of VTL for the TIMIT Database

In the case of P&B data, formant frequencies have been measured over the mid steady part of the vowels in the hVd context. In order to study the influence of context on the estimation of F_v, we consider the TIMIT training set of SA1 sentences comprising 326 male and 136 female speakers. In the utterance, "She had your dark suit in greasy wash water all year", the entire segment of vowels labelled 'iy', 'aa' and 'ao' are analyzed. A frame-wise linear prediction (LP) analysis of order 18 is performed on the pre-emphasized and windowed frames of 20 ms duration, at 100 frames per second. The histogram of F_v, determined from the log spectrum of the all-pole model, is shown in Fig. 3b. The mean values for the TIMIT database (see Table 1) and those of P&B data match very well, thus showing the consistency of the proposed method.

2.4. Robustness of the VTL Estimate for the P&B Data

White Gaussian random noise is added to the impulse response of a vowel synthesized using the P&B data to obtain 20 dB SNR. LP analysis is performed on the noisy signal and F_v is estimated. A similar experiment is conducted with the addition of babble noise also. The results for the noisy data, shown in Table 1, are very similar to those obtained for the clean (P&B) data. The analysis of telephone quality speech will not be an issue, since F_v is expected to lie in the range of 800-2600 Hz.

3. Applications of Valley Frequency

3.1. Normalization of P&B Formant Data

The estimated valley frequency, which is inversely related to the acoustic VTL, serves the purpose of speaker normalization. If F_v is closer to 1000 rather than 2000 Hz, then it is the first valley, F_{1v}; else, it is the second valley F_{2v}. Accordingly, we define the normalization frequency, V_n as,

$$V_n = 2F_{1v} \tag{1}$$

OR

$$V_n = F_{2v} \tag{2}$$

depending upon which of the valleys happen to be the deepest. All the formant frequencies are normalized by one and the same factor as (F_1/V_n, F_2/V_n). A plot of F_2 versus F_1 of P&B data is shown in Fig. 4a. A plot of normalized second formant frequency versus normalized first formant frequency is shown in Fig. 4b. Here, the dimensionless normalized values are multiplied by 2200 for the purpose of graphic comparison with the plot shown in Fig. 3a. We observe that gender differences are normalized effectively, especially for front vowels. Since V_n is estimated frame-wise for each vowel of each speaker, the normalization procedure is speaker and vowel intrinsic (self-normalization) [17].

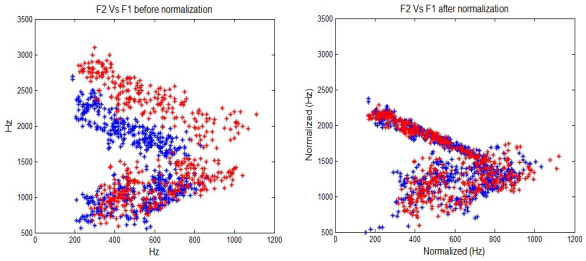

Figure 4: *Normalization of formant data using the frequency of significant spectral valley. Left: Original raw data of Peterson and Barney. Right: Formant data normalized using F_n. Male (blue) and female (red) data overlap and intra-speaker spread is reduced after the normalization.*

Graphically, the normalization appears to be as effective as extrinsic normalization procedures (z-score or S-ratio) [17]. When the P&B data of all male and female speakers are pooled together, using F_1, F_2 (in Hz) as the feature vector and Euclidean distance measure, a vowel classification accuracy of about 67.2% and 75.7% are obtained for the raw data and normalized formant data, respectively. A formant dependent normalization procedure gives a better clustering and a higher vowel classification accuracy [20].

