# ESTIMATION OF PLACE OF ARTICULATION OF FRICATIVES FROM SPECTRAL PARAMETERS USING ARTIFICIAL NEURAL NETWORK

*Thesis*
*submitted in partial fulfillment of the requirements*
*for the degree of*

**Doctor of Philosophy**

*by*

**K. S. Nataraj**
(Roll No. 144070003)

*under the supervision of*

**Prof. P. C. Pandey**


**Department of Electrical Engineering**
**Indian Institute of Technology Bombay**

**September 2021**

**Dedicated to my mother**

# Indian Institute of Technology Bombay
# Department of Electrical Engineering

## *Ph.D. Thesis Approval*

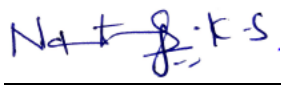
Thesis entitled **"Estimation of place of articulation of fricatives from spectral parameters using artificial neural network"** by **K. S. Nataraj (Roll No. 144070003)** is approved, after the successful completion of viva-voce examination, for the award of the degree of **Doctor of Philosophy**.

| | | |
|---|---|---|
| Supervisor: | PCPandey | (Prof. P. C. Pandey) |
| Internal Examiner: | Preeti Rao | (Prof. P. Rao) |
| External Examiner: | 16/09/2021 | (Prof. R. Sinha) |
| Chairman | D. N. Singh | (Prof. D. N. Singh) |

Date: 16th September, 2021

## DECLARATION

I declare that this written submission represents my ideas in my words and where ideas or words are taken from others, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and I have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.



K. S. Nataraj
(Roll No: 144070003)

Date: 15 September 2021

K. S. Nataraj / Supervisor: Prof. P. C. Pandey, "Estimation of place of articulation of fricatives from spectral parameters using artificial neural network", *PhD Thesis*, Department of Electrical Engineering, IIT Bombay, September 2021.

---

**Abstract**

Speech-training aids providing visual feedback of the place of articulation, quantified as the distance of maximum constriction from the lips, are useful for improving the consonant articulation of children with hearing impairment. For such feedback, the place of articulation needs to be estimated from the speech signal. The research objective is to develop a speaker-independent method for estimating the place of articulation during fricatives for visual speech training.

The relation between the place of articulation of fricatives and their spectral parameters is investigated using the simultaneously acquired speech signals and articulograms available in the X-ray microbeam database. An automated graphical technique is developed for estimating the place of articulation from the articulograms, with the estimated place of articulation values closely matching those obtained by manual marking. Investigation relating the place of articulation with the spectral parameters is carried out using several earlier reported spectral parameters and a set of proposed spectral parameters. Earlier reported parameters, including spectral moments and spectral peak frequency, and the proposed parameters, including maximum-sum segment centroid, normalized sum of absolute spectral slopes, and four spectral energy parameters, were found to be associated with the place of articulation. An investigation is carried out for ANN-based speaker-independent mapping from spectral parameters of the frication segments to the place of articulation, using a feedforward network with multiple hidden layers and different number of neurons, different training data sizes, different sets of spectral parameters as the input, the place of articulation estimated from the articulograms as the reference, a dataset with 10,112 utterances, and five-fold cross-validation. Networks with two hidden layers were found to be adequate for all input parameter sets. The estimation using the proposed set of parameters resulted in the smallest mean RMS error of 2.55 mm, with scope for improving the estimation by increasing the training data size. The errors for alveolar and palatal fricatives were comparable to the standard deviation of the reference values. The errors for labiodentals were larger than the standard deviation of the reference values but smaller than their distance from the alveolars. The results indicated that the proposed ANN-based speaker-independent estimation could be used for feedback of the place of articulation.

A significant part of the estimation error could be attributed to non-uniqueness in the mapping. A perceptual study on the relative importance of transition segments adjacent to the fricative and the frication showed the place perception to be determined by a combination of the frication and transition segments. Investigation using the spectral parameters computed from the vocalic transition adjacent to frication showed that the ANN-based place estimation could be improved by supplementing the frication information with the vocalic information represented by the transition and vowel parameters.

## CONTENTS

## List of Figures

## List of Tables

## List of Symbols

**Symbols**

| | |
|---|---|
| DSC | dominant spectral centroid (Hz) |
| $f(k)$ | frequency at frequency index $k$ in Hz |
| $f_S$ | sampling frequency in Hz |
| $i_{BG}(k)$ | beginning of the maximum-sum subarray |
| $i_{BL}(k)$ | beginning of the local subarray |
| $i_{EG}(k)$ | end of the maximum-sum subarray |
| $k$ | frequency index of spectrum |
| $l$ | band index of mel-filter bank energies |
| $L$ | FFT size |
| MLBE | average mid-frequency band energy (MLBE) expressed in dB with respect to the average low-frequency-band energy |
| MSSC | maximum-sum segment centroid (Hz) |
| NHBE | normalized average energy in the high-frequency band (dB) |
| NSASS | normalized sum of absolute spectral slope |
| NVLBE | normalized average energy in the very low-frequency band (dB) |
| PAE | peak energy expressed in dB with respect to the average energy |
| $P(k)$ | normalized squared magnitude spectrum |
| $q$ | joint vector formed by concatenation of the spectral and articulatory parameter vector |
| $S(k)$ | average magnitude spectrum |
| SM1 | First spectral moment (Hz) |
| SM2 | Second spectral moment (Hz) |
| SM3 | Third spectral moment |
| SM4 | Fourth spectral moment |
| $\alpha$ | warping factor for vocal tract length normalization |
| $\boldsymbol{\theta}^{(q)}$ | GMM parameter set of the probability density function of joint vector formed by concatenation of the spectral and articulatory parameter vector. |

## List of Abbreviations

**Abbreviations**

| | |
|---|---|
| ANN | artificial neural network |
| CV | consonant-vowel |
| DNN | deep neural network |
| DSC | dominant spectral centroid |
| MSSC | maximum-sum segment centroid |
| PAE | peak energy expressed in dB with respect to the average energy |
| EGG | electroglottograph |
| EMA | electromagnetic articulography |
| EPG | Electropalatography |
| FFT | fast Fourier transform |
| F1 | first formant frequency (Hz) |
| F2 | second formant frequency (Hz) |
| GMM | Gaussian mixture model |
| LP | linear prediction |
| LSF | line spectral frequencies |
| MFCC | mel-frequency cepstral coefficients |
| MLE | maximum likelihood estimation |
| MRI | magnetic resonance imaging |
| MSSC | maximum-sum segment centroid (Hz) |
| n-Amp | normalized amplitude (dB) |
| NHBE | normalized average energy in the high-frequency band (dB) |
| NSASS | normalized sum of absolute spectral slope |
| NVLBE | normalized average energy in the very low-frequency band (dB) |
| PoA-art | place of articulation from the articulatory data |
| RMS | root mean square |
| RMSE | RMS of errors |
| SPF | spectral peak frequency |
| SPS-6 | Spectral parameter set 6 |
| VCV | vowel-consonant-vowel |
| VTLN | vocal tract length normalization |
| XRMB | X-ray microbeam |

# Chapter 1
# INTRODUCTION

## 1.1 Problem overview

Auditory feedback helps children with normal hearing in speech acquisition process. It helps them to compare their sounds to those produced by other persons and acquire the ability to control the movement of the articulators (tongue body, tongue tip, tongue blade, velum, jaw, and lips) to produce intelligible speech (Kuhl 2000, McGowan *et al.* 2008, Shiller and. Rochon 2014). Children with prelinguistic hearing impairment experience great difficulty in learning to produce intelligible speech as they do not have the target acoustic pattern to imitate and compare with their production. The lack of auditory input causes several deficiencies in articulation, stress, and intonation patterns. The tongue movement is restricted during the production of sounds resulting in the target vowels' substitution by the mid vowels. The consonants produced with the tongue movements not visible from outside are indistinct and are substituted by glottal sounds. Fricative and affricate sounds are often misarticulated and substituted by stops. Thus the children with hearing impairment, despite having functional articulators, are generally unable to produce adequately intelligible speech (Nober 1967, Nickerson and Stevens 1973, Elfenbein *et al.* 1994, Moeller *et al.* 2007, McGowan *et al.* 2008).

For assisting the process of speech acquisition in persons with hearing impairment, the lack of auditory cues needs to be compensated by providing the appropriate cues through other sensory modalities. Speech training aids displaying important acoustic parameters, such as short-time energy, voicing, pitch, and spectral features, have been shown to help the articulation of persons with hearing impairment (Nickerson and Stevens 1973, Mahshie *et al.* 1988, Adams *et al.* 1989, Zahorian and Venkat 1990, Micro Video Corp. 2017). However, the effectiveness of these aids is limited as they provide information on the articulation's correctness compared to a reference but do not provide information for correcting the articulatory effort.

Speech therapists generally use a mirror to provide feedback on position of the tongue and the lips to improve articulation (Roth and Worthington 2011). Investigations on the articulation errors in the speech of children with hearing impairment have reported the bilabial consonants to be more intelligible than the lingual consonants and the vowels, and the intelligibility differences are attributed to the relative visibility of the articulators involved in producing the sounds (Huntington *et al.* 1968, Smith 1975). Speech training aids providing

visual feedback of the articulators' movements inside the mouth have been found to be useful in improving the articulation (Massaro and Light 2004, Bernhardt *et al.* 2008, Bacsfalvi and Bernhardt 2011, Engwall 2012, Wilson 2014, Katz and Mehta 2015). As it may be difficult to imitate the entire vocal tract shape, it may be more helpful to provide feedback on salient aspects of articulation, such as the place of maximum constriction in the oral cavity.

For visual feedback of the articulatory movements to the speaker, the articulatory parameters need to be estimated. Direct imaging techniques, such as X-ray imaging, electromagnetic articulography (EMA), electropalatography, magnetic resonance imaging (MRI), and ultrasonic imaging, can be used to obtain the location and movements of articulators (Hardcastle *et al.* 1991, Westbury 1994, Munhall *et al.* 1995, Zhang 1997, Gick 2002, Bacsfalvi and Bernhardt 2011, Narayanan *et al.* 2014). However, they interfere with speech production and are time-consuming. Several indirect methods have been proposed to estimate vocal tract shape using the acoustic measurements or by processing the speech signal (Schroeder 1967, Sondhi and Gopinath 1971, Wakita 1973, Hiroya and Honda 2004, Toda *et al.* 2008, Richmond 2006, Panchapagesan and Alwan 2011, Ji *et al.* 2014b, Liu *et al.* 2015, Ji *et al.* 2016, Illa and Ghosh 2018). Estimation based on acoustic measurements is not suitable for speech training, as it requires the speaker to articulate silently with an acoustic tube connected to the mouth. Estimation based on processing the speech signal does not pose any restrictions on speaking, and hence it is considered suitable for use in speech training.

The LP-based method proposed by Wakita (1973) is commonly used to estimate vocal tract shape from speech signal in speech training aids for vowel articulation training. The method is based on modeling the vocal tract as an all-pole filter. Despite several limitations, as outlined by Wakita (1979), the method works satisfactorily during non-nasalized vowels, diphthongs, and semivowels. However, it fails during stop closures due to the lack of spectral information to estimate the shape. Pandey and Shah (2009) used bivariate surface polynomial interpolation of the shape estimates during transitional segments preceding and following the stop closures of vowel-consonant-vowel utterances to estimate the shape during stop closure. The LP-based method is not useful during nasals and fricatives due to the presence of spectral zeros, and hence other methods for estimation of the vocal tract shape, particularly the place of articulation, during these segments need to be investigated.

Several investigations relating the place of articulation of fricatives to acoustic characteristics have been reported (Heinz and Stevens 1961, Shadle and Mair 1996, Jongman *et al.* 2000, Nissen and Fox 2005, Todd *et al.* 2011, Li *et al.* 2012, Zharkova 2016). These studies have generally investigated the relationship between the spectral characteristics and the categorical values of the place of articulation (labiodental, linguodental, alveolar, and

palatal). 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.

Earlier studies on spectral characteristics of English fricatives have reported that place perception for fricatives is determined by the adjacent vocalic segments in the case of non-sibilants (Harris 1958, Wagner *et al.* 2006). Hence, the use of spectral parameters during the vocalic segments may improve the place estimation. Earlier studies have also reported that the frication duration affects the place perception (Baum and Blumstein 1987, Jongman 1989). These studies have examined the effect of vocalic segments and duration on the place perception separately. As the vocalic segments' effect may vary with the frication duration, it may be useful to study the effect of the vocalic segments on the place perception of fricatives of different durations.

Machine learning methods to obtain the acoustic-to-articulatory mapping have become popular due to the availability of simultaneously acquired audio and articulatory data (Hiroya and Honda 2004, Richmond 2006, Toda *et al.* 2008, Ji *et al.* 2014b, Liu *et al.* 2015, Ji *et al.* 2016, Illa and Ghosh 2018). The articulatory data are acquired using direct imaging techniques such as EMA, MRI, X-ray imaging, and ultrasound imaging. These methods have been reported to provide reasonably good accuracy for speaker-dependent mapping for all types of sounds, including vowels, stops, fricatives, and nasals. The articulators' *x*-*y* positions, commonly used as the articulatory parameters, introduce significant variability in the acoustic-to-articulatory mapping (Ji *et al.* 2014b, Afshan and Ghosh 2015). Due to the speaker-dependent mapping, these methods perform poorly for acoustic data from an unseen speaker.

## 1.2 Research objective

The research objective is to develop a method for estimating place of articulation of the fricative segments by studying the relationship between the spectral characteristics and the place of articulation obtained from the simultaneously acquired audio and articulatory data.

The relationship between the spectral characteristics and the place of articulation is investigated, using the simultaneously acquired audio and articulograms of the English fricative utterances available in the X-ray microbeam database (Westbury 1994), to identify the spectral parameters suitable for estimation of the place of articulation. Investigation is carried out using several earlier reported spectral parameters and a set of proposed parameters. An automatic technique for estimating the axial curve of the oral cavity contour is

proposed, and the place of articulation is estimated as the position of the smallest oral cavity opening along the axial curve. A speaker-independent mapping based on an artificial neural network (ANN) is investigated for estimating the place of articulation, using the spectral parameters during the frication segments as the input feature vector and the place of articulation estimated from the articulograms as the target for network training. The place of articulation is used as the output parameter instead of the *x*-*y* positions of the pellets on the articulators to reduce the variability in the acoustic-to-articulatory mapping, which may occur due to the variability in the oral cavity size and pellet placement on articulators. The effects of the input feature vectors, the number of hidden layers, and the number of neurons in each layer are examined to obtain an optimal combination. Vocal tract length normalization based on frequency warping is investigated to compensate for variation in vocal tract length.

A perceptual study is carried out to investigate the relative importance of the transition and the frication on the place perception of the unvoiced fricatives /f, s, ʃ/ in vowel-consonant-vowel (VCV) sequences with natural transition segments and synthesized frication segments with durations of 50–300 ms. ANN-based estimation of the place of articulation using spectral parameters during transition segments is investigated to quantify the potential of spectral parameters during transition segments to improve the estimation of the place of articulation.

## 1.3 Thesis outline

The second chapter presents a brief review of the importance of visual feedback in speech-training aids for persons with hearing impairment, followed by a review of the techniques for vocal tract shape estimation. A description of the automated technique for the estimation of the axial curve and place of articulation from oral cavity contours, the investigation relating the place of articulation with spectral parameters, and ANN-based estimation of the place of articulation, along with the test results, are presented in the third chapter. The fourth chapter presents the details of the perceptual study on the effect of transition and frication on the place of articulation of fricatives in VCV utterances, listening test results, and ANN-based estimation of the place of articulation using the spectral parameters during the transition segment. The last chapter provides a summary of the investigations, conclusions, and suggestions for further work. A brief review of visual speech training aids for the hearing impaired is provided in Appendix A. An investigation on vocal tract length normalization for estimation of the place of articulation is presented in Appendix B. Appendix C presents an investigation for quantifying the non-uniqueness in the estimation of place of articulation from spectral parameters.

# Chapter 2
# ESTIMATION OF VOCAL TRACT SHAPE: A REVIEW

## 2.1 Introduction

The speech acquisition process in normal-hearing children involves imitating the sounds produced by others and self-correction by comparing their sounds with those of the others. Children with hearing impairment experience great difficulty in acquiring speech due to a lack of auditory feedback. Speech training aids providing non-auditory feedback can help the speech correction process in such children. Speech-training aids displaying acoustic parameters (such as short-time energy, voicing, pitch, spectral features) have been reported to help the speech acquisition process (Nickerson and Stevens 1973, Mahshie *et al.* 1988, Adams *et al.* 1989, Zahorian and Venkat 1990, Tiger DRS 1999, Carey 2004, Olson 2014, Micro Video Corp. 2017, Purr Programming 2017, Speechtools Ltd. 2019, DevExtras 2020). Investigations using these aids, summarized in Section A.2 of Appendix A, have shown that visualization of the acoustic parameters through interactive cartoons and games can motivate the children to continue practicing with the speech training systems. These aids provide information on correctness of the acoustic parameters to help the speech training process, but they do not provide feedback for correcting the articulatory effort (Park *et al.* 1994, Mahshie 1996, Neri *et al.* 2002, Wilson and Gick 2006).

Automatic speech recognition (ASR) derives the sequence of words in an utterance using various approaches, including spectral distance measure, template matching, hidden Markov model (HMM), and artificial neural network (ANN) model. Several ASR-based speech-training aids (described in Section A.3 of Appendix A) have been developed for improving the articulation of children with articulation difficulties (Kewley-Port *et al.* 1991, Vicsi *et al.* 2000, Ahmed *et al.* 2018). The effectiveness of these aids depends on the ASR accuracy, and the ASR systems trained using the speech from normal-hearing speakers may result in low accuracy for the speech from children with articulation difficulties (Bigham *et al.* 2017, Glasser *et al.* 2017, Ahmed *et al.* 2018). Further, the ASR-based aids provide information on the articulation's correctness but not the feedback for correcting it.

