INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Estimation of Place of Articulation of Fricatives from Spectral Characteristics for Speech Training

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

A visual feedback of the place of articulation is considered to be useful for speech training aids for hearing-impaired children and for learners of second languages in helping them in improving pronunciation. For such applications, the relation between place of articulation of fricatives and their spectral characteristics is investigated using English fricatives available in the XRMB database, which provides simultaneously acquired speech signal and articulogram. Place of articulation is estimated from the articulogram as the position of maximum constriction in the oral cavity, using an automated graphical technique. The magnitude spectrum is smoothed by critical band based median and mean filters for improving the consistency of the spectral parameters. Out of several spectral parameters investigated, spectral moments and spectral slope appear to be related to the place of articulation of the fricative segment of the utterances as measured from articulogram. The data are used to train and test a Gaussian mixture model to estimate the place of articulation with spectral parameters as the inputs. The estimated values showed a good match with those obtained from the articulograms.

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

1. Introduction

In children with normal hearing, speech correction during the learning process of speech is aided by auditory feedback of the sounds. Due to lack of auditory feedback, hearing-impaired children have difficulty in acquiring the ability to control the articulators involved in speech production. Speech training aids providing appropriate non-auditory feedback can help the process of speech correction in hearing-impaired children. Several such aids have been developed to provide a dynamic display of important acoustic parameters, such as short-time energy, voicing and pitch, spectral features, etc. [1]-[3]. Speech-training aids providing a visual feedback of articulatory efforts have been found to be useful in improving vowel articulation by the hearing-impaired children [4]-[5]. The visual feedback can also be used to help learners of a second language to improve their pronunciation [6].

Wakita's LPC-based method [7] is the commonly used method for estimation of vocal tract shape in speech training aids due to its low computational requirement [4]. It works satisfactorily during vowels, diphthongs, and semivowels as the vocal tract configuration in these cases can be modeled as an all-pole filter. Pandey and Shah [8] reported a method to estimate the vocal tract shape during stop closures of vowel-consonant-vowel (VCV) utterances by using a bivariate surface model fitted on the vocal tract shapes during the transition segments preceding and following the stop closure and release burst. Nasals and fricatives pose difficulties for the LPC-based method because of spectral zeros in the vocal tract filter model. We present investigations for estimating the place of articulation of fricatives from the spectral characteristics without using a filter model.

Fricatives are produced when a steady airflow is obstructed by a narrow constriction to create turbulence in the oral cavity. On the basis of the place of articulation (place of maximum constriction in the oral cavity), English fricatives are grouped into four classes: (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") [9]-[10].

The methods for vocal tract shape estimation during fricatives can be broadly grouped as based on analysis-by-synthesis and data-driven machine learning. In analysis-by-synthesis methods [11]-[13], the parameters of an articulatory synthesizer are iteratively adjusted to minimize the acoustic distance between the synthesized speech signal and input speech signal. An acoustic-to-articulatory codebook is used to initialize the iteration, but the codebook searches may not yield a good initial parameter set [12]. Machine learning methods use MFCC as the input features and x and y coordinates of the articulators as the output parameters [14]-[15]. These methods work well using speaker-dependent mapping, but speaker-independent mapping is needed for speech-training aids.

Several investigations relating the place of articulation of fricatives to acoustic characteristics have been reported [16]-[18]. Jongman *et al.* [16] reported that spectral peak location, spectral moments, and energy relative to the adjacent vowel could be used to distinguish all four places of articulation while F2 transition properties and noise duration were not helpful. Based on the observation of spurious spectral peaks due to voice bars in the voiced fricatives, auditory perception-based spectral features including spectral centroid, maximum normalized spectral slope, and most dominant peak location were used in [18]. The maximum normalized spectral slope was suitable for distinguishing labio-dental and sibilant fricatives.