3.2. Front/Back Classification

We propose to make use of F_v to determine if a vowel belongs to the Front or Back class. The histograms of the front vowels of female speakers and the back vowels of male speakers cross around 1500 Hz (Fig. 3). In other words, the two modes representing the front and back vowels show a clear separation around 1500 Hz, both for the P&B data and the SA1 samples of the TIMIT database. For any other database, the actual frequency to be used (in place of 1500 Hz) can be determined using a training set, if desired. We define normalized F_v, NFV, as

$$NFV = (F_v - 1500)/1500 \tag{3}$$

For front vowels, NFV is zero or negative, since in the histogram of F_v, almost all samples of front vowels lie to the left of 1500 Hz. For back vowels, NFV is positive. Using a threshold of 0 on NFV, we obtain a Front/Back classification accuracy of 98.5% for the P&B data. For the TIMIT database considered in Sec. 2.3, Front/Back classification accuracy of 96.1% has been obtained for 7861 frames of front vowels and 19730 frames of back vowels.

3.3. Gender Recognition

Since V_n is inversely related to VTL, it may be used for gender recognition. The accuracy of gender recognition is a measure to assess the performance of the proposed method. The gender recognition rule is different for front and back vowels. From Fig. 3a, we note that the histograms for the front (back) vowels of male and female speakers cross around 1300 Hz (2000 Hz). For front vowels, if $F_v \leq 1300$ Hz, the sample is assigned to a male speaker; else, to a female speaker. For back vowels, if $F_v \leq 2000$ Hz, the sample is assigned to a male speaker; else, to a female speaker. Using such a procedure along with the Front/Back classification as previously described in Sec. 3.2, we obtain a gender recognition accuracy of 93.7% for the P&B data. Most errors occur for rounded vowels. If we exclude rounded vowels, a gender recognition accuracy of 97% is obtained. A lower accuracy might have arisen since gender recognition itself is based on Front/Back classification. Also, there may be some male (female) speakers with VTL shorter (longer) than the average VTL of male (female) population. The threshold frequencies of 1300 and 2000 Hz may be fine tuned, if required, for a new database or a new population.

Using the same thresholds, a gender recognition accuracy of 83.3% is obtained for the TIMIT database, comprising 326 male and 136 female speakers. This is a decent result, considering the large number of speakers. The lower accuracy arises because of (i) significant overlap of the histograms of F_v (Fig. 2b) arising due to the contextual influence and (ii) the dependence on the Front/Back classification accuracy. A previous study on VTL estimation using higher formants reports an accuracy in the range of 25 to 37% depending on the gender and context [9]. A vowel-dependent gender recognition accuracy of 98.2% has been reported [21] using the second formant frequency for the P&B data. However, this study is not related to the estimation of VTL.

3.4. Normalized Phonetic Feature Space

Generally, the acoustic signal of a vowel is represented in F_2 versus F_1 space, as in Fig. 4a. We propose an alternative representation of vowels corresponding to the phonetic feature space of High/Low versus Front/Back. Assuming the VTL as 17 cm (for an adult male speaker), we note that F_1<500 Hz for high vowels /i/ and /u/ and F_1>500 Hz for low vowel /ɑ/ and F_2>1500 Hz for front vowel /i/, whereas F_2<1500 Hz for back vowels /ɑ/ and /u/. Generalizing this observation, we postulate that $F_1 < F_{1u}$ for high vowels and vice-versa; $F_2 > F_{2u}$ or $3F_{1u}$ for front vowels and vice-versa. Further, note that F_{1u}=$0.5F_{1v}$ or $0.25F_{2v}$. We define normalized, dimensionless variables representative of High/Low and Front/Back phonetic features as

$$HL = 1 - F_1/F_{1u} = 1 - F_1/(0.5F_{1v}) = 1 - F_1/(0.25F_{2v}) \quad (4)$$

$$FB = 1 - F_2/3F_{1u} = 1 - F_2/(1.5F_{1v}) = 1 - F_2/(0.75F_{2v}) \quad (5)$$

A plot of HL versus FB for the P&B formant data is shown in Fig. 5. This resembles the rotated F_2 versus F_1 space used by phoneticians and linguists [17]. Since F_v is measured for each vowel and each speaker, the samples of male and female speakers are normalized. The three corner vowels occupy relatively the correct positions in the phonetic feature space. Vowel /i/ occupies the top left corner corresponding to high and front phonetic features. Vowel /ɑ/ occupies the bottom central position corresponding to low and back phonetic features. Vowel /u/ occupies the top right corner corresponding to high and back