Speech therapists generally use a mirror to provide feedback about the lip movements (Roth and Worthington 2011, Grossinho *et al.* 2014). However, the mirror does not show movements of the articulators inside the oral cavity. It has been observed that the bilabial consonants produced by persons with hearing impairment tend to be more intelligible than the lingual consonants (Huntington *et al.* 1968, Smith 1975). Several speech-training aids

displaying the articulatory movements that are not externally visible have been reported to be useful in improving the articulation by the hearing-impaired children. A review of these aids (Crichton and Fallside 1974, Fletcher 1982, Pardo 1982, Black 1988, Dagenais *et al.* 1994, Massaro and Light 2004, Engwall *et al.* 2006, Martin *et al.* 2007, Bernhardt *et al.* 2008, Mahdi 2008, Bacsfalvi and Bernhardt 2011, Pickett 2013, Wilson 2014) is provided in the fourth section of Appendix A. For a speech-training aid to be effective, it should enable a visual comparison of the articulatory efforts of the student with that of the therapist or a reference speaker. For displaying the articulatory effort, the time-varying vocal tract configuration needs to be estimated. This estimation can be carried out directly using the imaging techniques or indirectly by acoustic measurements or processing the speech signal. This chapter presents a review of techniques for vocal tract shape estimation for use in speech training aids.

## 2.2 Vocal tract shape estimation by direct methods

Several techniques have been developed to measure the geometry of the vocal tract during speech production. These techniques include X-ray imaging (Chiba and Kajiyama 1941, Fant 1960, Westbury 1994), electropalatography (Hardcastle *et al.* 1991, Bacsfalvi and Bernhardt 2011), optopalatography (Fletcher 1982), ultrasound imaging (Gick 2002, Eshky *et al.* 2018), electromagnetic articulography (Zhang 1997, Ji *et al.* 2014a), and magnetic resonance imaging (Baer *et al.* 1991, Story *et al.* 1996, Narayanan *et al.* 2004, Narayanan *et al.* 2014).

Chiba and Kajiyama (1941) carried out 3D measurements of the vocal tract using X-ray photographs and vocal tract casts and used them to make mechanical models of the vocal tract. These models were used to generate vowel sounds using a telephone receiver as the excitation source. Fant (1960) reported use of X-ray based direct imaging to study vocal tract shapes of vowel sounds. This technique provided good visualization of the vocal tract configuration during articulation for use in speech research. As the tongue's soft tissue is difficult to image by X-rays, a contrasting material coating was used to obtain the tongue contour. In these studies, the speaker held the articulatory configuration during the X-ray image acquisition. In later studies, high-speed X-ray imaging was employed for imaging the moving vocal tract. A database consisting of 25 X-ray films, collected from Universite Laval and M.I.T, was created by Munhall *et al.* (1995). The database has the mid-sagittal views of the moving vocal tract during a reading of sentences by English and French speakers. It also has audio recordings synchronized with the articulatory movements, but they are severely affected by background noise. The use of X-ray based direct imaging in speech research was discontinued due to the risk of radiation hazards for the speakers (Munhall *et al.* 1995).

The X-ray microbeam (XRMB) technique (Westbury 1994) uses a narrow beam of X-rays for tracking the movement of articulators in the mid-sagittal plane using gold pellets glued on the tongue, the lips, and the jaw. In this technique, the X-ray beam path is sequentially adjusted to scan the location of each of the pellet points. The radiation exposure is reduced by sampling only the regions where the pellets are expected with an adequate sampling rate. A database (Westbury 1994) was developed at the University of Wisconsin, providing the *x* and *y* locations of four pellet points (T1-T4) on the tongue and one each on the upper lip (UL), the lower lip (LL), and the incisor (MNi), at 145 frames/s and the simultaneously obtained audio recordings. It has recordings for vowels, vowel-consonant-vowel syllables, sentences, and paragraphs, from 21 male and 26 female speakers. The database also provides the palatal outline for each speaker measured separately by scanning the pellets attached along the palate's midline. Although the technique has lower radiation exposure than the standard X-ray imaging, it is not considered suitable for multiple recordings from the same speaker (Westbury 1994).

Electropalatography (EPG) (Hardcastle *et al.* 1991), also known as palatometry, detects the contact between the tongue and the palate during speech production. The speaker wears a thin artificial palate, known as pseudopalate, molded specifically for the speaker. The pseudopalate has several electrodes embedded on its surface, with these electrodes and an electrode on the tongue body wired to a sensing hardware. As the tongue comes in contact with any of the palate electrodes, the low impedance path between the tongue electrode and the corresponding palate electrode is detected by the sensing hardware. The tongue-palate contact information during the production of obstruents is obtained in real-time and can be displayed using a computer.

The technique for ultrasound imaging of the tongue uses a transducer held below the chin for transmitting ultrasound waves towards the tongue and receiving the waves reflected from the boundary between the tongue surface and the air above it (Gick 2002). The images are reconstructed by measuring the time elapsed between transmission and reception, using a frame rate of 30 frames/s or higher. Automatic extraction of the tongue contour from the images is difficult due to noise and artifacts such as shadows created by hyoid bone and mandible (Qin *et al.* 2008). Manual extraction of the tongue contour is not suitable for speech therapy as it is time-consuming and requires specialized training (Eshky *et al.* 2018).

Electromagnetic articulography (EMA) tracks the location and movement of various articulators (tongue, jaw, lips, teeth, etc.) using electromagnetic induction without radiation risks (Zhang 1997). The setup comprises a set of transmitter coils mounted on a plastic helmet and a set of receiver coils placed on the articulators. The alternating voltage induced in a

receiver coil is inversely related to its distance from the corresponding transmitter coil. The voltage is used to calculate the receiving coil's $x$ and $y$ coordinates during articulation. The Multi-channel Articulatory (MOCHA-TIMIT) corpus (Wrench and William 2000), developed at the Edinburgh Speech Production Recording Facility, provides articulatory data at 500 frames/s and audio recordings for 460 sentences from one male and one female speaker. The Marquette University Electromagnetic Articulography Mandarin Accented English (EMA-MAE) database (Ji *et al.* 2014a) provides 3D articulatory data at 400 frames/s and audio recordings from 20 American English speakers and 20 Mandarin-accented English speakers reading words, sentences, and paragraphs, with about 45 minutes of recordings for each speaker.

The magnetic resonance imaging (MRI) technique can be used to capture images of the oral cavity (Baer *et al.* 1991, Narayanan *et al.* 1995, Story *et al.* 1996, Narayanan *et al.* 2004, Narayanan *et al.* 2014) in one or more orthogonal planes without radiation risk. Story *et al.* (1996) used it to obtain 3D vocal tract shapes for 12 vowels, three nasals, and three plosives by one male speaker. The vocal tract area function was used as input for articulatory synthesis based on the wave reflection model. The formant frequencies of the synthesized and natural speech signals were reported to be similar. Narayanan *et al.* (1995) used MRI to obtain 3D images of the vocal tract for sustained English fricatives by two male and two female speakers. The MRI technique has a large image acquisition time, and it can be used for imaging only sustained sounds (Baer *et al.* 1991, Narayanan *et al.* 1995, Story *et al.* 1996). Real-time MRI (rt-MRI) can be used to acquire images at a high frame rate. It permits simultaneous speech signal recording, but the signal is usually corrupted by the scanner noise (Narayanan *et al.* 2004). Bresch *et al.* (2006b) used two optical microphones to capture the noisy speech signal and the ambient noise separately. The noise in the speech signal was suppressed by offline processing using the ambient noise as a reference. The USC-TIMIT database (Narayanan *et al.* 2014) developed at the University of Southern California has MRI data at 23.18 frames/s and noise-suppressed audio recordings for 460 sentences from five male and five female speakers, and EMA data at 100 frames/s from two male and two female speakers.

The currently available direct methods are not suitable for estimating the vocal tract shape during speech training, as they require expensive equipment and skilled manpower and generally interfere with speech production. However, databases developed using these methods are useful for research and can be used as a reference for indirect methods.

## 2.3 Estimation of vocal tract shape by acoustic measurements

The vocal tract can be modeled as an acoustic tube with sections of varying cross-sectional areas with the vocal cords at the back end and the lip opening at the front end. An electrical analog or digital simulation of this model, along with the assumption of plane wave propagation and appropriate boundary conditions, can be used for speech synthesis with the area values as the inputs (Stevens *et al.* 1953, Kelly and Lochbaum 1962). Synthesis of the speech signal from the vocal tract area function is known as the direct problem. The inverse problem of obtaining the vocal tract area function from the speech signal is difficult as different vocal tract area functions can produce similar spectra. For addressing this non-uniqueness in the mapping between the speech spectrum and the vocal tract area function, methods for vocal tract shape estimation using the acoustic measurement of impedance or impulse response at the lips have been proposed. This section presents a review of these indirect methods.

Schroeder (1967) proposed a technique to estimate the vocal tract area function from the acoustic impedance measured at the lips. It was empirically verified that the first $2n$ coefficients in the Fourier expansion of the logarithm of the area function for a tube of a given length could be uniquely determined from the $n$ lowest-order poles and zeros of the acoustic impedance measured at the lips. The impedance measurement setup comprised an impedance tube, an electrodynamic driver unit, and two closely spaced condenser microphones. The electrodynamic driver unit, coupled to the tube's left end, provided periodic acoustic pulses with a repetition period of 10 ms and a bandwidth of 4 kHz. The tube's right end was connected to the speaker's mouth using a specially designed seal. The speaker articulated silently with the vocal cords closed. The two microphones picked up the incoming and reflected waves' pressure. The complex Fourier coefficients of the two microphone outputs were used to compute the input impedance at the lips. Area function was estimated iteratively using the poles and zeros of the input impedance and the plane wave equation, assuming a vocal tract length of 17 cm. It was reported that the correspondence between the measured and actual area functions for known fixed shapes was remarkably good.

Sondhi and Gopinath (1971) pointed out that the iterative procedure by Schroeder (1967) for obtaining the area function may not converge, and the assumptions of fixed vocal tract length and closed glottis may result in errors in the estimation. To avoid these limitations of the acoustic impedance technique, they proposed determining the vocal tract area function from the impulse response, measured as the pressure waveform at the lips in response to a unit impulse of volume velocity applied at the lips. The experimental setup for impulse response measurement had a tube connected to the speaker's mouth. The tube had a

microphone at the mouth end, a sound-absorbing wedge at the far end, and a sound source in the middle, with the sound moving towards the far end absorbed by the wedge and that moving towards the mouth picked up by the microphone. The impulse response at the lips was calculated from the pressure and volume measurements. They derived a mathematical relation between the impulse response and the vocal tract area function with the assumption of plane wave propagation with negligible losses. By using metal tubes with known dimensions as the vocal tract models, it was shown that the vocal tract area function could be uniquely obtained from the impulse response without involving the vocal tract length and boundary conditions.

The techniques based on acoustic measurements are not suited for speech training, as they require specialized instrumentation and unnatural silent articulation with the mouth sealed to the acoustic tube. No databases using these techniques have been reported.

## 2.4 Estimation of vocal tract shape using LP analysis

Wakita (1973) reported a technique for estimating the vocal tract area function from the speech signal by establishing equivalence between the vocal tract's acoustic tube model and the inverse filter obtained from the LP analysis. In this technique, the speech signal is modeled as generated by an excitation source and a filter representing the effects of the glottis, the vocal tract, and radiation at the lips. The excitation is assumed an impulse train for the voiced sounds, and the inverse filter coefficients are obtained by LP analysis using the least-mean-square error technique. The vocal tract is modeled as a lossless acoustic tube with sections of equal length and different cross-sectional areas and assuming plane wave propagation, a zero load at the lips, and a resistive load at the glottis. The reflection coefficients at the section interfaces are related to the partial correlation coefficients calculated using LP analysis by equating the responses of the LP-based inverse filter and the acoustic inverse filter. The ratios of areas of the adjacent sections are obtained from the reflection coefficients. A +6 dB/octave pre-emphasis is applied on the speech signal to account for the –12 dB/octave slope due to the glottal spectral envelope and +6 dB/octave slope due to the radiation impedance. Based on the results for the speech signal with a sampling frequency of 7 kHz, it was reported that the technique could be used for estimating fixed and transitional vocal tract configurations during voiced segments. The results for five vowels uttered by a male speaker showed the gross features of the estimated area functions to be similar to those reported earlier by Fant (1960).

Several limitations of the LP-based technique were described by Sondhi (1979) and Wakita (1979). Pre-emphasis does not eliminate the effect of glottal and radiation

characteristics. The assumption of a lossless vocal tract also contributes to the error in the estimated area functions. The assumption of a constant area at the glottis end introduces errors during dynamically varying vocal tract configuration due to variation in the glottis-end area during these transitions. The technique fails during nasals and fricatives due to zeros in the spectral envelope, as these cannot be modeled by the all-pole filter model used in the LP analysis. Further, the technique is not suitable during stop closures due to a lack of signal energy.

Pandey and Shah (2009) reported a method for estimating the vocal tract shape during stop closures of vowel-consonant-vowel (VCV) utterances using a bivariate surface polynomial interpolation of the LP-based estimates of the shapes during the transition segments preceding and following the stop release. The estimated places of closure for the /ʌCa/ utterances spoken by 20 male and 20 female speakers, involving stop consonants /b/, /d/, and /g/ were reported to have a good match with those obtained from the pellet locations in the XRMB database.

Deng *et al.* (2004) proposed a method to minimize the effect of varying glottal characteristics on the estimated vocal tract area functions. In this method, the closed glottal phases are marked by applying amplitude thresholding on the glottal wave signal obtained by LP analysis. The segments corresponding to these phases are used for estimating the vocal tract area functions. It was reported that the estimated vocal tract area functions for /a/ and /i/ spoken by a male and a female speaker were comparable to those obtained from MRI by Story *et al.* (1996). The vocal tract shapes estimated by Wakita's method during steady-state vowels show variability with the analysis frame's position. Nataraj *et al.* (2011) reported a method to select the frame positions to improve the consistency of the vocal tract shape estimation. Frame positions corresponding to the minima in the windowed energy index, calculated as the ratio of the windowed signal energy to the frame energy, were selected for computing the vocal tract area functions. It was reported that this selection reduced the variability in the estimated area functions during synthesized and natural steady-state vowels.

Nayak *et al.* (2012) proposed a method to estimate the scaling factor for converting the area ratios, obtained by Wakita's method, into area values during dynamically varying vocal tract configuration. The scaling factor was estimated from the lip opening area, obtained from a video recording of the speaker's face using a template matching approach. Use of the lip opening area for scaling the area ratios resulted in improved estimation of the area values at the places of maximum opening without significantly affecting the places of maximum constriction.

Wakita's LP-based inverse filtering method, despite its limitations, is used for estimating the vocal tract area functions for providing visual feedback in several speech-training aids. Several modifications to the LP-based method have been proposed to resolve some of the limitations (Pandey and Shah 2009, Deng *et al.* 2004, Nataraj *et al.* 2011, Nayak *et al.* 2012). However, the LP-based method is not useful during nasals and fricatives due to the spectral zeros' presence.

## 2.5 Vocal tract shape estimation using analysis-by-synthesis approach

Synthesis of the speech signal using a speech production model and articulatory parameters (tongue shape, lip opening, lung pressure, nasal coupling, etc.) is known as articulatory synthesis (Schroeter and Sondhi 1994, Ngo *et al.* 2020). In the methods using the analysis-by-synthesis approach for vocal tract shape estimation, an optimization algorithm is used to estimate the articulatory parameters by minimizing the spectral distance between the speech signal synthesized from the articulatory parameters and the input speech signal. As optimization algorithms find the local minimum of the cost function, initialization of the articulatory parameters plays an important role. Most methods use an articulatory codebook with a table of the corresponding acoustic and articulatory parameters generated using articulatory synthesis for initialization. A review of these methods is presented in this section.

Atal *et al.* (1978) reported a method using a set of five parameters, comprising distance of the place of maximum constriction from the glottis, cross-sectional area at the place of maximum constriction, area of the mouth opening, amount of the lip protrusion, and length of the vocal tract from the lips to the glottis, as the articulatory parameters and the first three formant frequencies as the acoustic parameters. The articulatory codebook was created by sampling the articulatory space at equidistant points within a specified range of the articulatory parameters and obtaining the corresponding acoustic parameters using a transmission line model of the vocal tract as described by Flanagan (1975). For estimating the articulatory parameters from the input acoustic parameters, the codebook was searched for the entry with the smallest distance from the input, and the corresponding articulatory parameters were used for initialization. A refined estimate of the articulatory parameters was obtained using an iterative procedure by searching in the previous estimate's neighborhood to reduce the difference between the synthetic and input acoustic parameters. The method was used for estimating the articulatory parameters for several vowels. The estimated mouth opening for /a/ was always larger than that for /u/, and the estimated area at the maximum constriction was smallest for /i/ and largest for /u/. It was also observed that there were many vocal tract

configurations in the codebook corresponding to the same set of acoustic parameters, indicating the non-uniqueness in the mapping.

Levinson and Schmidt (1983) reported a method using unconstrained optimization to minimize the difference between the natural speech spectra and the corresponding spectra calculated from the articulatory parameters using a lossy transmission line model, which was based on the articulatory model of Coker (1976). The vocal tract was modeled as a soft-walled acoustic tube having varying cross-sectional area, thermal and viscous friction losses, and plane wave propagation. The error function was calculated as the difference between the squared magnitudes of the LP spectral envelope of the input signal segment and the spectrum calculated from the articulatory parameters. Gradient-descent algorithm with no constraints on the vocal tract geometry was used to minimize the error function. Articulatory parameters were initialized to values corresponding to the vocal tract with uniform area function, and they were adjusted to minimize the error function. Evaluation was carried out for isolated vowels and diphthongs spoken by two male and two female speakers. The estimated articulatory configurations were reported to be anatomically reasonable. The distance between the spectrum calculated from the estimated articulatory parameters and the LP spectral envelope for vowels was approximately 2 dB. For diphthongs /OU/ and /eI/, one or more of the articulatory parameters remained fixed resulting in unrealistic articulatory configurations, possibly due to the optimization convergence to the local minimum.