Earlier studies have generally investigated relationship between the spectral characteristics and the categorical values of the place of articulation. As small changes in place of articulation may result in significant changes in spectral characteristics and may introduce error in the acoustic-to-articulatory mapping, there is a need to study the relationship between the spectral parameters computed from the speech signal and the place of maximum constriction as obtained by an imaging technique. The XRMB database [19] provides simultaneously acquired audio and articulogram (display of motion of articulators) and therefore we use it to study the relationship of place of articulation of the fricative segment of the utterances, measured from articulogram, with the spectral

http://dx.doi.org/10.21437/Interspeech.2017-1074

parameters. The magnitude spectrum is smoothed by critical band based filters for improving the consistency of the spectral parameters. The selected parameters are used to train and test a Gaussian mixture model to estimate the place of articulation.

The method and material are described in the second section. The third section presents investigations relating the place of articulation with spectral parameters. The experiments for estimation of place of articulation using selected spectral parameters and results are presented in the fourth section, followed by conclusion in the last section.

2. Material and method

The XRMB database [19] consists of simultaneous recordings of the articulatory data and speech signal for vowels, words, and sentences. In our study, recordings of the English fricatives involving voiced fricatives /v/, /z/, /ʒ/ and unvoiced fricatives /f/, /s/, /ʃ/ in the database were used. The speech signals in the database have sampling frequency of 21.739 kHz and these were down-sampled to 16 kHz. Place of articulation was obtained from the x-y articulatory plots and the spectrum was smoothed for improving the consistency of spectral parameters.

2.1. Place of articulation from x-y articulatory plots

The place of articulation during frication was extracted from the x-y articulatory plots in the XRMB database. These plots show four pellet points (T1-T4) on the tongue and one each on the upper lip (UL), lower lip (LL) and incisor (MNi), along with the palatal outline and posterior pharyngeal wall as shown in Fig. 1. Place of articulation was estimated using an automated technique reported in [20] by graphical processing of the upper and lower contours of the oral cavity. This method iteratively estimates the axial curve as an axis of symmetry of the oral cavity, such that the curve approximately bisects the normals to it. The distance between the contours along the normal to the axial curve gives the oral cavity opening and position of the smallest opening is taken as the place of articulation.

2.2. Spectral smoothing

The smoothed magnitude spectrum was calculated for the central one-third segment of each fricative utterance. For this segment, the magnitude spectra were calculated using a window length of 20 ms with 5 ms shift and FFT size $N = 512$ and were averaged. Spectra of multiple utterances of a fricative from the same speaker generally show random variations but nearly similar spectral envelopes. In order to improve the consistency of spectral parameters, the spectra were smoothed using two-step median-mean filtering, as described in [21]. Median filtering smooths out small variations without significantly disturbing large variations. The two-step median-mean filtering restores the peaks and valleys which may get distorted by single-step filtering. The number of samples used for median and mean filters at each spectral sample was equal to the number of samples in the critical band centered at the sample. The critical bandwidth $B(l)$ centered at spectral sample l with frequency $f(l)$, in kHz, was estimated as [22]

$$B(l) = 25 + 75(1 + 1.4(f(l))^2)^{0.69} \qquad (1)$$

The critical band based filtering was selected as the smoothing performed by it is similar to that in the auditory system. An example of the smoothing for fricatives /s/ and /ʃ/ is shown in Fig. 2, with the dotted and continuous curves indicating the averaged and the smoothed spectra, respectively. It is seen that spurious variations are suppressed without significantly distorting the peaks and valleys.