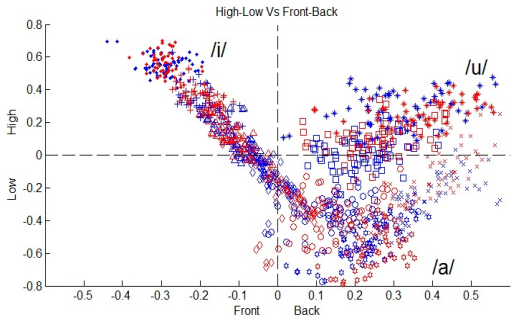

Figure 5: *Formant data of Peterson and Barney mapped to a normalized phonetic feature space.*

phonetic features. There seems to a bias in the estimated FB value and FB>0.1 may have to be considered as 'Back' based on the distribution seen for vowel /ae/.

Table 1: *The mean value in Hz and the standard deviation (SD) as a percentage of the mean for various data sets. P&B: Peterson and Barney data. SA1: The TIMIT database. WN-20 (WN-0): White noise, SNR=20 (0) dB. BN-20 (BN-0): Babble Noise, SNR=20 (0) dB.*

	Male		*Female*	
Data	*Mean*	*SD(%)*	*Mean*	*SD(%)*
$\mathbf{F_1}$ — *P&B*	500	32.0	586	35.0
$\mathbf{F_{1v}}$— *P&B*	1153	7.0	1427	11.0
$\mathbf{F_{1v}}$— *SA1*	1109	6.7	1342	13.2
$\mathbf{F_{1v}}$— WN-20	1180	15.0	1436	14.2
$\mathbf{F_{1v}}$— WN-0	1376	24.5	1614	20.7
$\mathbf{F_{1v}}$— BN-20	1241	12.5	1513	13.0
$\mathbf{F_{1v}}$— BN-0	1512	18.4	1760	13.6
$\mathbf{F_2}$ — *P&B*	1430	36.0	1700	40.0
$\mathbf{F_{2v}}$ — *P&B*	1795	6.0	2112	5.7
$\mathbf{F_{2v}}$ — *SA1*	1805	13.2	2083	12.4
$\mathbf{F_{2v}}$— WN-20	1735	8.0	2034	7.2
$\mathbf{F_{2v}}$— WN- 0	1671	12.1	1963	12.8
$\mathbf{F_{2v}}$— BN-20	1800	6.7	2116	6.3
$\mathbf{F_{2v}}$— BN- 0	1899	8.5	2299	7.89

4. Conclusion

We have proposed a method for estimating VTL using the frequency of significant spectral valley instead of the formant frequencies. The proposed method predicts distinctly different VTLs for the front and back vowels, a finding difficult to infer using formant based methods. Also, we have argued that the acoustic VTL is different from the anatomical VTL. For the formant data published by Peterson and Barney, we have demonstrated the effectiveness of the speaker normalization. Since VTL is estimated for each frame of a vowel sample, the normalization procedure is speaker as well as vowel intrinsic. We have applied the method for Front/Back classification, gender recognition and a procedure to map formant data into a normalized phonetic feature space.

Future work involves validating the proposed method and applying the method on a larger database. The frequency of the significant spectral valley may be used as an anchor for spectral warping procedures. Methods proposed in the literature for estimating VTL based on the formant data can as well be extended to the estimation of VTL using the frequency of spectral valley.