Soquet *et al.* (1990) reported an inversion method using a neural network to obtain the vocal tract area function. The method used one hidden layer with six neurons to map the first three formant frequencies as the acoustic parameters to eight vocal tract area values and the vocal tract length as the articulatory parameters. A transmission line model of the vocal tract reported by Liljencrants and Fant (1975) was used to compute the formant frequencies from the vocal tract area function. The difference between the synthetic and input formant frequencies was used to adjust the network weights. After convergence, the network outputs provided the vocal tract area function for the input formant frequencies. The method was evaluated for 11 French vowels, assuming a vocal tract length of 19 cm. With no constraint on the vocal tract's average volume, the network training converged in 40% cases, and some of the vocal tract area functions were anatomically unrealistic. The use of a volume constraint resulted in convergence in 80% cases and generally reasonable vocal tract area functions.

Panchapagesan and Alwan (2011) studied the acoustic-to-articulatory inversion for non-nasalized vowels using the articulatory model of Maeda (1990) and an articulatory codebook to initialize the articulatory parameters, with the first three formant frequencies as the input acoustic parameters. The palatal outline and the tongue and lip pellets' locations in the

XRMB database (Westbury 1994) were used to evaluate the estimated articulatory parameters. The articulatory model was used to convert articulatory parameters to vocal tract outlines, and the mid-sagittal heights obtained from the vocal tract outlines were converted to area function using a heuristically obtained exponential function. The synthetic formant frequencies were computed from the spectrum estimated from the vocal tract area functions using a hybrid time-frequency articulatory synthesizer proposed by Sondhi and Schroeter (1987). The articulatory parameters were estimated by minimizing the objective function involving the difference between the computed and input formant frequencies. The objective function also included the difference between the current and previous articulatory parameter estimates to maintain the estimated parameters' continuity. The articulatory codebook was designed using a large number of randomly selected articulatory parameters with a minimum vocal tract area greater than 0.05 $cm^2$. For evaluation, the speaker-dependent parameters of the articulatory model were calibrated for a male speaker in the XRMB database by minimizing the difference between the computed and input acoustic parameters for a set of speech frames for three cardinal vowels. There was an average 1.5 mm distance between the estimated vocal tract outlines and pellet locations for several vowels and diphthongs. It was also reported that the average error between the input and computed formant frequencies was less than 1%.

Busset and Laprie (2013) reported an acoustic-to-articulatory inversion with cepstral coefficients as the input acoustic parameters. They used four X-ray films with the synchronized audio recordings of one speaker in DocVacim database (Sock *et al.* 2011) to obtain an affine transformation on the acoustic parameters, for reducing the mismatch between the vocal tract outline obtained from X-ray imaging and that from the articulatory model. The same recordings were also used for performance evaluation. The articulatory model estimated the vocal tract outlines as a linear combination of seven articulatory parameters, including one parameter for jaw opening, four parameters for tongue shape, one for the lips, and one for the larynx. Cepstral coefficients were calculated from the spectrum estimated from the vocal tract area functions using the transmission line model of the vocal tract. An articulatory codebook search was carried out using the second formant as the input parameter. Articulatory parameters were estimated by minimizing the distance between the transformed cepstral coefficients of the input speech signal and the cepstral coefficients obtained from the synthesized speech. An average error of 1 mm was observed in the estimation of articulatory parameters for vowels.

The inversion methods using the analysis-by-synthesis approach have been reported to result in reasonably accurate vocal tract area functions. However, as the optimization

converges to a local minimum, the estimate is very sensitive to the articulatory parameters' initialization. Most of these methods were developed for inversion of only vowel sounds as the articulatory synthesis of these sounds is well established. These methods are not suitable for fricatives as the assumption of plane wave propagation is not valid due to significant energy beyond 4 kHz in the spectrum of these sounds.

## 2.6 Vocal tract shape estimation using machine learning

Methods for vocal tract shape estimation based on machine learning use simultaneously acquired acoustic and articulatory data to obtain the acoustic-to-articulatory mapping. Direct imaging techniques, such as EMA, MRI, XRMB imaging, and ultrasound imaging, have been used to acquire the articulatory data during speech production. Machine learning methods do not use vocal tract models for inversion and are thus not affected by these models' assumptions. With the availability of extensive parallel acoustic-articulatory data, several methods using machine learning have been investigated. A review of these methods is presented in this section.

Hogden *et al.* (1996) reported acoustic-to-articulatory mapping using look-up tables for the EMA data and speech signals from a male speaker producing /CVVC/ utterances with /g/ as the stop consonant and two vowels from a set of nine Swedish vowels and one English vowel. The articulatory vector comprised the *x*-*y* positions of the coils on lips, tongue, jaw, and nose-bridge. The acoustic vector comprised 128 samples of the cepstrally-smoothed spectrum. Three repetitions of 90 utterances were recorded, with the data from one repetition used for training, and the remaining data for testing. A codebook of reference acoustic vectors was created by vector quantization of the training data. For each reference acoustic vector, a corresponding articulatory vector was obtained by averaging the articulatory vectors corresponding to all the acoustic vectors that were quantized to it. A look-up table was created using the reference acoustic vectors and the corresponding articulatory vector obtained by averaging. The articulatory parameters were estimated from the input acoustic vector using vector quantization and look-up table search. The test results showed the coil positions estimated with a correlation of around 0.94 and an RMS error of 2 mm.

Hiroya and Honda (2004) reported an acoustic-to-articulatory inversion using the HMM-based speech production model to utilize the dynamic characteristics of the articulatory movements. Each phoneme was represented by a sequence of HMM states with each state corresponding to an articulatory vector. For each state, a linear mapping was used to obtain the acoustic vector from the articulatory vector. An optimum HMM state sequence was obtained for the input acoustic vector sequence using the Viterbi algorithm. The resulting

state sequence and maximum a posteriori estimation (MAP) were used to estimate the articulatory parameters for the input acoustic vector sequence. To constrain the articulatory trajectory to be smooth, the velocities and accelerations of the articulatory parameters were appended to the articulatory vector. The EMA data with *x*-*y* positions of the coils on lips, tongue, and velum were used as the articulatory parameters, and the first 16 MFCCs were used as the acoustic parameters. The evaluation used 358 sentences from three Japanese male speakers, with 342 sentences for training the HMM using the Baum-Welch algorithm and the remaining 16 sentences for testing. The estimated articulatory parameters had an average RMS error of 1.73 mm.

Toda *et al.* (2008) reported a method for estimating the articulatory movements from the acoustic parameters using a Gaussian mixture model (GMM). The joint probability density function of the acoustic and articulatory parameters from the MOCHA-TIMIT database was modeled as a mixture of multivariate Gaussian density functions. The parameters of 64-component GMM with diagonal covariance matrices were estimated using the expectation-maximization (EM) algorithm. The *x*-*y* positions of the coils on lips, tongue, incisor, and velum were used as the articulatory parameters, and the first 24 MFCCs were used as the acoustic parameters. Articulatory parameters were estimated from the acoustic parameters using maximum likelihood estimation (MLE). Evaluation was performed in a five-fold cross-validation manner, by dividing 460 sentences from two speakers into five sets with four sets for training and the fifth set for testing. The RMS error, averaged over five combinations of training and test data sets, was 1.39 mm.

Richmond *et al.* (2006) used a mixture density network followed by maximum likelihood parameter generation to estimate the articulatory parameters from the acoustic features, for the MOCHA-TIMIT database. The mixture density network estimated the GMM parameters (mean vector, covariance matrices, and weights) conditioned on the input acoustic parameters for each component of mixture density. MLE was used to estimate the articulatory parameters from the parameters of conditional probability density estimated from the networks, with one network trained for each articulatory parameter. The *x*-*y* positions of the coils on the articulators were used as the articulatory parameters, and 20 mel-scale filter-bank energies were used as the acoustic parameters. Every network had one hidden layer with 60 neurons. The acoustic vector comprised a concatenation of the acoustic parameters of 20 frames. The evaluation used 460 utterances by a female speaker, with 368, 46, and 46 utterances for training, validation, and testing, respectively. The average RMS errors for the proposed method and a multilayer perceptron method with a single hidden layer of 38 neurons were 2.02 mm and 2.22 mm, respectively.

Uria *et al.* (2012) investigated two deep neural network architectures for acoustic-to-articulatory inversion using the simultaneously acquired acoustic and EMA data. The first architecture was a deep regression network with five hidden layers, each with 300 neurons. The second architecture used a series of trajectory mixture density networks with six hidden layers, each with 300 neurons, and maximum likelihood parameter generation for estimating the articulatory parameters. The *x-y* positions of the coils on the articulators were used as the articulatory parameters, and 40 LSFs were used as the acoustic parameters. The evaluation used the MNGU0 database (Richmond *et al.* 2011) having 1263 sentences spoken by a male speaker, with 1137, 63, and 63 sentences used for training, validation, and testing, respectively. The training involved pre-training each of the hidden layers sequentially using restricted Boltzmann machines (RBMs) followed by full network training. With pre-training of the hidden layers using the deep regression network, the average RMS error was 0.94 mm. The error increased by 0.05 mm with the network initialized randomly, and it decreased by 0.06 mm with pre-training of the hidden layers using the mixture density networks.

For exploiting the correlation of the articulatory parameter at a frame with the acoustic parameters of the adjacent frames, Liu *et al.* (2015) used a deep bidirectional long short-term memory recurrent neural network (DBLSTM) and a deep recurrent mixture density network (DRMDN) for acoustic-to-articulatory mapping. The investigation used the MNGU0 database (Richmond *et al.* 2011), with the *x-y* positions of the coils on the articulators as the articulatory parameters and 40 LSFs as the acoustic parameters. Both the networks were pre-trained layer-wise for improving accuracy. In DBLSTM, the first two layers were 300-neuron feedforward layers and were used to extract the features. The last two layers were 100-neuron bidirectional LSTMs, and they were used to capture the dynamic property of the training data. DRMDN consisted of DBLSTM followed by a mixture-density output layer. For DRMDN, the mean of mixture component with the largest weight was used as the predicted articulatory parameter value. The evaluation was also carried out for a deep neural network (DNN) with four 300-neuron hidden layers. The DBLSTM, DRMDN, and DNN networks resulted in average RMS errors of 0.81 mm, 0.83 mm, and 1 mm, respectively.

The investigations on acoustic-to-articulatory inversion methods based on machine learning have generally used the EMA databases with the receiver coils' *x-y* positions as the articulatory parameters. The sensed *x-y* positions vary significantly across speakers due to the speaker's physiology and placement of the receiver coils on the articulators. The data do not include the palate information relative to the coil positions and thus are not representative of the vocal tract configuration. Ji *et al.* (2014b) proposed a set of articulatory parameters measured using the palate as a reference and normalized with respect to the speaker's

articulatory space. The articulatory parameters included the vertical distance between the tongue coils and the hard palate, horizontal positions of the tongue coils, the vertical distance between the lips, and the lip protrusion normalized using the distance between the center incisors and the middle point of the back molar. HMM-based acoustic-to-articulatory mapping was obtained using 13 MFCCs, 13 delta coefficients, and 13 delta-delta coefficients as the acoustic parameters. The evaluation used data of 198 sentences spoken by a female English speaker, with 178 and 20 sentences for training and testing, respectively. The normalized RMS error (ratio of the RMS error in a parameter to its standard deviation) in the *x-y* positions was 0.94. It reduced to 0.66 by using the proposed articulatory parameters.

Most of the acoustic-to-articulatory inversion techniques based on machine learning use speaker-dependent mapping, with the training and testing data obtained from the same speaker. Therefore, these techniques do not generalize to the unseen speaker's acoustic data. It is not practical to acquire large amounts of parallel acoustic and articulatory data from each speaker. Ji *et al.* (2016) proposed a method to obtain a new speaker's articulatory model using a small amount of acoustic data, based on the assumption of a high correlation between the speakers' similarity in the acoustic space and that in the articulatory space. The new speaker's acoustic model was obtained as a linear combination of the acoustic models of a set of reference speakers, and the corresponding articulatory model was obtained using a linear combination of the reference speakers' articulatory models, with the weights obtained during the acoustic model estimation. Normalized articulatory parameters measured with the palate as a reference, as used by (Ji *et al.* 2014b), were used as the articulatory parameters, and 13 MFCCs, 13 delta coefficients, and 13 delta-delta coefficients were used as the acoustic parameters. Speaker-dependent HMM models were obtained for each of the 20 English speakers from the EMA-MAE corpus. A set of reference speakers, having good accuracy for acoustic-to-articulatory inversion, were selected for adapting the test speaker data. Articulatory model and acoustic-to-articulatory mapping were obtained for each of the 20 speakers using only a small amount of their acoustic data and by choosing reference speakers among the remaining 19 speakers. The average correlation using the reference speaker weighting was 0.62 and that using the speaker-dependent model was 0.63.

Illa and Ghosh (2018) proposed a method to reduce the amount of speaker-specific parallel acoustic-articulatory data for acoustic-to-articulatory inversion. The EMA and acoustic data for 460 sentences from 30 speakers were divided into two sets: set-1 with data from 8 male and 7 female speakers, set-2 with data from 9 male and 6 female speakers. A generic acoustic-to-articulatory mapping (GBM-AAI) was obtained using a bidirectional long short-memory network (BLSTM) using all the data in set-1. For each speaker in set-2, a speaker-adaptive

acoustic-to-articulatory (SA-AAI) mapping was obtained using a separate BLSTM network starting with weights initialized from the GBM-AAI model, and the model was fine-tuned using the speaker-specific parallel data. The 13 MFCCs were used as the acoustic parameters, and the receiver coils' *x*-*y* positions were used as the articulatory parameters. The sentences recorded for every speaker were divided into sets of 80%, 10%, and 10% for training, validation, and test, respectively. The model for each speaker in set-2 was trained using 12.5%, 37.5%, 62.5%, and 100% of the data. A speaker-dependent acoustic-to-articulatory mapping (SD-AAI) was trained for each of the 30 speakers separately for comparing the results. Only 37.5% of the speaker-specific data was required for SA-AAI to match the SD-AAI performance, with an average RMS error of 1.36 mm. With 62.5% and 100% of the speaker-specific data, SA-AAI performed better than SD-AAI. The results indicated that initializing the adaptive model with the weights obtained from training on a set of reference speakers could improve the performance.

The use of machine learning methods for acoustic-to-articulatory inversion has become popular after the availability of reasonably large quantities of parallel acoustic-articulatory data. The methods employing deep neural networks estimate the articulatory parameters with reasonable accuracy for speaker-dependent mapping, but they do not generalize well to an unknown speaker's acoustic data. It has also been reported that the use of the EMA receiver coils' *x*-*y* positions as articulatory parameters results in increased variability across speakers leading to more non-uniqueness in the acoustic-to-articulatory mapping.

## 2.7 Summary

Visual feedback of the articulatory efforts not visible on the speaker's face can improve the articulation of children with hearing impairment. Speech training aids for this type of feedback require estimation of the vocal tract configuration. The methods employing direct imaging and acoustic measurements are expensive and interfere with speech production, and they are thus not suitable for speech training. For this purpose, the methods based on processing the speech signal are considered suitable. LP-based estimation of the vocal tract area function is commonly used for improving vowel articulation, but it is not suitable for nasals and fricatives due to the presence of zeros in the spectral envelope. Analysis-by-synthesis methods estimate the vocal tract area function during vowel articulation with reasonable accuracy, but not during the fricative utterances. Methods employing machine learning have become popular after the availability of large quantities of parallel acoustic-articulatory data. These methods are mostly based on the EMA data and provide reasonably good accuracy for speaker-dependent mapping for all types of sounds. However, they do not

generalize to the acoustic data from an unseen speaker. Also, estimation of the EMA receiver coils' *x*-*y* positions as articulatory parameters may not be suitable for speaker-independent acoustic-to-articulatory inversion. Further investigations may improve the accuracy of the speaker-independent mapping from the spectral parameters to the articulatory parameters as needed for speech training.

# Chapter 3

# PLACE OF ARTICULATION OF FRICATIVES FROM SPECTRAL PARAMETERS DURING FRICATION SEGMENTS

## 3.1 Introduction

Fricatives are produced by airflow through a narrow constriction in the oral cavity. Obstruction of the airflow by the constriction causes turbulence in its vicinity, and the resulting noise-like excitation of the oral cavity is known as frication. Based on the place of articulation, English fricatives are grouped into four classes: (i) labiodental (/f/ as in "fine", /v/ as in "vine"), (ii) linguodental (/θ/ as in "thing", /ð/ as in "then"), (iii) alveolar (/s/ as in "sue", /z/ as in "zoo"), and (iv) palatal (/ʃ/ as in "shoe", /ʒ/ as in "measure") (Ladefoged 1982, O'Shaughnessy 2000). The fricative /h/ with semi-vowel-like characteristics is generally considered a voiceless equivalent of the adjacent vowel.

The methods proposed to estimate the place of articulation and other articulatory parameters during fricatives may be broadly grouped as the methods using model-based analysis-by-synthesis and those based on machine learning. In the analysis-by-synthesis methods, the articulatory parameters are estimated by minimizing the spectral distance between the synthesized and input speech signals (Shirai and Masaki 1983, Riegelsberger 1997). Shirai and Masaki (1983) used the vocal tract modeled as a concatenation of acoustic tubes with varying cross-sectional areas and excited by a white noise source at the place of maximum constriction to synthesize the fricatives. A fixed vocal-tract function was used as the initial estimate, and the spectral distance of the synthesized fricative from the input fricative, measured using 64 samples (0−10 kHz) of the log magnitude spectrum obtained for 12.8-ms speech segments, was iteratively minimized to estimate the articulatory parameters. The spectral distance between /s/ and /ʃ/ synthesized from the estimated vocal tract area functions was comparable to that between the corresponding input signals. Riegelsberger (1997) used a hybrid time-frequency articulatory synthesizer proposed by Sondhi and Schroeter (1987) and weighted cepstral distortion measure to estimate the articulatory parameters, with an acoustic-to-articulatory codebook to initialize the inversion. The results showed that the values of the place of maximum constriction estimated for English fricatives spoken by two male and two female speakers were significantly distinct across different classes. Even though the analysis-by-synthesis methods result in reasonable vocal tract shapes, the codebook searches do not always yield initial parameter sequences for the optimization process to converge to the global minima (Panchapagesan and Alwan 2011).