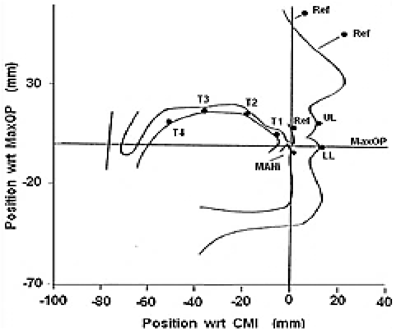

Figure 1: P*osition of pellet points in XRMB database.*

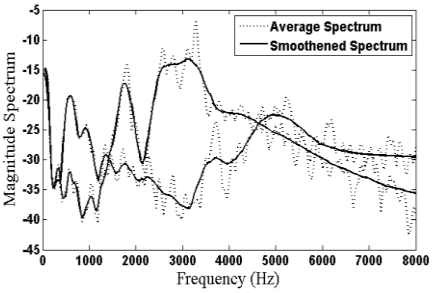

Figure 2: *Example of spectral smoothing.*

3. Relationship of place of articulation with spectral parameters

The XRMB database has 48 speakers. Investigations were carried out using vowel-consonant-vowel utterances (VCV) of the type /ΛCa/ involving the voiced fricatives /v, z, ʒ/ and the unvoiced fricatives /f, s, ʃ/. Data from two speakers were non-useable due to missing pellet positions. Some of the speakers did not have full set of fricatives in the recordings. A total of 221 utterances from 46 speakers (21 male and 25 female) were used in the study. Place of maximum constriction was obtained from the x-y articulatory plots and its distance from the lips, measured in mm, was taken as the place of articulation. Computation of the spectral parameters and their relation with place of articulation is described in the following subsections.

3.1. Spectral moments

The normalized magnitude spectrum $P(k)$, with k as frequency index, is calculated from the smoothed average magnitude spectra $S(k)$ as

$$P(k) = S(k) / \sum_{k=1}^{N/2} S(k) \qquad (2)$$

The first moment (spectral centroid) and the second moment (measure of spectral bandwidth) are calculated as

$$m_1 = \sum_{k=1}^{N/2} kP(k) \qquad (3)$$

$$m_2 = \left[\sum_{k=1}^{N/2} (k - m_1)^2 P(k) \right]^{1/2} \qquad (4)$$

The plots of the spectral centroid and standard deviation versus the place of articulation are shown in Fig. 3(a) and Fig. 3(b), respectively. The values of the place of articulation for the palatal, alveolar, and labio-dental fricatives lie in the range of 27 – 40, 18 – 26, and 0 – 1 mm, respectively. The spectral centroid decreases as the place of constriction increases from

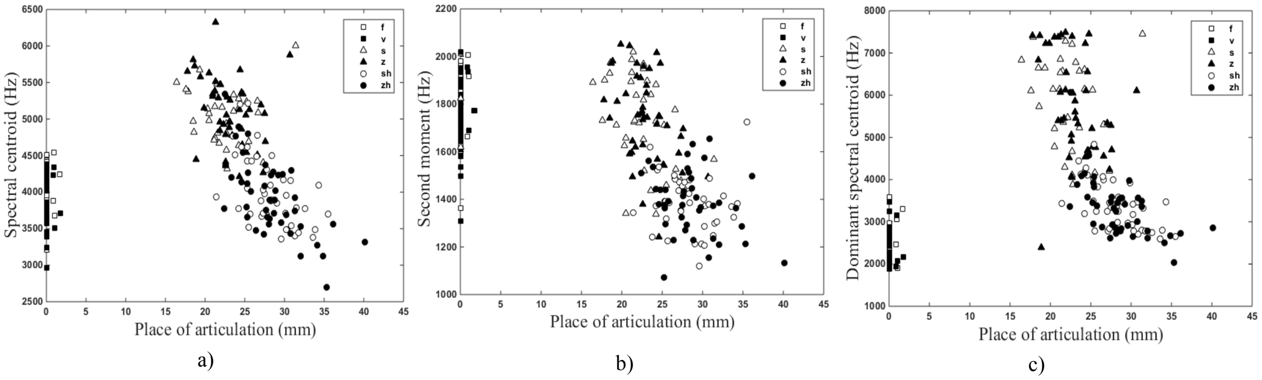

Figure 3: *Relationship between the spectral parameters and the place of constriction for voiced /v, z, ʒ/ and unvoiced /f, s, ʃ/ fricatives: (a) spectral centroid, (b) standard deviation, (c) dominant spectral centroid.*

alveolar position to palatal position. It may be due to the occurrences of distinct spectral peaks at low and high frequencies in the palatal and alveolar fricatives, respectively. The labio-dentals and palatals have similar spectral centroids. The standard deviation decreases as place of constriction increases from alveolar position to palatal position. It may be attributed to spectral flatness of alveolar fricatives. The labio-dentals and alveolars have similar values of standard deviation. The skewness and kurtosis did not show any distinct relationship with the place of articulation.

