Estimation of Place of Articulation of English Fricatives Using the Modified Dominant Spectral Centroid and Slope as the Spectral Parameters

K. S. Nataraj, Prem C. Pandey, and Hirak Dasgupta

Dept. of Electrical Engineering, Indian Institute of Technology Bombay, Mumbai 400076, India

natarajks@ee.iitb.ac.in, pcpandey@ee.iitb.ac.in, hirakdgpt@ee.iitb.ac.in

Abstract

Visual feedback of the place of articulation, the place of maximum constriction in the oral cavity, can be useful for speech training of the hearing-impaired children and second language learners. This application requires a speaker-independent estimation of the place of articulation from the speech signal. For fricatives, which are produced by excitation of the oral cavity by the turbulence at the place of maximum constriction, the place perception is reported to be primarily related to the spectral characteristics of the frication. In our earlier work, place of articulation of fricative utterances in the XRMB database were estimated using spectral moments and slope parameters. In this work, modification of the earlier reported spectral parameters is presented. Results show the reduction in the RMS error by 2.1 mm in the estimated values of place of articulation using the proposed spectral parameters compared to the earlier ones.

Index Terms: speech training, spectral characteristics, fricatives, place of articulation.

1. Introduction

Persons with hearing impairement have great difficulty in acquiring correct articulation of speech sounds due to lack of auditory feedback. Non-auditory feedback such as displaying the short-time energy, spectral features, voicing etc. can contribute to their learning of correct pronunciation. It has been reported that computer-based speech training aids, displaying the vocal tract shape during speech-production, can be used for improving the vowel articulation of hearing-impaired [1]-[2]. The visual feedback can also be used to guide the second language learners to adjust the articulators to improve their pronunciation instead of displaying acoustic differences or pronunciation score [3].

For providing visual feedback, vocal tract shape or articulatory information needs to be estimated from the speech signal. Estimation of vocal tract shape is an ill-posed problem due to non-uniqueness of acoustic-to-articulatory mapping. LPC-based methods are commonly used for estimation of the vocal tract shape for articulation training of vowels, diphthongs, and semivowels; as the vocal tract filter during these sounds can be modeled as an all-pole filter [1]. LPC-based methods fail during stop-closures due to lack of spectral information [4]. Pandey and Shah [5] reported a method for estimating the vocal tract shape during stop closures of vowel-consonant-vowel (VCV) utterances by interpolating the LPC-based estimates of shapes during the transition segments preceding and following the stop closure and release burst. However, the LPC-based methods cannot be used during nasals and fricatives due to spectral zeros in the vocal tract filter model.

Earlier methods proposed to estimate the vocal tract shape during fricatives are based on analysis-by-synthesis and data-driven machine learning. In the analysis-by-synthesis methods [6]-[8], articulatory parameters are estimated by minimizing the spectral distance between the synthesized speech signal and the input speech signal. These methods suffer from the problem of optimization converging to local minima leading to erroneous estimates [7]. Machine learning methods [9]-[10] are reported to work well for speaker-dependent estimation of articulatory parameters. However, speech-training applications require speaker-independent mapping.

Fricatives are produced by forming a narrow constriction in the oral cavity, which divides the oral cavity into front and back cavities. During frication, a turbulence is created by obstruction of the steady airflow due to the narrow constriction. English fricatives have four places of articulation: (i) labio-dental (/f/ as in "fine", /v/ as in "vine"), (ii) linguo-dental (/θ/ as in "thing", /δ/ as in "then"), (iii) alveolar (/s/ as in "sue", /z/ as in "zoo"), (iv) palatal (/ʃ/ as in "shoe", /ʒ/ as in "measure") [11]-[12].

Several studies have been reported on relating the place of articulation of fricatives to acoustic characteristics including duration, energy, and spectral characteristics [13]-[15]. Jongman *et al*. [13] showed that spectral moments and spectral peak location could be used to distinguish four places of articulation of fricatives. The spectral moments capture the spectral shape while spectral peak location is related to the length of the front cavity. Ali *et. al.* [15] used the spectral slope parameter to distinguish labial fricatives from the sibilant ones. These studies relate the acoustic characteristics to the categorical place of articulation and not to the place of articulation on a continuum as needed for the visual feedback for speech training.