Most of these approaches assume plane wave propagation in the vocal tract, which may not be valid during the fricative utterances with significant energy in the high-frequency region (Sondhi 1979).

Several machine learning methods have been proposed for acoustic-to-articulatory mapping, using hidden Markov model (HMM), Gaussian mixture model (GMM), and artificial neural network (ANN) (Hiroya and Honda 2004, Richmond 2006, Toda *et al.* 2008, Uria *et al.* 2012, Liu *et al.* 2015). These methods use MFCC or LSF vectors as the input acoustic features, articulator positions as the output parameters, and large quantities of simultaneously acquired acoustic and articulatory data for training. These methods are reported to work well for speaker-dependent mapping, but speech-training applications require speaker-independent mapping. The use of articulator positions as the articulatory parameters results in increased across-the-speakers variability, leading to more non-uniqueness in the acoustic-to-articulatory mapping (Ji *et al.* 2014b, Afshan and Ghosh 2015).

Earlier studies have investigated the importance of several acoustic characteristics, including spectral peak location, spectral moments, spectral slope, duration, and band energies, for characterizing the place of articulation of fricatives (Hughes and Halle 1956, Heinz and Stevens 1961, Baum and Blumstein 1987, Forrest *et al.* 1988, Shadle and Mair 1996, Jongman *et al.* 2000, Nissen and Fox 2005, Li *et al.* 2012, Zharkova 2016, Anjos *et al.* 2020). These studies have used fricatives belonging to a particular place group as having the same place of articulation. However, the place of articulation varies with the length of the speaker's oral cavity, and there are context-dependent and utterance-to-utterance variations. Thus, there is a need to study the relationship between the spectral parameters computed from the speech signal and the place of articulation as obtained using an imaging technique. An investigation for speaker-independent acoustic-to-articulatory mapping is carried out using simultaneously acquired speech signals and articulograms displaying the motion of articulators for fricatives spoken by a large number of speakers in the XRMB database (Westbury 1994). The relationship between the place of articulation of the fricative segment of the utterances measured from the articulograms and the corresponding spectral characteristics is analyzed, using some of the earlier reported spectral parameters and a few proposed parameters. An ANN-based method is investigated for speaker-independent estimation of the place of articulation from the spectral parameters.

The following section provides a review of the studies on the acoustic characteristics of the fricatives. A description of the automated technique for the estimation of the axial curve and place of articulation from oral cavity contours and the investigation relating the place of articulation with the spectral parameters are presented in Sections 3.3 and 3.4, respectively.

An ANN-based estimation of the place of articulation and the results are presented in Sections 3.5 and 3.6, respectively, followed by a discussion in the last section.

## 3.2 Acoustic characteristics of the fricatives

The place of maximum constriction in the oral cavity divides the oral cavity into two cavities. The cavity behind the constriction, known as the back cavity, acts as a trap for the turbulence and contributes resonances and anti-resonances. The cavity in front of the constriction, known as the front cavity, is excited by the turbulence and introduces resonances or peaks in the spectrum. It has been reported that the back cavity does not have a significant effect on the signal spectrum due to a tight constriction during fricative production (Heinz and Stevens 1961, Johnson 2003). Several acoustic parameters, including the spectral peak location, spectral moments, spectral slope, relative amplitude, relative energy in spectral bands, MFCC, locus equations of the second formant (F2) frequency of the adjacent vowel, and duration have been studied for characterizing the place of articulation of fricatives. A review of these studies is presented in this section.

Hughes and Halle (1956) analysed the spectra of the fricatives /f, v, s, z, ʃ, ʒ/ and reported that the palatal fricatives with the longest front cavity had spectral peaks at 2.5–3 kHz, and the alveolar fricatives with a shorter front cavity had spectral peaks at 4–5 kHz. Relatively flat spectra without dominant spectral peaks characterized the labiodental fricatives with very short front cavity. They developed a procedure to identify the fricatives using three spectral energy measurements on the dB scale: energy in the 0.72–10 kHz band with reference to that in the 4.2–10 kHz band, energy in the 0.72–6.5 kHz band with reference to that in the 0.72–2.1 kHz band, and energy in the 500 Hz band centered at the spectral peak in 1.5–4 kHz band with reference to that in the 0.72–1.3 kHz band. The procedure to identify fricatives was developed based on the examination of these energy measurements for several utterances. Utterances with first energy measurement lower than 2 dB were identified as /s/ if the second energy measurement was larger than 10 dB and were identified as /f/ if the second energy measurement was lower than 5 dB. Utterances with first energy measurement larger than 2 dB were identified as /ʃ/ if the third measurement was larger than 5 dB and were identified as /f/ if the third measurement was lower than 2 dB. The fricative identifications showed a good agreement with the responses by human listeners.

Heinz and Stevens (1961) showed that spectra of the unvoiced fricatives /f, θ, s, ʃ/ could be characterized by two complex-conjugate pole pairs and one complex-conjugate zero pair. The locations of poles and zeros were related to the length of the front cavity and the constriction, respectively. The pole locations were found to be dependent on the speaker and the vowel

following the frication. The /f, θ/ spectra were characterized by a broad low-frequency noise in addition to the high-frequency poles. The sounds synthesized using an electrical circuit excited by white noise and characterized by one complex-conjugate pole pair and one complex-conjugate zero pair, with the pole frequency varied from 2 kHz to 8 kHz, were perceived as /ʃ/ for the lowest pole frequency. The perception shifted to /s/ and then to /f, θ/ as the pole frequency increased, indicating the pole frequency being related to the front cavity length.

Baum and Blumstein (1987) reported that the voiced fricatives had smaller overall frication duration than the unvoiced fricatives. However, analysis of consonant-vowel (CV) syllables involving fricatives /f, v, θ, ð, s, z, ʃ, ʒ/ and vowels /i, e, a, o, u/ from three male speakers showed a significant overlap between the durations of voiced and unvoiced fricatives. The frication duration was found to be not an indicator of the place of articulation.

For quantifying the spectral shape, Forrest *et al.* (1988) used spectral moment analysis by treating the squared magnitude spectrum as a probability density function. The first four moments were calculated, with the centroid indicating the average energy-concentration location, the variance indicating the spectral bandwidth around the centroid, the skewness indicating the asymmetry, and the kurtosis indicating the peakiness of the spectrum. A discriminant analysis using the moments for the unvoiced fricatives /f, θ, s, ʃ/, in the syllable-initial position and spoken by five male and five female speakers, resulted in a good classification of sibilant fricatives (85% for /s/ and 95% for /ʃ/), while non-sibilants were poorly classified (58% for /θ / and 75% for /f/).

Jongman *et al.* (2000) studied several static and dynamic acoustic characteristics of the English fricatives /f, v, θ, ð, s, z, ʃ, ʒ/ from 20 speakers for classification of the place of articulation. The static characteristics included spectral peak location, spectral moments, frication duration, normalized amplitude, and F2 onset frequency. The dynamic characteristics included relative amplitude and F2 locus equations. The frequency with highest amplitude in the spectrum was used as the spectral peak location, and the spectral moments were calculated as by Forrest *et al.* (1988). Normalized amplitude was calculated as the frication noise RMS in dB with reference to the adjacent vowel RMS over three consecutive pitch periods at the point of maximum vowel amplitude. Relative amplitude was calculated as the fricative spectral magnitude in dB with reference to the adjacent vowel at the frequency corresponding to third formant for sibilants and fifth formant for nonsibilants. Locus equations were obtained as a linear regression of F2 at the vowel onset and at the vowel midpoint. It was reported that the spectral peak location, the spectral moments, the normalized amplitude, and the relative amplitude could discriminate all four places of

articulation, but F2 onset and locus equations could not. Classification using discriminant analysis resulted in good accuracy (88%) for sibilants but low accuracy (66%) for nonsibilants using spectral peak location, spectral moments, normalized amplitude, and relative amplitude.

Shadle and Mair (1996) studied the spectral bandwidth's effect in calculating the spectral parameters for discriminating the fricatives /f, v, θ, ð, s, z, ʃ, ʒ/. Spectral parameters were calculated for six bands (50–16950 Hz, 50–10000 Hz, 50–5000 Hz, 200–16950 Hz, 200–10000 Hz, 100–6000 Hz) for two sets of utterances spoken by two speakers. The first set comprised sustained fricatives preceded by vowel /a/, and the second set comprised ten repetitions of vowel-fricative-vowel utterances with vowels /a, i, u/. Spectral moments were calculated as by Forrest et al. (1988), and dynamic amplitude was calculated as the minimum spectral magnitude in 0–2 kHz expressed in dB with respect to the maximum spectral magnitude in 0.5–17 kHz. Spectral slope was calculated by fitting a line to the spectrum magnitude values expressed in dB in the corresponding spectral bands. Spectral moments showed significant band-dependent variation, and they did not reliably discriminate between the fricatives. Dynamic amplitude could discriminate between the sibilants and nonsibilants.

Nissen and Fox (2005) studied the acoustic characteristics of the unvoiced fricatives /f, θ, s, ʃ/ by calculating first four spectral moments, normalized amplitude, fricative duration, and spectral slope for CV utterances with syllable-initial fricatives and the vowels /a, i, u/, produced by 10 adult speakers and 30 typically developing children of 3−6 years. Each of the parameters showed variation as a function of the categorical place of articulation, but only the spectral variance could discriminate all the four articulation places for all speakers. Spectral slope (slope of the regression line fitted on spectra), spectral centroid, and spectral skewness could discriminate three articulation places with labiodental and linguodental having similar values for these parameters. The discrimination between /s/ and /ʃ/ increased with age, with a significant change at approximately five years.

Todd *et al.* (2011) used the spectral centroid and peak locations to study the differences between the fricative production of 43 children of 2−7 years with normal hearing and 39 children of 4−9 years with cochlear implants. Nine utterances with /s/ in the initial position and nine utterances with /ʃ/ in the initial position with a vowel chosen from /a, i, u/ after the initial fricative were recorded from each of the children. The centroids and the peak locations of the fricative /s/ for children with cochlear implants were lower than those for children with normal hearing. However, these values of the fricative /ʃ/ were similar for the two groups, indicating that the children with cochlear implants produced the word-initial fricatives /s/ and /ʃ/ with less contrast than the normal-hearing children.

Li *et al.* (2012) conducted perceptual experiments to isolate cues for the perception of /f, v, θ, ð, s, z, ʃ, ʒ/, using a '3-dimensional deep search method' involving stimuli obtained by modifying the duration, frequency range, and signal-to-noise ratio. The stimuli included 48 CV utterances, each with a fricative from /f, v, θ, ð, s, z, ʃ, ʒ/ followed by vowel /a/ from three male and three female speakers. For the minimum duration for fricative identification, the speech stimuli were obtained by truncating the utterances from the end of CV transition in eight 5-ms steps, followed by twelve 10-ms steps, and subsequently 20-ms steps until the onset of the consonant. For the frequency range for fricative identification, the speech stimuli were obtained by filtering the utterances using nine highpass and nine lowpass filters with cutoff frequency variation of 0.25−8 kHz. For checking robustness to white noise, the speech stimuli were obtained by adding white noise to the utterances at eight signal-to-noise ratios from −21 dB to +12 dB. The required duration, frequency range, and signal-to-noise ratio were obtained as the respective values for 90% identification. The perceptual experiments involved listening tests on 12 subjects for obtaining minimum duration, 18 subjects for obtaining the frequency range, and 24 subjects for checking the robustness to white noise. The perceptual experiment results showed that the required frequency ranges for alveolars, palatals, and labiodentals were 3.6−8 kHz, 1.4−4.2 kHz, and 0.5−2 kHz, respectively. The minimum durations for the palatals, alveolars, and labiodentals were 110, 105, and 65 ms, respectively. The minimum signal-to-noise ratios for the alveolars, palatals, and labiodentals were +5, −1, and +10 dB, respectively.

Zharkova *et al.* (2016) studied the variability in tongue shapes obtained from ultrasound imaging and spectral characteristics of the fricatives /s, ʃ/ produced by 15 children of 10–12 years and 15 adults. The spectral centroid during frication and the F2 frequency at the fricative offset were calculated for the CV utterances with the fricatives /s/ and /ʃ/ and the vowels /a/ and /i/. The tongue shape was characterized by the amount of tongue excursion in relation to the ends of the tongue curve (termed as bunching) and the location of the most bunched part of the tongue measured from the end of the tongue curve. The spectral centroid during frication of /s/ was higher than that of /ʃ/ across the speakers and the vowel contexts. The F2 frequencies at the fricative offset for /ʃ/ were significantly higher than those for /s/. The tongue bunching location was better than the amount of tongue bunching in discriminating between the two fricatives and consistently correlated with the spectral centroid. However, the tongue bunching location was measured only using the tongue curve and did not provide information about the place of constriction relative to the palate.

In summary, several acoustic characteristics have been shown to be related to the place of articulation of the fricatives. Spectral moments have been used to characterize the place of

articulation, despite their variation with the spectral bandwidth used for their computation. Spectral peak location, spectral slope, normalized amplitude, relative amplitude, and F2 frequency at the vowel onset also have been reported to be good indicators of the place of articulation. However, computations of the relative amplitude and F2 frequency at the vowel onset require an accurate formant tracking. It may be noted that the studies on acoustic characteristics of fricatives relate the spectral characteristics with the categorical place of articulation instead of the place of articulation obtained from direct imaging. For use in speech training of the persons with hearing impairment and the second language learners, the changes in the place of articulation of the speaker need to be estimated to provide visual feedback. Therefore, an investigation is carried out for relating the place of articulation obtained from direct imaging with the spectral parameters. An automated technique for the estimation of the place of articulation from oral cavity contours is presented in the following section, followed by an investigation relating the place of articulation with the spectral parameters using earlier reported parameters and proposed parameters. Subsequently, an investigation is presented to estimate the place of articulation from the input spectral parameters using an ANN-based model with the place of articulation values estimated from direct imaging as the reference.

## 3.3 Estimation of place of articulation from oral cavity contours

The place of articulation corresponds to the position of maximum constriction, i.e. the smallest opening, in the oral cavity measured from the lips. Its estimation from direct imaging of the oral cavity during speech production can be used to validate the estimation from the speech signal, and it can be used as a reference for acoustic-to-articulatory mapping based on machine learning. Direct imaging provides upper and lower contours of the oral cavity. Irregular shapes of the contours cause difficulty in consistently locating the smallest opening and finding its distance from the lips. As a solution to this problem, an automated technique using graphical processing of the contours is presented. It iteratively estimates an axial curve between the two contours, uses the length of the normal to this axial curve between the contours as a measure of the oral cavity opening, and estimates the place of articulation as the distance along the axial curve between the lips and the smallest oral cavity opening.

A review of the earlier techniques for estimating the place of articulation from oral cavity contours obtained by direct imaging, the proposed technique to estimate the place of articulation from oral cavity contours, and test results are presented in the following subsections.

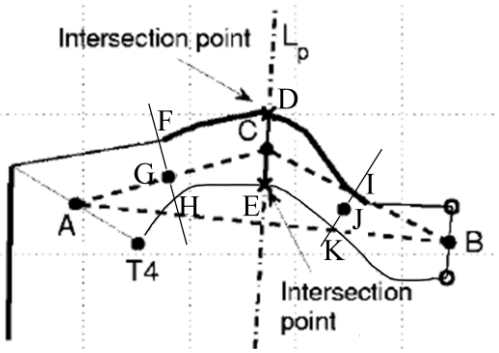



**Figure 3.1** Iterative bisection method used by Story (2007).

### *3.3.1 Earlier methods for estimating the place of articulation from oral cavity contours obtained by direct imaging*

Several methods have been reported to estimate the oral cavity shape and place of articulation from the images in the direct imaging databases (Story 2007, Bresch *et al.* 2006a, Jagabandhu 2012). In these methods, an axial curve of the oral cavity is estimated from the images. The oral cavity opening is measured as the distance between the upper and lower contours along the normal to the axial curve. Story (2007) used an iterative bisection method to obtain an axial curve of the oral cavity contours from the articulograms in the XRMB data. As shown in Figure 3.1, the points A and B are marked as the midpoints between the upper and lower contours at the two ends of the oral cavity. The iterative segmentation by bisection is started using the line segment AB. The oral cavity is segmented by perpendicular bisector Lp of the line segment AB, cutting the upper and lower contours at D and E, respectively. The segment DE is taken as the mouth opening at the interface of the two segments. The midpoint C of the segment DE provides the line segments AC and CB. Repetition of the segmentation provides the four segments AG, GC, CI, and IB. The process is iterated for obtaining the desired numbers of segments. The technique was applied, with five iterations giving a 33-point axial curve, on the oral cavity shapes for 11 vowels and vowel-vowel sequences from four speakers. Bresch *et al.* (2006a) applied the iterative bisection method, with three iterations giving a 9-point axial curve and a cubic spline for smoothing the curve, on the oral cavity contours from the MRI images. Oral cavity configuration for the vowel /a/ was obtained and the corresponding resonant frequencies, estimated using VTAR software (Zhou *et al.* 2004), matched well with those obtained from the speech signal.

Jagbandhu (2012) used a segmentation method for estimating the place of articulation from the articulograms in the XRMB database. In this method, the upper and lower contours are divided into 50 equal segments. The midpoints of line segments joining the corresponding points on the two contours are obtained, and the line segments joining these points form an

axial curve. The normal to the axial curve at a point is drawn by using the mean of the slopes of the left and right segments as the slope of the curve. The length of the normal between the two contours marks the oral cavity opening at the point. The method was applied for finding the place of articulation for /ΛCa/ utterances with stop consonants /b/, /d/ and /g/ from 35 speakers. The results were validated by comparing them with those obtained by manual measurement along the oral cavity length, using the 'tongue profile' method (Pandey and Shah 2009) by adding the straight-line distances along the curve joining the lower lip marker and the pellet markers.