The alveolars in many cases have relatively flat spectrum and thus have low spectral centroid, similar to the labio-dentals. However, alveolars have energy concentrated in the high frequency region and the labio-dentals have it in the low frequency region. To exploit this characteristics, the dominant spectral centroid (DSC) is calculated as the centroid of the values of the magnitude spectrum above the 80 percentile of the distribution, i.e,

$$\text{DSC} = \sum_{k\in \text{D}} kM(k) \tag{5}$$

where $\text{D} = \{k, \text{ such that } S(k) > P_{th}\}$ and P_{th} is the 80 percentile value of the smoothed magnitude spectral values. The D-normalized spectrum is calculated as

$$M(k) = S(k) / \sum_{k\in \text{D}} S(k) \tag{6}$$

The plot of the DSC versus the place of articulation is shown in Fig. 3(c). The DSC value decreases as the place of articulation changes from alveolar position to palatal position. Compared to the spectral centroid, the DSC values of the labio-dentals are much smaller than those of alveolars because of the localized spectral centroid which removes the effect of insignificant spectral samples. Also the spread of the DSC values is reduced compared to the spectral centroid. However, the DSC values of labio-dentals and palatals have significant overlap and thus need other parameters to separate them.

3.2. Spectral slope

Maximum normalized spectral slope (MNSS), somewhat similar to that in [18], is calculated as the maximum value of first difference of the smoothed spectrum and normalized with respect to the sum of the magnitude values of the spectrum, i.e,

$$\text{MNSS} = \max_{k\in\{2,\dots,N/2\}} (S(k) - S(k-1)) / \sum_{l=1}^{N/2} S(l) \tag{7}$$

The plot of the maximum normalized spectral slope versus the place of articulation is shown in Fig. 4(a). The alveolars and labio-dentals have similar slopes. The palatals have larger slopes than the alveolars and labio-dentals.

The spectra of labio-dentals have downward tilt. Thus, the first difference values of the smoothed spectrum of labio-dentals are mostly negative. The first difference of the spectra of alveolars and palatals have both positive and negative values. To represent this characteristic, the normalized sum of spectral slopes (NSSS) is calculated as the sum of the first difference values of smoothed spectrum normalized with respect to the sum of the magnitude values of the spectrum, i.e,

$$\text{NSSS} = \sum_{k=2}^{N/2} (S(k) - S(k-1)) / \sum_{l=1}^{N/2} S(l) \tag{8}$$

The plot for the NSSS values is shown in Fig. 4(b). The values are negative for the labio-dentals. For palatals and alveolars, they are either positive or zero. Thus, palatals can be separated from the labio-dentals using this parameter.

As the first difference of smoothed spectrum of the labio-dentals will be mostly negative, the sum of the positive values of the first difference will be small compared to that for palatals and alveolars which have distinct spectral peaks. To exploit this characteristics, normalized sum of positive spectral slope (NSPSS) is defined as the sum of the positive values of the first difference of smoothed spectrum normalized with respect to the sum of the magnitude values of the spectrum, i.e,

$$\text{NSPSS} = \sum_{k\in(Q(k)>0)} Q(k) / \sum_{k=1}^{N/2} S(k) \tag{9}$$

where $Q(k) = S(k) - S(k-1)$. The plot of the NSPSS versus the place of articulation is shown in Fig. 4(c). The NSPSS increases with the place of articulation and the NSPSS values of the palatals and labio-dentals are well separated.