An earlier investigation [16], using a speaker-independent mapping, based on a Gaussian mixture model (GMM), of spectral parameters during frication to the place of articulation showed the dominant spectral centroid and the spectral slope to be related to place of articulation, but with significant errors in the estimated values of place of articulation of labio-dental and palatal fricatives. The error may be attributed to the inadequacy of the spectral parameters to capture the spectral information related to the place of articulation and/or lack of vocal tract length normalization.

In this paper, an investigation is presented to improve the estimation of the place of articulation, by using two modifications on the earlier reported dominant spectral centroid and slope parameters. The second section presents the spectral parameters used in the investigation. The third section describes the material and method for the evaluation. The results are presented in the fourth section, followed by a conclusion in the last section.

2. Estimation of spectral parameters

Two spectral parameters namely spectral centroid and spectral slope are estimated. The spectral centroid is related to the length of the front cavity and thus captures the place of articulation of alveolar and palatal fricatives having different front cavity lengths. The place of articulation of labial fricatives is near the lips and thus does not have large variations in the spectral centroid. As the labial fricatives have spectral energy concentration at low frequencies, the spectral centroid can be used to separate them from alveolar fricatives which have energy concentrated in the high-frequency region. Spectral slope parameter is used to separate the palatal fricatives from labial fricatives as the labial fricatives have a relatively low spectral slope even though both have energy concentration in the low-frequency region.

Investigations were carried out using the utterances involving the voiced fricatives /v, z, ʒ/ and the unvoiced fricatives /f, s, ʃ/. The magnitude spectrum, for speech with a sampling frequency (f_s) of 16 kHz was calculated using a window length of 30 ms with 5 ms shift and FFT size L = 512. The average magnitude spectrum $S(k)$ was calculated using the magnitude spectrum values over the central one-third segment of each fricative utterance. Computation of the spectral parameters is described in the following two subsections.

2.1. Spectral centroid using maximum sum subarray

It has been reported that spectral moments are related to place of articulation in case of fricatives [13]. It has been shown in [16] that spectral centroid is not effective in discriminating between the labial and alveolar fricatives, although these fricatives have energy concentrations at different frequencies. It was further shown in [16] that the dominant spectral centroid (DSC), calculated as the centroid of significant values of the magnitude spectrum obtained using values above the 80-percentile, reduced the effect of insignificant values and resulted in a better discrimination between labial and alveolar fricatives. But, it did not serve as a useful parameter if the dominant spectral samples were far apart on the frequency axis. Critical-band based smoothing reduces this problem but results in excessive smoothing at high frequencies and may distort the corresponding peak locations. To resolve this problem, a contiguous subset of spectral samples with maximum sum is searched along the frequency axis.

To obtain the contiguous maximum sum subset, Kadane's one-pass optimal search algorithm is used [17]-[18]. The Kadane's algorithm is applicable for arrays with bipolar values. It is applied on the median-subtracted spectrum obtained as

$$S_{MS}(k) = S(k) - S_{median} \tag{1}$$

where $S_{median} = \text{median}(S(1), S(2), \ldots, S(L/2))$. The algorithm calculates a local sum $S_{LS}(k)$ and a global sum $S_{GS}(k)$. The global sum is updated whenever a local sum becomes more than the global sum. The algorithm is expressed using the following difference equations:

$$S_{LS}(k) = \max(S_{MS}(k),\ S_{LS}(k-1) + S_{MS}(k)) \tag{2}$$

$$S_{GS}(k) = \max(S_{GS}(k-1),\ S_{LS}(k)) \tag{3}$$

for spectral samples $k \in \{1, 2, \ldots\ldots, L/2\}$. The local sum $S_{LS}(k)$ and the global sum $S_{GS}(k)$ are initialized with the value of the first spectral sample $S_{MS}(1)$. The beginning of the maximum sum subset $i_{BG}(k)$, the end of the maximum sum subset $i_{EG}(k)$, and the beginning of the local subset $i_{BL}(k)$ are updated as

$$i_{BL}(k) = \begin{cases} k, & S_{MS}(k) > S_{LS}(k-1) + S_{MS}(k) \\ i_{BL}(k-1), & \text{otherwise} \end{cases} \tag{4}$$

$$i_{BG}(k) = \begin{cases} i_{BL}(k), & S_{LS}(k) > S_{GS}(k-1) \\ i_{BG}(k-1), & \text{otherwise} \end{cases} \tag{5}$$