The earlier methods (Story 2007, Bresch *et al.* 2006a, Jagabandhu 2012) were used to estimate the axial curves of the oral cavity contours from MRI images to compare the estimations. The oral cavity contours on the MRI image, as used in the investigation by Bresch *et al.* (2006a), were manually marked, and axial curve estimates for /a/ are shown in Figure 3.2 as an example. In Figure 3.2(a) for the iterative bisection method (Story 2007), the normals to the axial curve (dot-dashed trace) deviate from the bisector lines (dark trace). This method starts with a global slope and estimates the local axial slopes in subsequent iterations. Hence, it may not correctly estimate the local axial slopes at bisectors in the initial iterations. Further, the bisector lines after a few iterations start crisscrossing at sharp bends in the axial curve. The sharp bends are avoided in the method used by Bresch *et al.* (2006a), by limiting the number of iterations and applying a smoothing spline, as shown in Figure 3.2(b). A significant deviation between the smoothed axial curve (dot-dashed trace) and the curve joining midpoints of the normals (dark trace) indicates that the axial curve may not divide the oral cavity symmetrically. Figure 3.2(c) shows the result of the segmentation method (Jagabandhu 2012). The estimated axial curve (dot-dashed trace) deviates from the curve joining midpoints of the normals (dark trace) in the region where the two contours are not symmetric. Similar problems were observed in the application of these methods on many other images. Hence, there is a need for an automated technique for estimating the axial curve that approximately divides the normals into two equal segments, such that the deviation between the axial estimate and curve joining midpoints of the normals is minimal.

### *3.3.2 Proposed technique*

An automated technique is developed for estimating the place of articulation by graphical processing of the upper and lower contours of the oral cavity. It assumes the acoustic wave

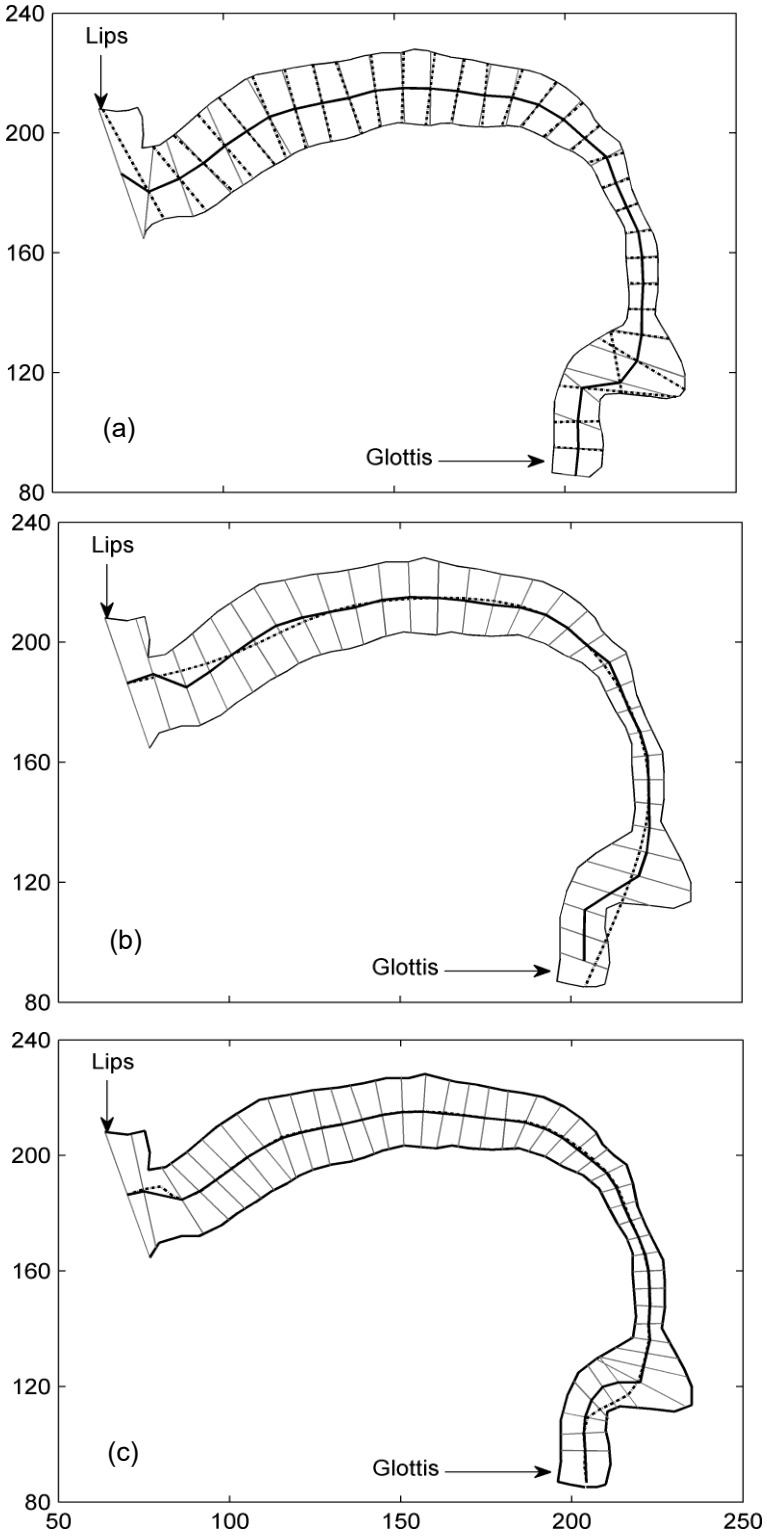



**Figure 3.2** Estimation of oral cavity opening from the lower and upper contours in an MRI image for /a/ using (a) iterative bisection method (Story 2007), (b) iterative bisection method and smoothening spline (Bresch *et al.* 2006a) with smoothening factor of 0.99, and (c) segmentation method (Jagabandhu 2012), with the *x* and *y* distances in number of pixels.

propagation to be along an axial curve between the two contours with the normal to the curve representing the wavefront. The axial curve is iteratively estimated as an axis of symmetry, approximately bisecting the normals to it. The length of the normal between the two contours provides an estimate of the oral cavity opening.

A set of equidistant points are marked on the lower and upper contours. Midpoints of the lines joining the corresponding points on the two contours form the initial estimate of the axial curve. A B-spline based on least-squares approximation is fitted on these points for smoothing the curve by reducing sharp bends. For this purpose, the Matlab function 'spap2', with an internal generation of the knot vector for a specified number of knots is used. The smoothing controls the curve locally, where it does not bisect the normals without significantly affecting it at other places. The normals are drawn at a set of equidistant points on the smoothed curve. Points of intersection of the normals with the upper and lower contours are obtained, and their midpoints form the revised estimate of the axial curve. The process of smoothing and revising the axial curve is repeated to improve the approximation until the RMS difference between the successive estimates is less than a specified fraction of the mean oral cavity opening.

An increase in the number of knots in the B-spline decreases the approximation error but reduces the axial curve's smoothing, leading to crisscrossing of the normals at sharp bends. The iterations are started with the knot vector length of 12 and stopped if the RMS difference is less than 5% of the mean oral cavity opening and the number of crisscrossing is five or less. If this criterion is not satisfied after 30 iterations, the knot vector length is decremented by one and the iterative process is repeated. It was found that the condition for stopping the iterations was satisfied with a knot vector length of five or higher.

Figure 3.3 shows an example of the processing for /a/. Figure 3.3(a) shows the initial axial curve as a continuous curve and the revised estimate as a curve with circle markers. The final iteration result is shown in Figure 3.3(b), with the estimated axial curve as the dot-dashed trace and curve formed by joining midpoints of the normals as the dark trace. The two curves are almost superimposed, the deviation being much smaller than the corresponding deviations observed in Figure 3.2 as obtained by the earlier methods.

The axial curve and oral cavity opening can also be estimated from the articulograms in the XRMB database. These articulograms show three reference pellet points (Ref), four pellet points (T1−T4) on the tongue, one pellet point each on the upper lip (UL), the lower lip (LL), and the incisor (MANi), the palatal outline, and the posterior pharyngeal wall as shown in Figure 3.4. Two reference pellet points on the nose and one reference pellet point on the buccal surface of the maxillary incisors were used to remove the head motion effect on the

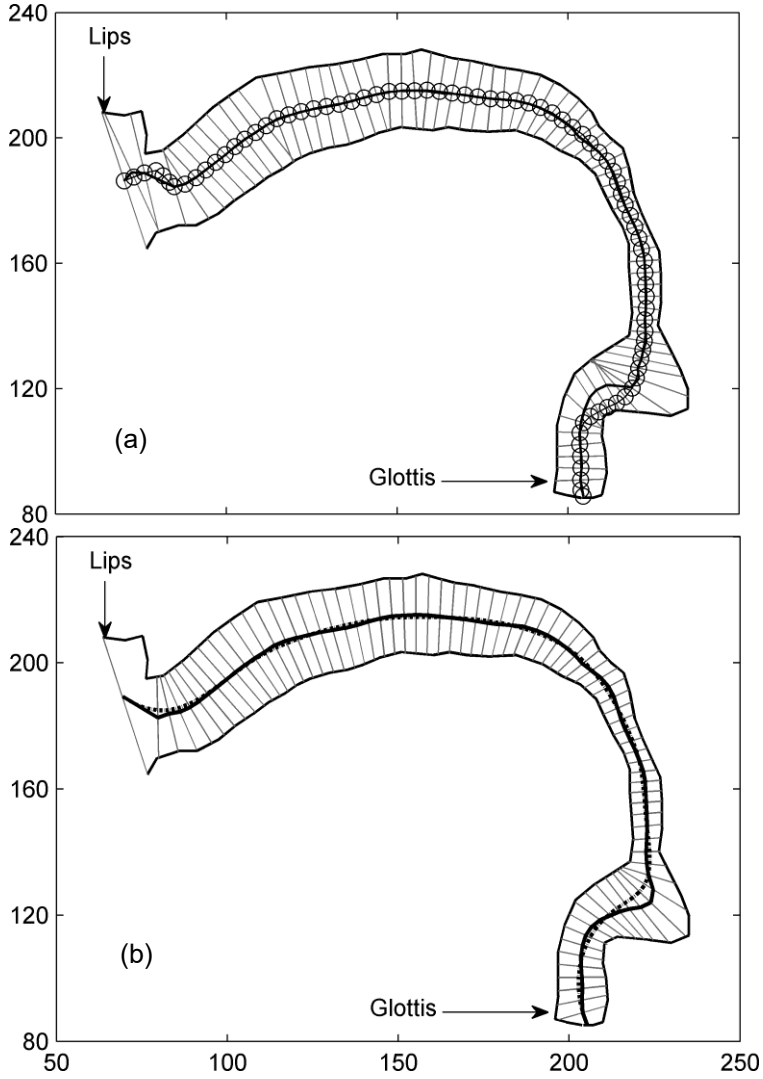



**Figure 3.3** Application of the proposed Iterative Axial Curve method for estimation of the oral cavity opening from the lower and upper contours in an MRI image of /a/: (a) first iteration, (b) final iteration, with the x and y distances in number of pixels.

position of other pellets. The origin of the plot is centered on the space between the two central maxillary incisors (CMI). The x-axis is intersection of the mid-sagittal plane and the maxillary occlusal plane (MaxOP) formed by tips of the central incisors and two other maxillary teeth on opposite sides of the mouth. The y-axis is along the CMI normal to the MaxOP plane at the origin. For removing the offset in measurement of the lip pellet points due to the pellet positioning, a set of phantom points were obtained for the upper and lower lips by adding an offset calculated as half of the distance between the upper lip pellet point and the lower lip pellet point, for the utterance /Λba/ during the closure. A piecewise cubic Hermite interpolation was used to connect the upper phantom point with the palatal outline to form the upper contour and to connect the lower phantom point with the incisor and tongue

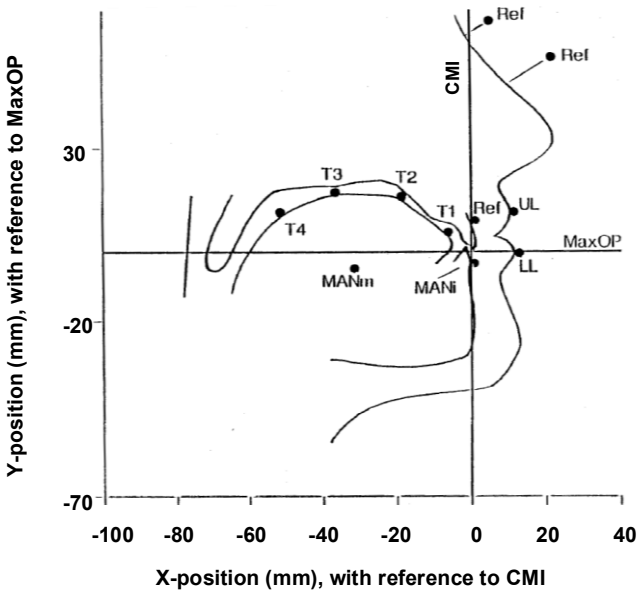



**Figure 3.4.** Position of pellet points in the XRMB database (Westbury 1994).

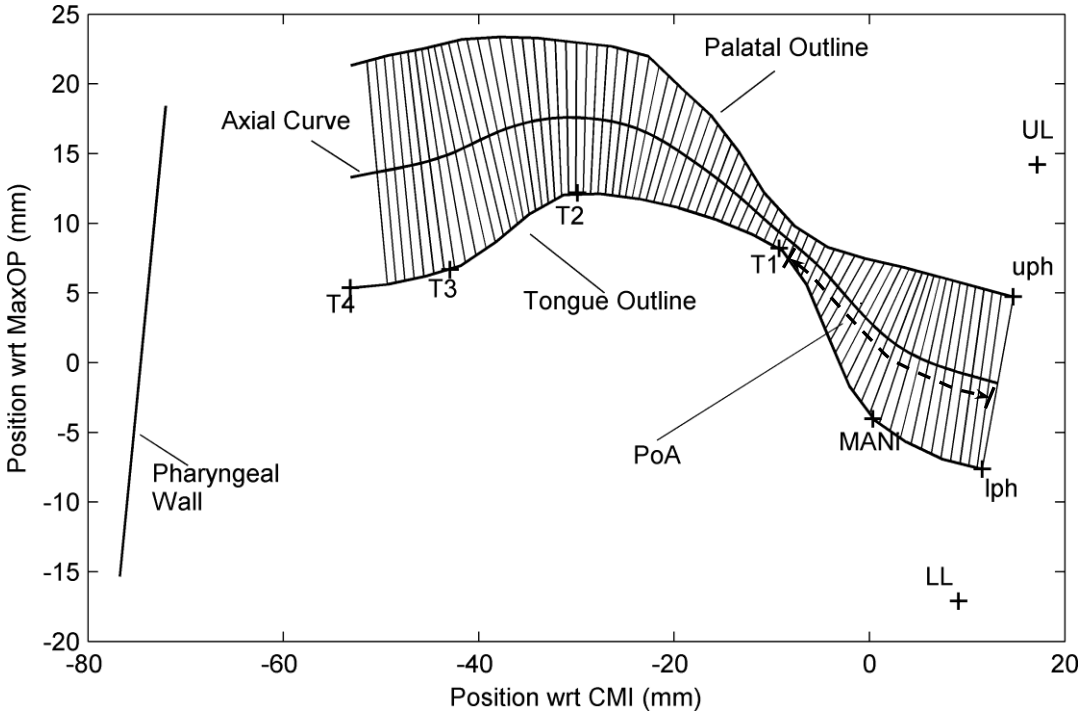



**Figure 3.5** Example of axial curve estimation from the pellet points (UL, LL, MANi, T1, T2, T3, T4).

pellet points to form the lower contour. These contours form an oral cavity image with the interpolation providing the smooth contour between the pellet points.

Figure 3.5 shows an example of the estimated axial curve with normal drawn to it. The pellet points are the corresponding points in Figure 3.4, with the points 'uph' and 'lph' are the upper and lower phantom points, and PoA is the place of articulation measured from lips.

### *3.3.3 Test results*

The proposed technique was evaluated for estimating the place of articulation. The test material comprised oral cavity images for 8 phonemes (3 vowels, 3 stops, and 2 fricatives) by two speakers (one male, one female) from the XRMB database (Westbury 1994) and two speakers (one male, one female) from the MRI database (Narayanan *et al.* 2014). Estimation accuracy was examined with reference to the manually marked and measured values of the visually estimated place of maximum constriction in prints of the oral cavity images. An axial curve was traced for symmetrically dividing the cavity into upper and lower halves, and the place of maximum constriction was marked on this curve. The distance of the maximum constriction from the lips along the axial curve was measured as the place of articulation. The distances measured on the print were scaled back to mm in the XRMB articulograms and to pixels in the MRI images. Means of the values as measured by five observers (fellow researchers in the lab) were used as the reference values.

A consistent manual marking of the place of articulation for the medial vowel /a/ was difficult due to a relatively large opening and lack of a well-defined place of maximum constriction. The places of maximum constriction could be consistently marked for the other seven phonemes. For these phonemes, the values of the place of articulation obtained by manual measurement and the estimation errors using the proposed automated technique are given in Table 3.1. For comparing the proposed technique with the earlier techniques for estimating the place of articulation, the estimation errors using the techniques proposed by Story (2007) and Bresch *et al.* (2006a) are also given in Table 3.1. Very small standard deviations of the manually measured values indicate consistency of manual marking across observers. A relatively large standard deviation for /i/ from Speaker 2 is due to two nearly similar places of maximum constriction. The standard deviation for the bilabial stop /p/ with the place of articulation visible at the lips is negligible.