4. Estimation of place of articulation

The place of articulation is estimated from the input spectral parameters using the method based on the Gaussian mixture model proposed by Toda *et al.* [15]. In this method, a joint vector is obtained by concatenating the spectral parameter vector with the corresponding articulatory parameter vector. As described in the previous section, DSC captured the variation of place of articulation with lesser spread and it also separated the clusters of alveolars and labio-dentals. However, it could not separate the labio-dentals and palatals. NSSS and NSPSS enhanced the separation of the clusters of palatals and labio-dentals compared to the MNSS. Based on these results, the spectral parameter vector is formed of DSC, NSSS, and NSPSS

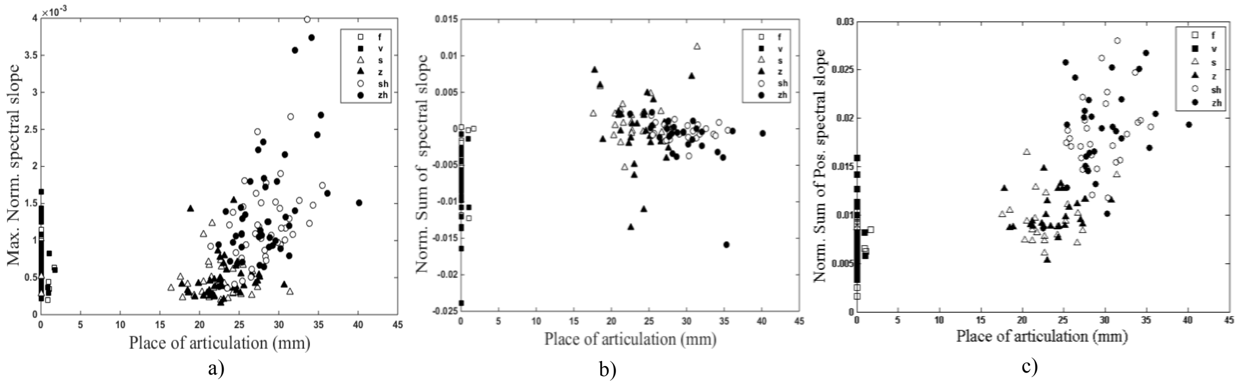

Figure 4: *Relationship between spectral parameters and place of articulation for voiced /v, z, ʒ/ and unvoiced /f, s, ʃ/ fricatives: (a) maximum normalized spectral slope, (b) normalized sum of spectral slope, (c) normalized sum of positive spectral slope.*

and articulatory parameter vector consists of place of articulation. The joint probability desnity function of spectral parameters and place of constriction is modelled using Gaussian mixture model. The GMM parameters are obtained using expectation-maximization algorithm.

The maximum likelihood estimator of the articulatory parameters $\mathbf{y}_t$ given the spectral parameters $\mathbf{x}_t$ is expressed as

$$\hat{\mathbf{y}}_t = \arg\max_{\mathbf{y}_t} p(\mathbf{y}_t|\mathbf{x}_t,\boldsymbol{\theta}^{(q)}) \qquad (10)$$

where $\boldsymbol{\theta}^{(q)}$ is the parameter set for joint probability density function, consisting of mean vector, covariance matrices, and Gaussian component weights. The modelling was carried out using 20 mixture components and full covariance matrices.

4.1 Experiments

The dataset for training and testing the GMM based technique for estimating the place of articulation consisted of the fricative segments extracted from the material available in the XRMB database with VCV utterances, words, and sentences. Data from one speaker could not be used because of missing pellet positions. A total of 3737 fricative segments involving voiced fricatives (/v/, /z/, /ʒ/) and unvoiced fricatives (/f/, /s/, /ʃ/) from 47 speakers were extracted. After removing incorrectly pronounced utterances, a total of 3133 fricative utterances were used in the study. Utterances from randomly selected 35 speakers were used for training and those from the remaining 12 speakers were used for testing, with 2371 and 762 utterances in the training and testing datasets, respectively

4.2 Results

For the training dataset, the RMS error for estimation of the place of articulation was 4.7 mm and the correlation coefficient was 0.91. For testing data, the scatter plot of the estimated place of articulation versus the place of articulation obtained from the XRMB data is shown in Fig.5. The RMS error and correlation coefficient were 6.4 mm and 0.84, respectively.