$$i_{EG}(k) = \begin{cases} k, & S_{LS}(k) > S_{GS}(k-1) \\ i_{EG}(k-1), & \text{otherwise} \end{cases} \tag{6}$$

where $i_{BL}(k)$ is an auxiliary variable used to obtain $i_{BG}(k)$. The values $i_{BG}(k)$, $i_{EG}(k)$, and $i_{BL}(k)$ are initialized to 1 (the first sample of the array). The final values of $i_{BG}(k)$ and $i_{EG}(k)$ at $k = L/2$ are used as the beginning and end of the maximum sum subset of spectral samples. The modified dominant spectral centroid (MDSC) is calculated using the contiguous spectral samples of the average magnitude spectrum in the maximum subset as

$$\text{MDSC} = \sum_{k=i_{BG}(L/2)}^{i_{EG}(L/2)} kS(k) \Bigg/ \sum_{k=i_{BG}(L/2)}^{i_{EG}(L/2)} S(k) \tag{7}$$

2.2. Spectral slope

Spectral slope parameter is used to separate the palatal and labial fricatives. In [16], the normalized sum of spectral slopes (NSSS) calculated as the sum of the first difference values of the smoothed spectrum and the normalized sum of the positive spectral slope (NSPSS) calculated as a sum of the positive values of the first difference of the smoothed spectrum are used. To improve the consistency of the slope estimation, the spectrum $S(k)$ is smoothed using a two-step median-mean filtering, to suppresses the spurious variations without disturbing the peaks and valleys. The median-mean filtering is similar to that in [16] and is carried out using the samples equal to the critical bandwidth $B(k)$ centered at spectral sample k using [19]

$$B(k) = 25 + 75(1 + 1.4(f(k))^2)^{0.69} \tag{8}$$

where $f(k) = kf_s / L$. The smoothed spectrum is referred to as $S_{CB}(k)$. In the present investigation, a single slope parameter termed as the normalized sum of absolute spectral slopes (NSASS) is calculated as the sum of the absolute values of the first difference of $S_{CB}(k)$ and is given as,

$$\text{NSASS} = \sum_{k=2}^{L/2} |S_{CB}(k) - S_{CB}(k-1)| \Bigg/ \sum_{l=1}^{L/2} S_{CB}(l) \tag{9}$$

This parameter combines the changes in the spectrum in both positive and negative directions, and thus gives a better representation of the spectral flatness compared to the earlier proposed ones.

3. Material and method

In this study, a total of 3133 English fricatives involving voiced fricatives /v/, /z/, /ʒ/ and unvoiced fricatives /f/, /s/, /ʃ/ from 47 speakers available in the XRMB database [20] were used. This database has been developed by the University of Wisconsin, by using a narrow X-ray beam to track and record the motion of 2-3 mm diameter gold pellets attached to the

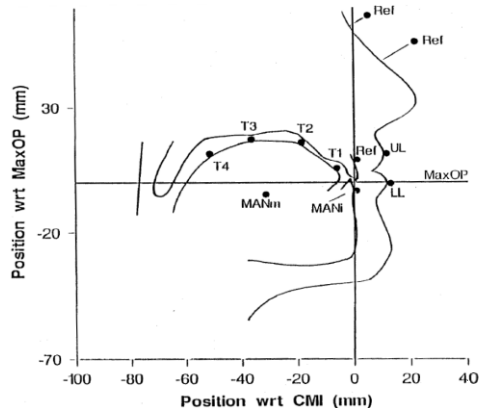

Figure 1: *Position of pellet points in the XRMB database.*

tongue, jaw, lips, and soft palate during speech production. The database consists of recordings of the x-y coordinates of the pellets and speech signal for isolated words, sentences, vowels, and vowel-consonant-vowels for all speakers.

Fig. 1 shows the example of locations of the pellets in the mouth during speech production. There are four pellet points (T1-T4) on the tongue and one pellet each on the lower lip (LL), the upper lip (UL), and mandibular incisor (MNi). The database also provides the palatal outline and posterior pharyngeal wall data. The x-y coordinates of these pellet points are tracked during speech production and are provided with respect to a reference plane. An active oral cavity image can be obtained by connecting, the upper lip with palatal outline forming the upper contour, the lower lip with the incisor and tongue pellet points to form lower contour. The place of articulation (PoA-XRMB) was estimated by a graphical processing of these contours by employing an automatic method proposed in [21]. In this method, an axial curve is estimated such that it divides the oral cavity into two equal parts. The oral cavity openings along the axial curve are obtained by measuring the distance between the two contours along the normals to the axial curve. PoA-XRMB is estimated as the position of the smallest oral cavity opening measured from the lips along the axial curve.