The techniques by Story (2007) and Bresch *et al.* (2006a) result in errors comparable to standard deviation of manually measured values for some utterances and higher error for many of the utterances. The errors using the proposed techniques are comparable to the standard deviation of the manually measured values for all the utterances, indicating a close match between the estimated and manually measured values for utterances both the databases. The large error for /i/ from Speaker 2 was due to two nearly similar places of maximum

**Table 3.1**: Place of articulation: (a) manually measured (mean and standard deviation (S.D.) of the values obtained by 5 observers) and (b) estimation error using the three automated graphical techniques (T1: proposed iterative axial curve technique, T2: technique proposed by Story (2007), T3: technique proposed by Bresch *et al.* (2006a)).

(a) Place of articulation for the XRMB utterances in mm

| Pho-neme | Speaker 1 (F, XRMB) | | | | | Speaker 2 (M, XRMB) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Manually Measured | | Estimation Error | | | Manually Measured | | Estimation Error | | |
| | Mean | S.D. | T1 | T2 | T3 | Mean | S.D. | T1 | T2 | T3 |
| /i/ | 53.9 | 0.8 | 0.3 | 3.5 | 3.8 | 35.8 | 10.2 | 17.9 | −5.0 | −4.5 |
| /u/ | 60.2 | 0.7 | 0.3 | 1.4 | 3.6 | 69.4 | 0.6 | 0.8 | −0.6 | 7.5 |
| /p/ | 0.0 | 0.0 | 0 | 0.0 | 0.9 | 0.0 | 0.0 | 0 | 0.0 | 0.9 |
| /t/ | 30.7 | 0.4 | −0.4 | −4.1 | −3.7 | 23.3 | 0.5 | −0.4 | 0.3 | 0.9 |
| /k/ | 63.4 | 0.4 | −0.2 | −3.7 | −3.2 | 61.9 | 0.4 | 0.2 | 0.2 | 1.2 |
| /s/ | 27.2 | 1.9 | −2.4 | −3.1 | −1.5 | 23.3 | 0.5 | −0.7 | −1.9 | −0.9 |
| /ʃ/ | 31.9 | 0.7 | 0.8 | −0.5 | −0.1 | 26.1 | 0.5 | −0.7 | 0.6 | 1.7 |

b) Place of articulation for the MRI utterances in pixels

| Pho-neme | Speaker 3 (F, MRI) | | | | | Speaker 4 (M, MRI) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Man. | | Err. | | | Man. | | Err. | | |
| | Mean | S.D. | T1 | T2 | T3 | Mean | S.D. | T1 | T2 | T3 |
| /i/ | 47 | 4 | 0 | 4 | 1 | 108 | 2 | −3 | −4 | 3 |
| /u/ | 124 | 1 | 2 | 5 | 4 | 108 | 2 | −2 | −3 | 4 |
| /p/ | 8 | 1 | 2 | 0 | 6 | 6 | 1 | −1 | 4 | 2 |
| /t/ | 36 | 2 | 0 | 2 | 3 | 49 | 5 | 0 | 3 | 4 |
| /k/ | 135 | 7 | −2 | 3 | 8 | 119 | 4 | 0 | −4 | 5 |
| /s/ | 36 | 2 | −1 | 3 | 5 | 47 | 1 | −1 | −1 | 1 |
| /ʃ/ | 47 | 1 | 3 | 0 | 3 | 60 | 1 | 1 | −1 | 5 |

constriction for this utterance, but such errors did not occur in the case of stops and fricatives. The manual and proposed automated marking are both based on the same approach of obtaining an axis of symmetry. Thus, the close match between the estimated values and manually measured ones may be attributed to the success of the graphical iteration technique in approximating the axial curve obtained by visual tracing.

The investigation relating the place of articulation values estimated using the automated technique with the spectral parameters is presented in the next section.

## 3.4 Spectral parameters for estimating the place of articulation

The investigation relating the spectral parameters with the place of articulation obtained from direct imaging requires simultaneously acquired speech signal and articulatory information from a large number of speakers. Several databases providing the simultaneously acquired audio signals and articulatory data are available, including X-ray film database (Munhall *et al.*

1995), MOCHA-TIMIT database (Wrench and Hardcastle 2000), XRMB database (Westbury 1994), MNGU0 database (Richmond *et al.* 2011), and USC-TIMIT database (Narayanan *et al.* 2014). The X-Ray film database provides 55-minute simultaneously acquired X-ray imaging and audio signals for sentences spoken by five English speakers and nine French speakers. Audio recordings in these films are severely affected by background noise, and X-ray images are affected by variable intensity, making it difficult to extract the tongue contours. The MOCHA-TIMIT, USC-TIMIT and MNGU0 databases provide simultaneously acquired audio signals and articulatory data using electromagnetic articulography (EMA) for 460 sentences spoken by two speakers, 460 sentences spoken by four speakers, and 1300 sentences spoken by one speaker, respectively. The MNGU0 and USC-TIMIT databases also provide simultaneously acquired audio signals and articulatory data using MRI for one speaker and ten speakers, respectively. The audio signal acquired during the MRI is affected by the scanner noise, making acoustic analysis difficult. As these databases provide data for a small number of speakers, they are not suitable for obtaining the relationship between spectral parameters and place of articulation. The XRMB database comprises simultaneous recordings of the *x*-*y* coordinates of gold pellets' attached on the tongue, lips, and mandibular incisors, tracked using a narrow beam of high-energy X-ray and the audio signal. It has recordings for isolated words, sentences, vowels, and vowel-consonant-vowels spoken by 47 speakers (21 male and 26 female speakers). As this database has recordings from several speakers, it is more suitable than others for obtaining the relationship between spectral parameters and place of articulation. Therefore, the fricative data in the XRMB database are used in the current study.

The voiced fricatives (/v/, /z/, /ʒ/) and the unvoiced fricatives (/f/, /s/, /ʃ/) extracted from the VCV utterances, words, and sentences of all the speakers in the XRMB database were used in the investigation. The database has recordings for 118 tasks, including a sequence of words, sentences, and VCV utterances, with parallel recordings of audio signal and articulatory data for each task by a speaker. Not all the speakers recorded all the tasks, and the articulatory data for some tasks are missing due to the pellets having fallen off the articulators during recording. The articulatory data are available at 145 frames/s. The audio signal in the database has a sampling frequency of 21.739 kHz, and it was down-sampled to 16 kHz. The fricative segment boundaries were extracted from the audio recordings using Penn Phonetics Lab Forced Aligner (Yuan and Liberman 2008), and these were manually checked and adjusted. The fricatives /θ/ and /ð/ were not included in the current study, as many of these fricative utterances in the database had stop-like characteristics, making it difficult to obtain segment boundaries for them. A total of 10,112 fricative segments were extracted from the

recordings of isolated words, sentences, and vowel-consonant-vowels in 24 tasks (TP003, TP010, TP011, TP016, TP017, TP020, TP024, TP029, TP039, TP043, TP050, TP051, TP053, TP056, TP060, TP067, TP078, TP079, TP080, TP081, TP088, TP089, TP097, TP098). The extracted segments from the XRMB database included at least 1500 utterances for each fricative class, except that the palatal fricative /ʒ/ had very few utterances in the database. The places of articulation during frication were estimated from the articulograms in the XRMB database using the automated technique described in the previous section. The length of the axial curve from the lips to the smallest oral cavity opening is used as the place of articulation estimated from the articulatory data, and it is referred to as 'PoA-art' in further description.

### *3.4.1 Computation of spectral parameters*

The magnitude spectra of the speech signals with the sampling frequency of 16 kHz were calculated using a 30-ms Hanning window and 512-point FFT. The average magnitude spectrum for each fricative utterance was calculated using the magnitude spectra with a 5-ms window shift over the central one-third segment. These spectra were used to calculate the spectral parameters. The relationship between the place of articulation of the fricative utterances measured from the articulatory data and the spectral features was analyzed using some of the earlier reported parameters and some additional proposed parameters, as described here.

*Earlier reported parameters*

Based on the earlier studies on the acoustic parameters related to the place of articulation of fricatives, spectral moments, spectral peak frequency, normalized amplitude, and MFCC were selected for the investigation.

*Spectral Moments (SM1, SM2, SM3, SM4):* The first four spectral moments (Forrest *et al.* 1988) are computed from the average magnitude spectra. These moments characterize the spectral shape variations corresponding to changes in place of articulation. The average magnitude spectrum $S(k)$, with $k$ as frequency index, is used to calculate the normalized squared magnitude spectrum $P(k)$ as

$$P(k) = S^2(k) \Big/ \sum_{k=0}^{L/2} S^2(k) \tag{3.1}$$

where $L$ is the FFT size. The first moment (SM1) spectral centroid indicating the average energy concentration location, is calculated as

$$\mathrm{SM1} = \sum_{k=0}^{L/2} kP(k) \tag{3.2}$$

The second spectral moment (SM2), spectral variance indicating the spectral bandwidth around the spectral centroid, is calculated as

$$\mathrm{SM2} = \left[ \sum_{k=0}^{L/2} (k - \mathrm{SM1})^2 P(k) \right]^{1/2} \tag{3.3}$$

The third moment (SM3), skewness indicating the asymmetry, is calculated with its value normalized with reference to SM2 as the following:

$$\mathrm{SM3} = \left[ \sum_{k=0}^{L/2} (k - \mathrm{SM1})^3 P(k) \right] \Big/ \mathrm{SM2}^3 \tag{3.4}$$

The fourth moment, kurtosis indicating the peakiness of the spectrum, is calculated with its value normalized with reference to SM2 as the following:

$$\mathrm{SM4} = \left[ \sum_{k=0}^{L/2} (k - \mathrm{SM1})^4 P(k) \right] \Big/ \mathrm{SM2}^4 \tag{3.5}$$

*Spectral Peak Frequency (SPF):* It is calculated as the frequency with the highest value in the averaged magnitude spectrum $S(k)$. This parameter is an indicator of spectral resonance in the alveolar and palatal fricatives.

*Normalized Amplitude (n-Amp):* The normalized amplitude (n-Amp) is calculated as the frication noise RMS in dB with reference to the adjacent vowel RMS over three consecutive pitch periods at the point of maximum vowel amplitude as described by Jongman *et al.* (2000). This parameter characterizes the /f, v/ fricatives with a lower normalized amplitude.

*Mel-Frequency Cepstral Coefficients (MFCC):* These coefficients represent the smooth spectral shape of a segment and have been widely used as the input parameters for acoustic-to-articulatory mapping (Hiroya and Honda 2004, Toda *et al.* 2008, Ghosh and Narayanan 2010, Mitra *et al.* 2010, Ji *et al.* 2014b). They are calculated by applying the discrete cosine transform on the log of the smoothed spectrum obtained by applying a mel filterbank on the average magnitude spectrum. The filterbank consists of mel-band triangular filters, with 20 bands over 0−8 kHz. For avoiding negative overflow during log calculations., a floor is applied to the mel filterbank outputs. This floor value is selected as 30 dB below the maximum of the band outputs. The zeroth coefficient of the discrete cosine transform, represents the signal level. The following twelve coefficients are used as the MFCC set to represent the spectral shape.

*Proposed parameters*

Based on a visual examination of average magnitude spectra corresponding to different values of the place of articulation for several fricative utterances in the XRMB database, a set of

additional parameters were investigated for their relationship with the place of articulation: maximum-sum segment centroid, spectral slope, and spectral energy.

*Maximum-Sum Segment Centroid (MSSC)*: The alveolar fricatives have broadband energy above 3.5 kHz, the palatal fricatives have relatively narrowband energy above 2 kHz, and the labiodental fricatives have broadband low-frequency energy with a downward slope above 1.5 kHz (Heinz and Stevens 1961, Stevens 2000). The spectral centroid (SM1) has been reported to be a useful indicator of these spectral characteristics (Jongman *et al.* 2000). However, it gets affected by the outlying spectral samples, resulting in reduced discrimination. For decreasing the effect of outlying spectral samples, dominant spectral centroid (DSC) is calculated as the centroid of the spectral samples above the 80-percentile of the distribution (Nataraj *et al.* 2017). We form a set of frequency indices $\mathrm{D} = \{k, \text{ such that } S(k) > P_{th}\}$ with $P_{th}$ as the 80-percentile value of $S(k)$. The DSC is calculated as

$$\mathrm{DSC} = \sum_{k \in \mathrm{D}} kS(k) \Big/ \sum_{k \in \mathrm{D}} S(k) \tag{3.6}$$

The DSC values were found to be better than the SM1 values in discriminating between the labial and alveolar fricatives. However, they showed inconsistent estimates when the dominant spectral samples were far apart, particularly in relatively flat spectra. For resolving this problem, a spectral segment as a contiguous subset of the spectral samples having the maximum sum of the median subtracted magnitudes is located, and its centroid is calculated and termed as the maximum-sum segment centroid (MSSC). This parameter is not affected by spurious spectral peaks outside the maximum-sum segment. To locate the maximum-sum segment, Kadane's one-pass optimal search algorithm (Bentley 1984, Takaoka 2002) for finding the maximum-sum subarray of the median-subtracted spectrum as the input array is used. The input array is obtained as

$$S_{MS}(k) = S(k) - \mathrm{median}(S(0), S(1), \ \ldots, S(L/2)) \tag{3.7}$$

The algorithm calculates a local sum $S_{LS}(k)$ and a global sum $S_{GS}(k)$. The global sum is updated if the local sum exceeds the global sum. The operations of the algorithm are expressed as the following difference equations

$$S_{LS}(k) = \max(S_{MS}(k),\ S_{LS}(k-1) + S_{MS}(k)\ ), \quad 1 \le k \le L/2 \tag{3.8}$$

$$S_{GS}(k) = \max(S_{GS}(k-1),\ S_{LS}(k)) \tag{3.9}$$

For $k$ = 0, the local sum $S_{LS}(k)$ and the global sum $S_{GS}(k)$ are initialized with the first spectral sample $S_{MS}(0)$. The beginning of the maximum-sum subarray, $i_{BG}(k)$, the end of the

maximum-sum subarray, $i_{EG}(k)$, and the beginning of the local subarray, $i_{BL}(k)$, are updated as

$$i_{BL}(k) = \begin{cases} k, & S_{MS}(k) > S_{LS}(k-1) + S_{MS}(k) \\ i_{BL}(k-1), & \text{otherwise} \end{cases} \tag{3.10}$$

$$i_{BG}(k) = \begin{cases} i_{BL}(k), & S_{LS}(k) > S_{GS}(k-1) \\ i_{BG}(k-1), & \text{otherwise} \end{cases} \tag{3.11}$$

$$i_{EG}(k) = \begin{cases} k, & S_{LS}(k) > S_{GS}(k-1) \\ i_{EG}(k-1), & \text{otherwise} \end{cases} \tag{3.12}$$

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 0 for $k = 0$. 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 subarray. MSSC is calculated using the contiguous spectral samples of the average magnitude spectrum in the maximum-sum segment as

$$\text{MSSC} = \sum_{k=i_{BG}(L/2)}^{i_{EG}(L/2)} kS(k) \Bigg/ \sum_{k=i_{BG}(L/2)}^{i_{EG}(L/2)} S(k) \tag{3.13}$$

Figure 3.6 shows an example of the MSSC calculation for the fricative /s/, with the subset of spectral samples corresponding to the maximum sum, MSSC, DSC, and SM1. It may be noted that the maximum-sum segment locates all the samples of dominant high-frequency energy, and MSSC marks the center of high-frequency energy, while DSC and SM1 deviate from this location.

*Normalized Sum of Absolute Spectral Slopes (NSASS):* The labiodental fricatives have broadband low-frequency energy with a slow decay above 1.5 kHz, and the palatal fricatives have narrowband energy above 2 kHz with steep rising and falling spectral segments. These spectral characteristics can be used to separate the labiodental and palatal fricatives by calculating the normalized sum of absolute spectral slopes (NSASS) of the smoothed spectrum. The average magnitude spectrum $S(k)$ is smoothed using a two-step median-mean filter (Rabiner *et al.* 1975) to obtain a smoothed spectrum $S_{CB}(k)$, suppressing the spurious variations affecting the slope calculation without disturbing the peaks and valleys. 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 frequency index corresponds to the critical bandwidth $B(k)$ centered at the frequency index $k$ and calculated using the relation given by Zwicker (1961) as

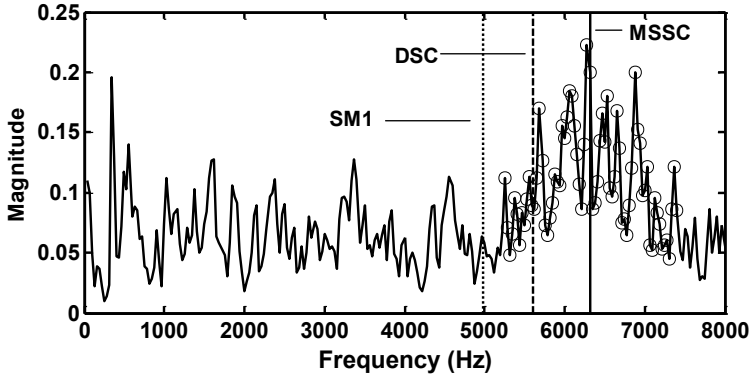



**Figure 3.6** Example of MSSC calculation from the average spectrum of an /s/ utterance, with the maximum-sum subset samples marked as circles and MSSC (continuous), DSC (dashed), SM1 (dotted) marked as vertical lines.

$$B(k) = 25 + 75(1 + 1.4(f(k))^2)^{0.69} \tag{3.14}$$

where $f(k) = kf_s / L$ and $f_s$ is the sampling frequency in kHz. The NSASS is calculated as the sum of the absolute values of the first difference of $S_{CB}(k)$ normalized using the RMS active speech level, $\text{RMS}_{\text{ASL}}$, of the recording. It is given as

$$\text{NSASS} = \sum_{k=1}^{L/2} |S_{CB}(k) - S_{CB}(k-1)| / \text{RMS}_{\text{ASL}} \tag{3.15}$$

The $\text{RMS}_{\text{ASL}}$ calculated using the audio recording for the database task containing the fricative utterance is used for normalization. It is used instead of the sum of the samples in the average spectrum to avoid the erroneously large slope values for labiodental fricatives due to their low signal level. Further, active speech level calculated in accordance with method B of International Telecommunications Union (1993), using the code provided by Loizou (2017), avoids the effect of silences in the audio recordings.