For the testing dataset, the means and standard deviations of the estimated place of articulation and those obtained from the articulograms along with the means and standard deviations of the differences are given in Table 1. The differences for alveolars are small. The differences are larger for labio-dentals and palatals, which can be attributed to the overlapping of the DSC of the labio-dentals with that of palatals. The large differences for the voiced palatals /zh/ may be due to small number of utterances available in the database. The large standard deviations of the estimated values may be attributed to the lack of vocal tract normalization.

Table 1: *Comparison of estimated place of articulation (PoA-Est.) with those obtained from the XRMB (PoA-XRMB).*

Frica-tive	**No. of utter-ances**	**PoA-XRMB (mm)**		**PoA-Est. (mm)**		**Mean diff. (mm)**	**S.D. of diff. (mm)**
		Mean	**S.D.**	**Mean**	**S.D.**		
/f/	137	0.2	0.5	2.9	8.0	-2.7	8.0
/v/	68	0.1	0.3	2.0	6.9	-1.9	6.9
/s/	218	23.7	2.6	22.9	4.6	0.8	4.7
/z/	194	23.8	2.5	23.5	4.4	0.3	4.7
/sh/	136	29.8	3.8	26.8	5.7	3.0	7.0
/zh/	9	28.9	4.7	21.7	12.0	7.2	14.2

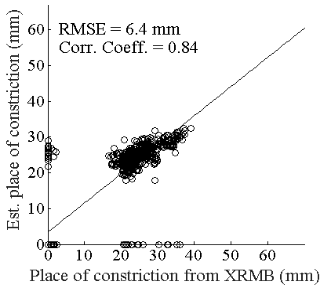

Figure 5: *Scatter plot of estimated place of constriction using GMM versus the actual ones from XRMB database.*

5. Conclusion

Investigations on the relationship between the place of articulation and spectral characteristics of the VCV utterances with English voiced fricatives /v/, /z/, /ʒ/ and unvoiced /f/, /s/, /ʃ/ fricatives in the XRMB database showed that the dominant spectral centroid, normalized sum of spectral slope, and normalized sum of positive spectral slope were related to place of articulation. The spectral parameters and place of articulation values were used to train and test a Gaussian mixture model to estimate the place of articulation. The estimated values of place of articulation of alveolar fricatives showed a good match with those obtained from the articulograms. However, the errors in estimation of place of articulation of palatal and labio-dental fricatives were larger, indicating a need for vocal tract normalization in the estimation technique.

Acknowledgements

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

References

[1] S. B. Davis H. Levitt, J. M. Pickett, and R. A. Houde, (Eds.), “Speech training aids,” part VII *in Sensory Aids for the Hearing Impaired.* New York: IEEE Press, 1980, pp. 349-419.

[2] R. S. Nickerson and K. N. Stevans, “Teaching to a deaf: can a computer help ?” *IEEE Trans. Audio Electroacoust.,* vol. AU-21, no. 5, pp. 445-455, 1973.

[3] S.A. Zahorian and S. Venkat, "Vowel articulation training Aid for the deaf," in *Proc. lnt. Conf. on Acoust., Speech, and Signal Process.*, New York, 1990, pp. 1121-1124.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

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

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

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

[12] 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.

[13] 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.

[14] 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.

[15] 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.

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

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] L. R. Rabiner, M. R. Sambur, and C. E. Schmidt, “Applications of a nonlinear smoothing algorithm to speech processing,” *IEEE Trans. Acoust., Speech, Signal Processing,* vol. ASSP-23, pp. 552- 557, Dec. 1975.

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