4. Evaluation results

Fricative utterances from 35 randomly selected speakers were used for training and the utterances from the remaining 12 speakers were used for testing, with 2371 and 762 utterances in the training and testing datasets, respectively. A speaker-independent mapping from the spectral parameters to the place of articulation, based on a Gaussian mixture model (GMM), is used to estimate the place of articulation [10].

During training, a joint probability density of the spectral parameters and the PoA-XRMB is modeled using a mixture of 20 Gaussians with full covariance matrices. A joint vector for each utterance in the training data is obtained by concatenating the corresponding spectral parameters and the PoA-XRMB. The GMM parameter set $\boldsymbol{\theta}^{(q)}$ (mean vector, covariance matrices, and Gaussian component weights) of the joint probability density function of these joint vectors are estimated using the expectation-maximization algorithm.

The place of articulation a is estimated for the utterance with a given spectral parameter set $\mathbf{r}$ by maximizing the likelihood of conditional probability density $p(a|\mathbf{r},\boldsymbol{\theta}^{(q)})$, using the GMM parameter set $\boldsymbol{\theta}^{(q)}$ of the joint probability density function as obtained during the training. The estimator is expressed as

$$\hat{a} = \arg\max_{a} p(a|\mathbf{r},\boldsymbol{\theta}^{(q)}) \tag{10}$$

The expectation-maximization algorithm is used to maximize the above likelihood function.

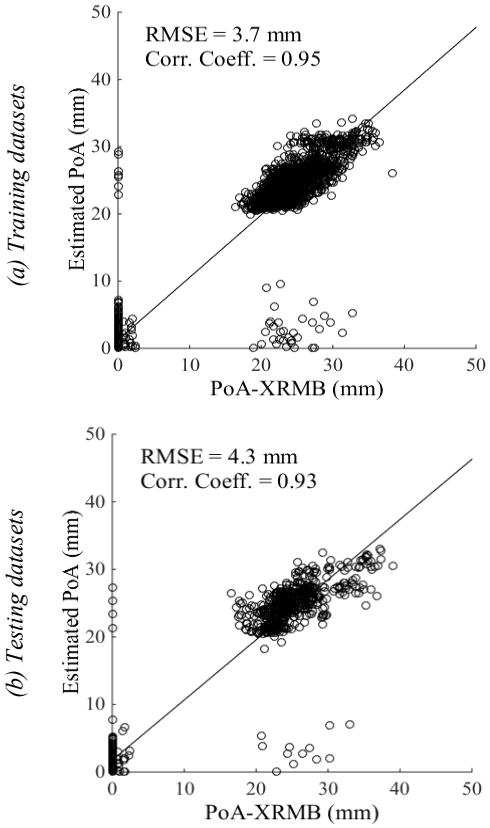

Figure 2: *Scatter plot of estimated place of articulation using the proposed set of spectral parameters versus the graphically obtained ones: (a)training datasets, (b) testing datasets.*

Table 1: *Mean error (μ), standard deviation of errors (σ), and RMS error (RMSE) for different sets of spectral parameters.*

Parameter set	μ (mm)	σ (mm)	RMSE (mm)
[DSC, NSSS, NSPSS]	0.3	6.4	6.4
[MDSC, NSSS, NSPSS]	0.4	5.8	5.8
[DSC, NSASS]	0.5	5.1	5.1
[MDSC, NSASS]	**0.3**	**4.3**	**4.3**

The scatter plot of the estimated place of articulation using the proposed set of spectral parameters versus the graphically obtained place of articulation, PoA-XRMB, for the training and testing datasets are shown in Fig. 2. It can be observed that the error patterns are similar across the two datasets, indicating a good fit of the data. The RMS error for estimation of the place of articulation was 3.7 mm for the training data and 4.3 mm for the testing data.