*Spectral Energy Parameters (NHBE, NVLBE, PAE, MLBE):* Normalized energy in different spectral bands can be used to characterize the fricative spectra. Mel filterbank outputs $M(l)$ with $1 \le l \le 20$, calculated as the outputs of the 20 triangular filters applied on the average magnitude spectrum $S(k)$ normalized using the sum of all the values of $S(k)$, are used to calculate the spectral energy parameters. Based on a visual examination of the mel filterbank outputs corresponding to different values of the place of articulation for several fricative utterances, four spectral energy parameters are calculated to characterize the spectral shape variations.

The alveolar fricatives have high-frequency energy concentration. This spectral characteristic can be used to discriminate between the alveolar and labiodental fricatives by calculating the normalized average energy in the high-frequency band (NHBE) in dB as

$$\mathrm{NHBE} = 20\log_{10}\left[\sum_{l=j_{HB}}^{j_{HE}} \frac{M(l)}{j_{HE} - j_{HB} + 1}\right] \tag{3.16}$$

The band beginning and end, $j_{HB}$ and $j_{HE}$, are set to 18 and 20, respectively, which correspond to the bands centered at 5.36 kHz and 7.01 kHz, respectively.

The labiodental fricatives have low-frequency energy concentration. This spectral characteristic can be used to discriminate between the labiodental and other fricatives by calculating the normalized average energy in the very-low-frequency band (NVLBE) in dB as

$$\mathrm{NVLBE} = 20\log_{10}\left[\sum_{l=j_{VB}}^{j_{VE}} \frac{M(l)}{j_{VE} - j_{VB} + 1}\right] \tag{3.17}$$

The band beginning and end, $j_{VB}$ and $j_{VE}$, are set to 5 and 8, respectively, which correspond to the bands centered at 0.57 kHz and 1.12 kHz, respectively.

The energy at spectral peaks of alveolar and palatal fricatives is higher than the average energy in low-frequency bands. The labiodental fricatives have broadband low-frequency energy without distinct spectral peaks. This spectral characteristic can be used to discriminate the alveolar and palatal fricatives from the labiodental fricatives by calculating the peak energy expressed in dB with respect to the average energy (PAE) as

$$\mathrm{PAE} = 20\log_{10}\left[\max_{l\in\{j_{PB},\ldots,j_{PE}\}} M(l) \Big/ \sum_{l=j_{AB}}^{j_{AE}} \frac{M(l)}{j_{AE} - j_{AB} + 1}\right] \tag{3.18}$$

The beginning and end of the bands for finding the peak energy, $j_{PB}$ and $j_{PE}$, are set to 11 and 20, corresponding to the bands centered at 1.92 kHz and 7.01 kHz, respectively. The beginning and end of the bands for finding average energy, $j_{AB}$ and $j_{AE}$, are set to 8 and 10, corresponding to the bands centered at 1.12 kHz and 1.62 kHz, respectively. The average energy values are larger for the labiodental fricatives compared to the palatal and alveolar fricatives due to their energy concentration in low-frequency regions. The peak energy values of labiodental fricatives are lower than those of the palatals and alveolars, and thus PAE may be useful in separating the labiodental fricatives from the palatal and alveolar fricatives.

Some of the alveolar and palatal fricatives have low values of MSSC and NSASS resulting in overlap with the labiodental fricatives. However, the alveolar and palatal fricatives have distinct spectral peaks, and the labiodental fricatives have broadband low-frequency energy with a slow decay above 1.5 kHz. This spectral characteristic can be used to discriminate

between the alveolars and palatal fricatives from the labiodental fricatives by calculating the average mid-frequency band energy (MLBE) expressed in dB with respect to the average low-frequency-band energy as

$$\text{MLBE} = 20\log_{10}\left[\sum_{l=j_{MB}}^{j_{ME}} \frac{M(l)}{j_{ME}-j_{MB}+1} \Big/ \sum_{l=j_{LB}}^{j_{LE}} \frac{M(l)}{j_{LE}-j_{LB}+1})\right] \tag{3.19}$$

The mid-frequency band beginning and end, $j_{MB}$ and $j_{ME}$, are set to 14 and 17, respectively, which correspond to the bands centered at 3.05 kHz and 4.68 kHz, respectively. The low-frequency band beginning and end, $j_{LB}$ and $j_{LE}$, are set to 11 and 15, which correspond to the bands centered at 1.92 kHz and 3.53 kHz, respectively.

For avoiding the negative overflow during the log calculations, a floor is applied to the mel filterbank outputs as in the MFCC calculation.

### *3.4.2 Relationship of place of articulation with the spectral parameters*

For each fricative utterance, the place of articulation was obtained from its articulogram, as described in Section 3.3. Figure 3.7 shows the place of articulation values for fricative utterances of the male and female speakers. The values of place of articulation for the labiodentals are lower than the alveolars and the palatals for both speaker groups. The lowest values for the alveolars are lower than the palatals of both speaker groups. There is an overlap between the values for the alveolars and the palatals. The lowest values for the alveolars and the palatals of the female speakers are lower than the male speakers, and the highest values for the palatals of the female speakers are lower than the male speakers. This difference could be attributed to the smaller vocal tract length of female speakers than male speakers. The mean value for all utterances of each fricative by the same speaker was calculated, and the means and standard deviations of these values were used to examine across-the-speakers distribution for each fricative. These values along with the corresponding number of utterances are given in Table 3.2. The mean values for the labiodentals are approximately zero. The mean values for the alveolars and the palatals of the female speakers are lower than the male speakers. The standard deviations indicate an overlap between the values for the male speakers' alveolars and the female speakers' palatals.

For examining the relation between the place of articulation and spectral parameters, the place of articulation (PoA-art) values were quantized in 1-mm steps, and the mean and standard deviation of the spectral parameters corresponding to each of the quantization steps were calculated. Figure 3.8 shows plots of the mean and standard deviation of the spectral moments (SM1, SM2, SM3, SM4) versus the place of articulation and normalized histograms

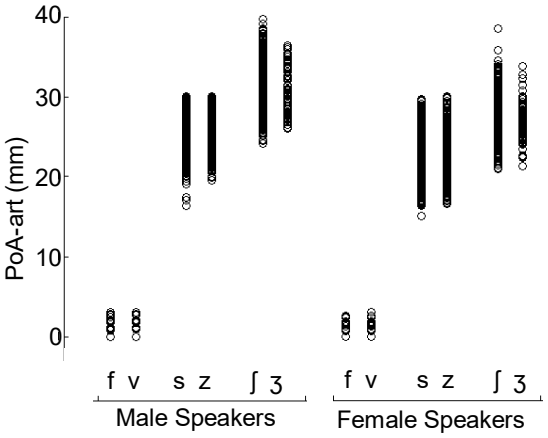



**Figure 3.7** Place of articulation values obtained from the articulogram (PoA-art) for fricative utterances for male and female speakers.

**Table 3.2**: Mean and standard deviation (S.D.) of place of articulation obtained from the articulograms (PoA-art) for fricative utterances using the automated graphical technique (N = number of utterances in the dataset).

| Fricative | POA-art (mm) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Male Speakers | | | Female Speakers | | | All Speakers | | |
| | N | Mean | S.D. | N | Mean | S.D. | N | Mean | S.D. |
| f | 705 | 0.19 | 0.36 | 806 | 0.14 | 0.26 | 1511 | 0.16 | 0.30 |
| v | 676 | 0.28 | 0.28 | 859 | 0.43 | 0.43 | 1535 | 0.37 | 0.38 |
| s | 1339 | 25.80 | 1.74 | 2030 | 22.51 | 1.68 | 3369 | 23.94 | 2.36 |
| z | 733 | 26.23 | 1.64 | 1177 | 23.06 | 1.58 | 1910 | 24.44 | 2.24 |
| ʃ | 680 | 31.32 | 2.47 | 952 | 26.99 | 1.96 | 1632 | 28.87 | 3.07 |
| ʒ | 67 | 31.06 | 2.51 | 88 | 26.96 | 2.28 | 155 | 28.74 | 3.13 |

of all fricatives. In these plots, the PoA-art values can be grouped into two regions, a region below 5 mm corresponding to the labiodentals and a region above 15 mm corresponding to alveolars and palatals. The SM1 values decrease as the PoA-art values increase above 15 mm, corresponding to a change in the frequency of energy concentration in the spectrum with change in the place of articulation. The labiodentals have energy concentrated in low frequency with no distinct spectral peaks, and the SM1 values are approximately 2 kHz for the PoA-art values less than 5 mm. The histograms show that SM1 values for the alveolars range 3−6 kHz with a peak around 5 kHz and these values are higher than those for the labiodental and the palatals. There is a significant overlap between the SM1 values for labiodentals and palatals. The SM2 values decrease as the PoA-art values increase above 15 mm and are approximately 1.5 kHz for the PoA-art values less than 5 mm. The histograms show SM2 values for the labiodentals and the alveolars are higher than that of the palatals,

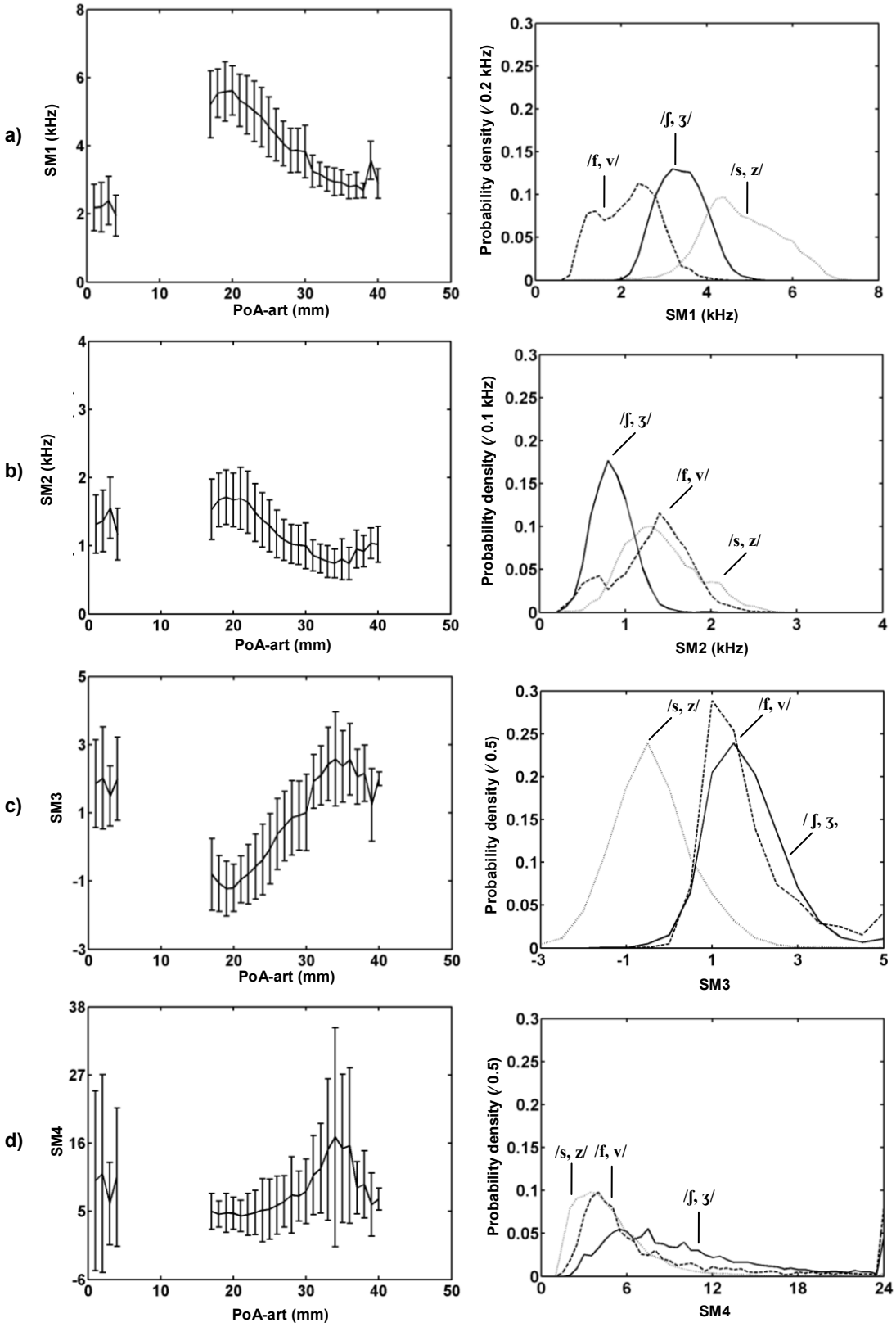



**Figure 3.8** Mean and standard deviation of spectral moments as a function of place of articulation (left side) and normalized histograms (right side) for fricatives /f, s, ʃ, v, z, ʒ/: (a) SM1, (b) SM2, (c) SM3, and (d) SM4.

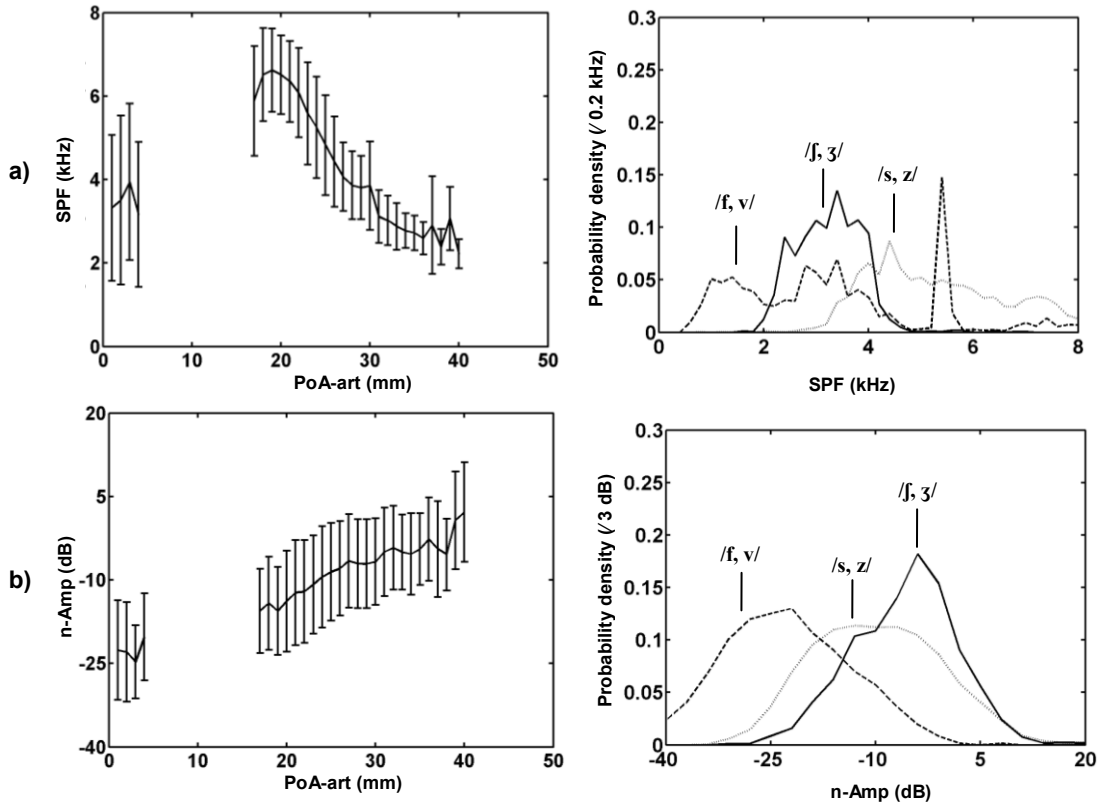



**Figure 3.9** Mean and standard deviation of spectral peak frequency (SPF) and normalized amplitude (n-Amp) as a function of place of articulation (left side) and normalized histograms (right side) for fricatives /f, s, ʃ, v, z, ʒ/: (a) SPF and (b) n-Amp.

with a significant overlap between the labiodentals and the alveolars. The SM3 values increase as the PoA-art values increase above 15 mm and are approximately 2 for the PoA-art less than 5 mm. The SM3 values for the labiodentals and the palatals are higher than the alveolars, with a significant overlap between the labiodentals and the palatals. The SM4 values increase as the PoA-art values increase above 15 mm and show significant variability for PoA-art less than 5 mm. The histograms show a significant overlap between the SM4 values for all fricatives.