5. References

[1] T. V. Ananthapadmanabha, A. G. Ramakrishnan, and S. Sharma, *Significance of the levels of spectral valleys with application to front/back distinction of vowel sounds*. https://arxiv.org/abs/1506.04828, 2015.

[2] ——, "An objective critical distance measure based on the relative level of spectral valley," *Proc. Interspeech 2017, Stockholm, Sweden*, pp. 641–644, 2017.

[3] G. Peterson and H. Barney, "Control methods used in a study of the vowels," *Journal of Acoustical Society of America*, vol. 24, no. 2, pp. 175–184, 1952.

[4] G. Fant, *Speech sounds and features.* The MIT Press, 1973.

[5] ——, *Acoustic Theory of Speech Production,*. Hague: Mouton, 1960.

[6] J. L. Flanagan, *Speech analysis synthesis and perception*. Springer Science & Business Media, 2013.

[7] A. Paige and V. Zue, "Calculation of vocal tract length," *IEEE Transactions on audio and electroacoustics*, vol. 18, no. 3, pp. 268–270, 1970.

[8] T. Claes, I. Dologlou, L. ten Bosch, and D. Van Compernolle, "A novel feature transformation for vocal tract length normalization in automatic speech recognition," *IEEE Trans. Speech and Audio Processing*, vol. 6, no. 6, pp. 549–557, 1998.

[9] V. Sorokin and I. Geraskin, "Vocal-tract length estimation," *Journal of Communications Technology and Electronics*, vol. 58, no. 12, pp. 1292–1301, 2013.

[10] H. Wakita, "Normalization of vowels by vocal-tract length and its application to vowel identification," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 25, no. 2, pp. 183–192, 1977.

[11] R. Kirlin, "A posteriori estimation of vocal tract length," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 26, no. 6, pp. 571–574, 1978.

[12] A. C. Lammert and S. S. Narayanan, "On short-time estimation of vocal tract length from formant frequencies," *PloS one*, vol. 10, no. 7, 2015.

[13] L. Lee and R. C. Rose, "Speaker normalization using efficient frequency warping procedures," in *Acoustics, Speech, and Signal Processing, ICASSP, Proceedings IEEE International Conference on*, vol. 1, pp. 353–356.

[14] L. F. Uebel and P. C. Woodland, "An investigation into vocal tract length normalisation," in *Sixth European Conference on Speech Communication and Technology*, 1999.

[15] P. Mermelstein, "Determination of the vocal-tract shape from measured formant frequencies," *The Journal of the Acoustical Society of America*, vol. 41, no. 5, pp. 1283–1294, 1967.

[16] M. R. Schroeder, "Determination of the geometry of the human vocal tract by acoustic measurements," *The Journal of the Acoustical Society of America*, vol. 41, no. 4B, pp. 1002–1010, 1967.

[17] A. H. Fabricius, D. Watt, and D. E. Johnson, "A comparison of three speaker-intrinsic vowel formant frequency normalization algorithms for sociophonetics," *Language Variation and Change*, vol. 21, no. 3, pp. 413–435, 2009.

[18] Peterson and Barney, *Vowel formant frequency database*. http://www.cs.cmu.edu/Groups/AI/areas/speech/database/pb/, Last visited 14 March, 2018.

[19] B. S. Atal, J. J. Chang, M. V. Mathews, and J. W. Tukey, "Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting , technique," *The Journal of the Acoustical Society of America*, vol. 63, no. 5, pp. 1535–1555, 1978.

[20] T. V. Ananthapadmanabha and A. G. Ramakrishnan, "Intrinsic-cum-extrinsic normalization of formant data of vowels," *The Journal of the Acoustical Society of America*, vol. 140, no. 5, pp. EL446–EL451, 2016.

[21] D. G. Childers and K. Wu, "Gender recognition from speech. part ii: Fine analysis," *The Journal of the Acoustical society of America*, vol. 90, no. 4, pp. 1841–1856, 1991.