The performance of the proposed set of parameters was compared with those of the earlier ones. Table 1 shows the

Table 2: *Comparison of estimation of place of articulation for different fricatives using different sets of spectral parameters: (a) DSC, NSSS, NSPSS,* (*b) MDSC, NSSS, NSPSS,* (c*) DSC, NSASS, (d) MDSC, NSASS.*

Frica-tive	No. of utter-ances	PoA-XRMB (mm)		Parameter set							
				[DSC, NSSS, NSPSS]		[MDSC, NSSS, NSPSS]		[DSC, NSASS]		[MDSC, NSASS]	
		Mean	S.D.	Mean error (mm)	S.D. of error (mm)	Mean error (mm)	S.D. of error (mm)	Mean error (mm)	S.D. of error (mm)	Mean error (mm)	S.D. of error (mm)
/f/	137	0.2	0.5	-2.7	8.0	-1.6	5.4	-3.0	7.5	-1.6	3.9
/v/	68	0.1	0.3	-1.9	6.9	-2.3	7.5	-0.8	3.2	0.2	3.4
/s/	218	23.7	2.6	0.8	4.7	0.5	4.1	0.7	4.2	0.2	3.5
/z/	194	23.8	2.5	0.3	4.7	0.1	4.1	0.4	4.0	0.4	4.7
/sh/	136	29.8	3.8	3.0	7.0	3.4	6.9	1.7	3.3	1.8	4.3
/zh/	9	28.9	4.7	7.2	14.2	7.5	14.5	4.4	10.2	4.5	10.0

mean error and the standard deviation of errors as the measures of the bias and inconsistency of the estimation, respectively, with reference to the PoA-XRMB values, for the testing data. The RMS of errors (RMSE) is also given in the table as a composite single measure of error. The estimation bias is negligible in all cases. The parameter set [DSC, NSSS, NSPSS] resulted in the largest error. Compared to it, the set [MDSC, NSSS, NSPSS], i.e. replacement of DSC by MDSC resulted in a decrease in the error by 0.6 mm. The set [DSC, NSASS], i.e. replacement of NSSS and NSPSS by NSASS, resulted in a decrease in error by 1.3 mm. The proposed set [MDSC, NSASS] resulted in a decrease in the error by 2.1 mm.

Table 2. provides the mean error and the standard deviation of errors for different fricatives, for estimation using different sets of spectral parameters. It can be seen that errors are largest for fricatives /f/, /s/, /z/, and /sh/ using the parameter set [DSC, NSSS, NSPSS]. When the DSC is replaced by MDSC without changing the slope parameters NSSS and NSPSS, the errors are reduced for fricative /f/, /s/, /z/, and /sh/, indicating that MDSC is able to better capture the place of articulation as it removes the effect of the spectral samples which are not in continuum on the frequency axis. Use of the parameter set [DSC, NSASS] reduced the errors for all the fricatives, as compared with the set [DSC, NSPSS, NSSS], indicating that use of both the positive and negative slopes in slope calculation captures the slope information more effectively. The lowest error for the parameter set [MDSC, NSASS] for all the fricatives indicates the effectiveness of the proposed set of spectral parameters.

The large errors for the voiced palatals /zh/ may be due to a small number of utterances available in the database. The errors for the alveolar voiced fricatives did not decrease and it can be attributed to the mixed excitation of the voiced fricatives resulting in occurrences of formants in the low-frequency region along with the frication noise. In these cases, the spectral centroid shifts towards a lower value due to large magnitudes of the formants. The large errors in the estimated values may be attributed to the mixed excitation during the voiced fricatives and the lack of vocal tract normalization.

5. Conclusions

Investigations for estimating the place of articulation of English voiced fricatives /v/, /z/, /ʒ/ and unvoiced /f/, /s/, /ʃ/ fricatives showed that the modified dominant spectral centroid and the normalized sum of absolute spectral slope parameter capture the place information more effectively compared to the parameters proposed earlier. A GMM-based estimation resulted in a reduction in the RMS error by 2.1 mm. Large errors in the voiced fricatives can be due to the mixed excitation of these utterances. The error in estimation may also be due to the lack of vocal tract normalization. Further investigations may help in reducing the error.

6. Acknowledgements

The research is supported by the project “National Programme on Perception Engineering,” sponsored by the Department of Electronics & Information Technology, Government of India.

7. References

[1] R. G. Crichton and F. Fallside, “Linear prediction model of speech production with applications to deaf speech training,” *Proc. IEE Control Sci.*, vol. 121, pp. 865–873, 1974.

[2] S. H. Park, D. J. Kim, J. H. Lee, and T. S. Yoon, “Integrated speech training system for hearing impaired,” *IEEE Trans. Rehab. Eng.*, vol. 2, no. 4, pp. 189–196, Dec. 1994.

[3] A. Neri, C. Cucchiarini, H. Strik, and L. Boves, "The pedagogy –technology interface in computer assisted pronunciation training," *Computer Assisted Language Learning,* vol. 15, pp. 441 – 467, 2002.

[4] H. Wakita, “Direct estimation of the vocal tract shape by inverse filtering of acoustic speech waveforms,” *IEEE Trans. Audio Electroacoust.,* vol. AE-21, no. 5, pp. 417–427, 1973.

[5] P. C. Pandey and M. S. Shah, “Estimation of place of articulation during stop closures of vowel-consonant-vowel utterances,” *IEEE Trans. Audio, Speech, Language Process.*, vol. 17, no. 2, pp. 277-286, 2009.

[6] V. N. Sorokin, "*Inverse problem for fricatives," Speech Commun.*, vol. 14, no. 3, pp. 249-262, 1994.

[7] S. Panchapagesan and A. Alwan, "A study of acoustic-to-articulatory inversion of speech by analysis-by-synthesis using chain matrices and the maeda articulatory model," J. *Acoust. Soc. Am.*, vol. 129, no. 4, pp. 2144–2162, 2011.

[8] K. Shirai, and S. Masaki “An estimation of the production process for fricative consonants*,” Speech Commun.*, vol. 2, no.2-3, pp. 111-114, 1983.

[9] S. Hiroya and M. Honda, “Estimation of articulatory movements from speech acoustics using an HMM-based speech production model,” *IEEE Trans. Speech Audio Process,* vol. 12, pp. 175–185, 2004.

[10] T. Toda , A. Black and K. Tokuda, "Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model," *Speech Commun.*, vol. 50, no. 3, pp. 215-227, 2008.

[11] P. Ladefoged, *A Course in Phonetics,* 2nd ed. New York: Harcourt Brace Jovanovich, 1982.

[12] D. O’Shaughnessy, *Speech Communications: Human and Machines*, 2nd ed. Piscataway, NJ: IEEE Press, 2000.

[13] A. Jongman, R. Wayland, and S. Wong, "Acoustic characteristics of English fricatives," *J. Acoust. Soc. Am.*, vol.108, pp. 1252–1263, 2000.

[14] S. R. Baum, and S. E. Blumstein, "Preliminary observations on the use of duration as a cue to syllable-initial fricative consonant voicing in English*," J. Acoust. Soc. Am.*, vol. 82, no. 3, pp. 1074–1077, 1987.

[15] A. M. A. Ali et al., "Acoustic-phonetic features for the automatic classification of Fricatives," *J. Acoust. Soc. Amer.*, vol. 109, pp. 2217–2235, 2001.

[16] K. S. Nataraj, P. C. Pandey, and H. Dasgupta, "Estimation of place of articulation of fricatives from spectral characteristics for speech training," in *Proc. 18th Annual Conference of the International Speech Communication Association (INTERSPEECH 2017), Stockholm,* 2017, pp. 339-343.

[17] J. Bentley, "Programming pearls. Algorithm Design Techniques," *ACM* vol. 27, no. 9, pp. 865-871, 1984.

[18] T. Takaoka, "Efficient algorithms for the maximum subarray problem by distance matrix multiplication," *Proc. CATS 2002, ENTCS*, vol. 61, pp 191–200, 2002.

[19] E. Zwicker, "Subdivision of the audible frequency range into critical bands (Freqenzgruppen*)," J. Acoust. Soc. Am.,* vol. 33, no. 2, pp. 248, 1961.

[20] J. R. Westbury, "X-ray microbeam speech production database user's handbook (version 1.0)," 1994 [Online]. Available: www.haskins.yale.edu/staff/gafos_downloads/ubdbman.pdf.

[21] K. S. Nataraj and P. C. Pandey, ''Place of articulation from direct imaging for validation of its estimation from speech analysis,'' in *Proc. 5th National Conf. Comput. Vision, Pattern Recognition, Image Process. and Graph. (NCVPRIPG 2015),* Patna, 2015, paper no. 88.