Figure 3.9 shows plots of the mean and standard deviation of the spectral peak frequency (SPF) and normalized amplitude (n-Amp) versus the place of articulation and normalized histograms of all fricatives. The SPF values decrease as the PoA-art values increase above 15 mm. As the labiodentals do not have distinct spectral peaks, the SPF values show a large variation. The histograms show that the SPF values for the alveolars are higher than the palatals, and the SPF values for the labiodentals are distributed across the frequency range as they do not have distinct spectral peaks. The mean values of the n-Amp for the palatals are higher than those of the alveolars. The histograms show a significant overlap between the

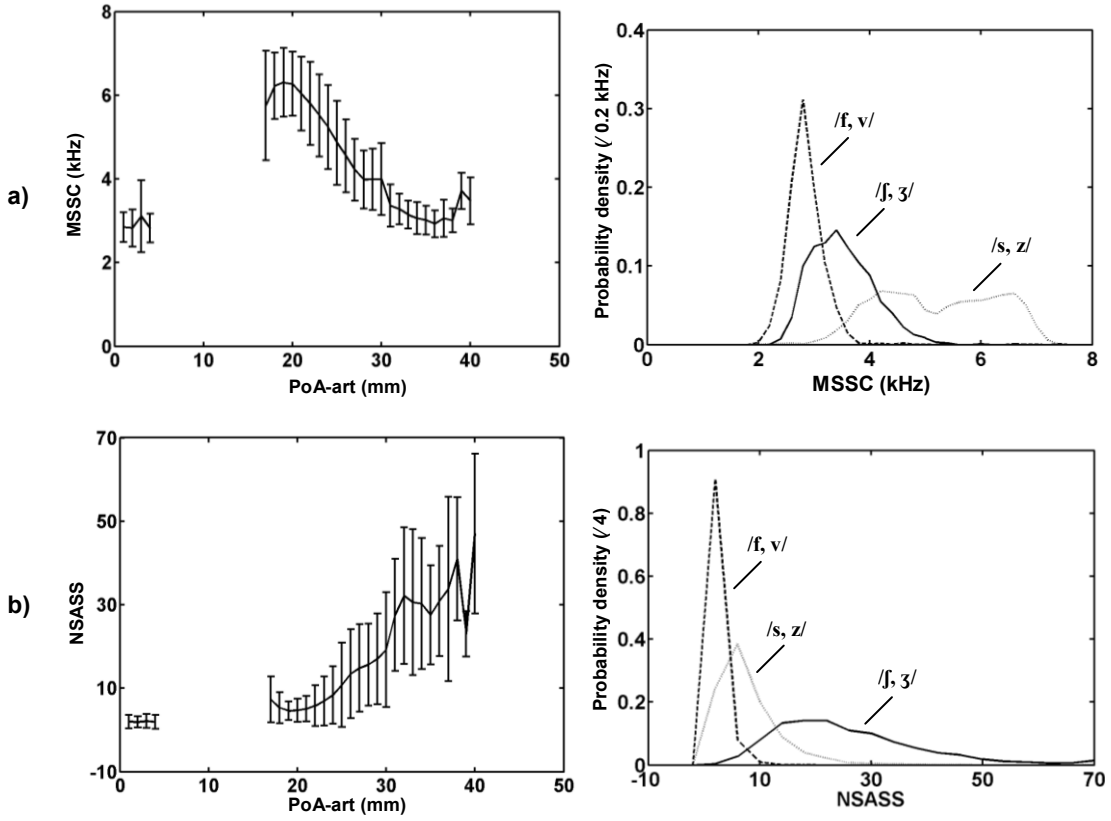



**Figure 3.10** Mean and standard deviation of MSSC and NSASS as a function of place of articulation (left side) and normalized histograms (right side) for fricatives /f, s, ʃ, v, z, ʒ/: (a) MSSC and (b) NSASS.

n-Amp values for all fricatives. From these results, it is observed that the spectral parameters SM1, SM2, SM3, and SPF are related to the place of articulation, but relationships are not distinct. The spectral parameters SM4 and n-Amp values have a significant spread across fricatives.

Figure 3.10 shows plots of the mean and standard deviation of the MSSC and NSASS versus the place of articulation and normalized histograms of all fricatives. The MSSC values decrease as the PoA-art values increase above 15 mm and is approximately 3 kHz for the values PoA-art less than 5 mm. The MSSC histogram for the alveolars is flat in the 3.5–7 kHz band, capturing the variability in the location of energy concentration in the spectrum. The labiodentals have low-frequency energy with less variability across speakers, resulting in the MSSC values localized to a low frequency narrow band. The MSSC histograms for the labiodentals and the palatals have an overlap. The NSASS values increase as the PoA-art values increase above 15 mm and are approximately 2 for PoA-art less than 5 mm. The NSASS histograms show that the labiodentals have the lowest values, while the palatals and

the alveolars have higher values. The labiodentals and the palatals are well separated, and there is an overlap between the labiodentals and the alveolars.

Figure 3.11 shows plots of the mean and standard deviation of the spectral energy parameters (NHBE, NVLBE, PAE, and MLBE) versus the place of articulation and normalized histograms of all fricatives. The NHBE values decrease as the PoA-art values increase above 15 mm and are approximately −35 dB for PoA-art less than 5 mm. The NHBE values are high for the alveolars as they have high-frequency energy. The labiodental and the palatals have lower values as they have energy concentrated in low-frequency bands. The NVLBE values decrease as the PoA-art values increase above 15 mm and are highest for the PoA-art values less than 5 mm. The NVLBE values are high for the labiodentals, as they have low-frequency energy. The values are low for the alveolars and the palatals due to high-frequency energy concentration. The PAE values for the alveolars and the palatals are significantly higher than the labiodentals as they have the high-frequency energy concentratation and have low energy in the band where the average energy is calculated. The MLBE values are high for the PoA-art values in the range 20−30 mm, corresponding to the fricatives with peaks in the mid-frequency bands. The MLBE values are low for the labiodentals, and high for the alveolars and the palatals, as they have spectral peaks in the mid-frequency band. Some of the palatals also had low values due to low-frequency spectral peaks. Thus the spectral energy parameters may help place estimation by providing additional spectral shape information.

The histogram plots show a relationship between the spectral parameters and the place of articulation of fricatives, but with significant overlap. Therefore, an investigation needs to be carried out to estimate the place of articulation from a combination of these parameters.

## 3.5 Estimation of the place of articulation using artificial neural network

Several models based on machine learning have been proposed to estimate the articulatory parameters from the acoustic parameters, including codebook based model, HMM, GMM, and ANN (Hiroya and Honda 2004, Toda *et al.* 2008, Richmond 2006, Uria *et al.* 2012, Liu *et al.* 2015). The ANN-based models have been reported to estimate the articulatory parameters with good accuracy (Liu *et al.* 2015). An investigation using feedforward ANN for speaker-independent estimation of the place of articulation from the spectral parameters is presented. The ANN model uses a fully connected network with every neuron of a layer connected to every neuron of the next layer. The input parameters, number of hidden layers,

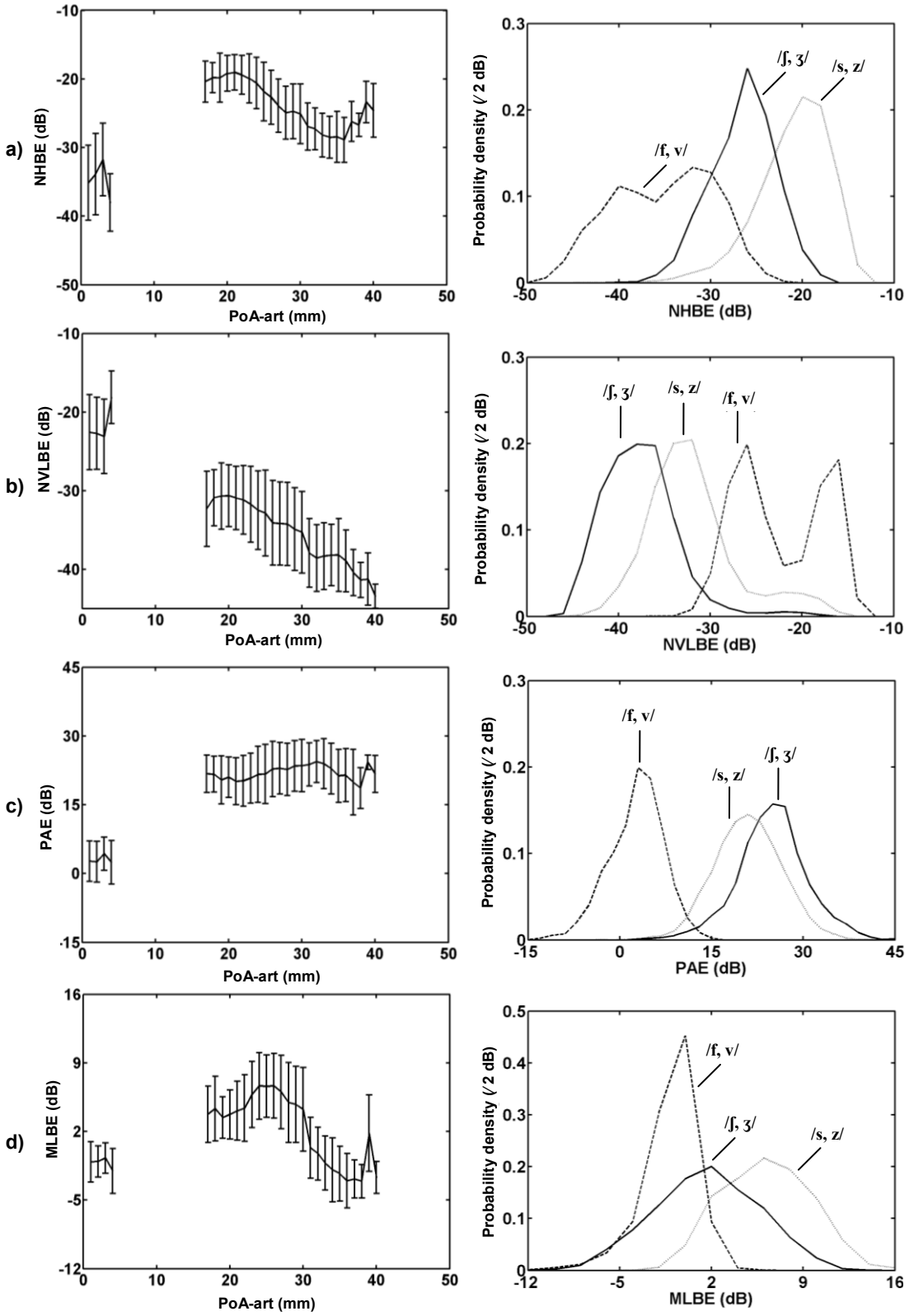



**Figure 3.11** Mean and standard deviation of spectral energy parameters as a function of place of articulation (left side) and normalized histograms (right side) for fricatives /f, s, ʃ, v, z, ʒ/: (a) NHBE, (b) NVLBE, (c) PAE, and (d) MLBE.

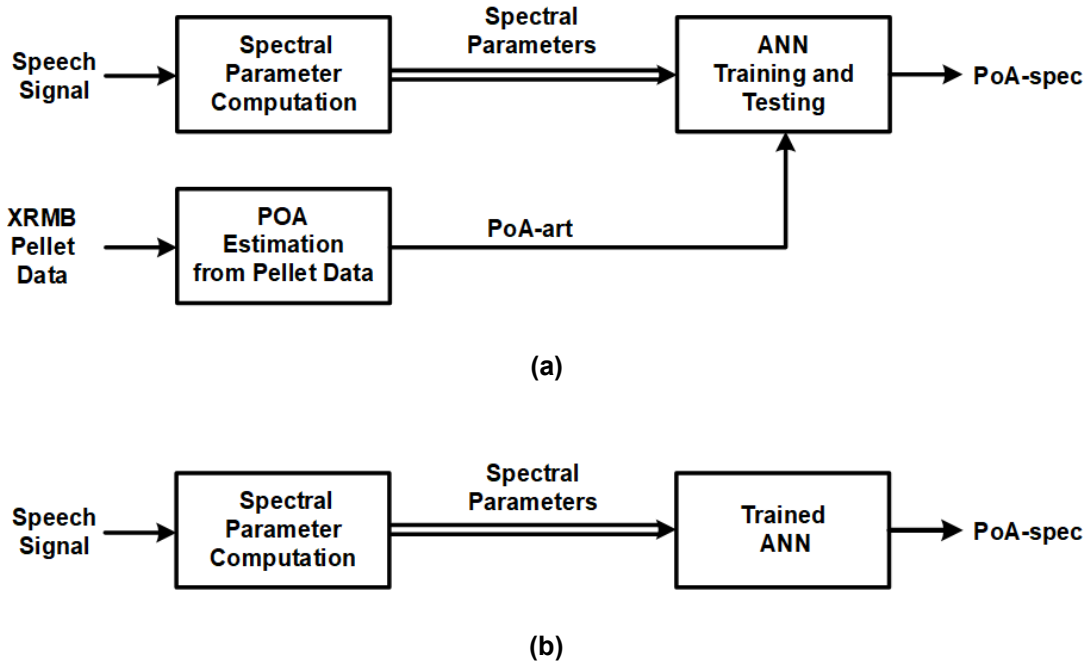



**Figure 3.12** ANN-based estimation of place of articulation: (a) ANN training using place of articulation values obtained from XRMB database, (b) estimation using the trained ANN.

number of neurons in each hidden layer, activation function in the hidden layers, and training algorithm are selected empirically to reduce the estimation error. A small number of neurons or layers may result in under-fitting, as the network might not model the training data. A large number of neurons or hidden layers may result in over-fitting, as the network tends to remember the training data but is unable to generalize to unseen test data. A suitable criterion needs to be selected for stopping the network training to avoid the over-fitting problem. The network needs to be cross-validated to estimate its performance by training on different subsets of the dataset and testing on the remaining data.

A block diagram of the ANN-based estimation of the place of articulation using speech signal and articulatory data of the fricatives in the XRMB database is shown in Figure 3.12. The ANN training and testing uses spectral parameters calculated from the speech signal as the input feature vector and the place of articulation obtained from the articulogram PoA-art as the reference to output the estimated place of articulation, as shown in Figure 3.12(a). The place of articulation estimated from the spectral parameters is referred to as 'PoA-spec' in further description. The trained ANN is subsequently used for estimating the PoA-spec from the spectral parameters calculated from the speech signal as the input feature vector as shown in Figure 3.12(b). A feedforward fully connected network with multiple hidden layers was implemented using MATLAB Neural Network Toolbox Release 2014a (MathWorks, Inc.,

Natick, Mass., USA). The hyperbolic tangent activation function was used in all the neurons in the hidden layers, and the Levenberg-Marquardt method was used for updating the weights and bias values of the network. The input and reference parameters were normalized to have zero mean and unity variance to speed up the training by making the learning rate uniform across the network weights.

The evaluation was carried out using the frication segments corresponding to the fricatives /f, v, s, z, ʃ, ʒ/ from 47 speakers in the XRMB database. As described earlier in Section 3.4, a total of 10,112 utterances were used. These utterances were divided, maintaining a balance in terms of gender and speakers into five utterance sets: (i) set-1 with four male and five female speakers and 1989 utterances, (ii) set-2 with four male and five female speakers and 1990 utterances, (iii) set-3 with four male and five female speakers and 1995 utterances, (iv) set-4 with four male and six female speakers and 2047 utterances, and (v) set-5 with five male and five female speakers and 2091 utterances. The data for each utterance, comprising spectral parameters and the corresponding PoA-art, is referred to as a data sample. The evaluation was performed using five-fold cross-validation, with the data from the four sets used for training, and the data from the fifth set used for validation in each step. This method provides a speaker-independent validation, as the data from a speaker is used either for training or for validation. It helps in examining the data bias with respect to the data selection from different speakers for training. The network weights were initialized using random numbers sampled from a uniform distribution between −1 and +1. Obtaining the estimated values for all the data samples in the training set constituted a training epoch. After each epoch, network weights were updated using the batch gradient descent algorithm to minimize the mean square error in the estimated values for all the data samples in the training set. A decrease in the mean square error over the training set and an increase or no change in the error over the validation set was considered as validation failure or over-fitting. The training was stopped after six successive validation failures to avoid over-fitting. If the training did not stop due to the validation failure, training was continued for a maximum of 1000 epochs.

### *3.5.1 Investigations*

The investigation for estimating the place of articulation was carried out to evaluate the effects of (i) input parameter set, number of hidden layers, and number of neurons, (ii) size of training data, and (iii) pellet locations as output parameters, as described in the following subsections.

*Input parameter set*, *number of hidden layers, and number of neurons:* The investigation used the following six spectral parameter sets (SPS-n):

(i) spectral parameter set SPS-1 with four spectral moments (SM1, SM2, SM3, SM4),

(ii) spectral parameter set SPS-2 with three spectral moments (SM1, SM2, SM3),

(iii) spectral parameter set SPS-3 with three spectral moments (SM1, SM2, SM3) and spectral peak frequency SPF,

(iv) spectral parameter set SPS-4 with four spectral moments (SM1, SM2, SM3, SM4), spectral peak frequency SPF, and normalized amplitude n-Amp,

(v) spectral parameter set SPS-5 with MFCC (12 coefficients), and

(vi) spectral parameter set SPS-6 with the proposed spectral parameters MSSC, NSASS, NHBE, NVLBE, PAE, and MLBE.

For examining the effect of the number of hidden layers, networks with one, two, and three hidden layers were used. The effect of the number of neurons was examined using networks differing in the number of neurons in the hidden layers.

*Training data size:* This investigation was carried out using one, two, three, and four utterance sets, corresponding approximately to 20%, 40%, 60%, and 80% of the dataset, for training and the fifth set, corresponding approximately to 20% of the dataset, for validation. The investigation used spectral parameter sets SPS-5 (MFCCs) and SPS-6 (proposed parameter set) with the optimal number of hidden layers and neurons as obtained in the earlier investigation.

*Pellet locations as the ANN output parameters:* Some studies have used acoustic-to-articulatory mapping to estimate the *x*-*y* locations of the markers on the articulators (Toda *et al.* 2008, Richmond 2006, Uria *et al.* 2012, Liu *et al.* 2015). An investigation was carried out with the *x*-*y* locations of the pellets on the lips (upper lip and lower lip), the incisor, and the tongue (four pellets) as the ANN output parameters. Networks differing in terms of the number of hidden layers and neurons were used to obtain the optimal architecture. The place of articulation was estimated from the estimated pellet locations using the graphical processing technique as described earlier in Section 3.3.

## 3.6 Results

The error in estimating the place of articulation for each data sample was calculated as the difference between the PoA-spec and PoA-art values. The mean of errors and the standard deviation of errors were calculated as measures of bias and random errors in the estimation, respectively. The RMS of errors was calculated as a composite error measure. Further, the

correlation coefficient between the PoA-spec and PoA-art values was calculated as a measure of association of the estimated values with the reference values.

### *3.6.1 Effects of input parameter set, number of hidden layers, and number of neurons*

The effects of the six input parameter sets (SPS-1, SPS-2, … SPS-6) using networks with different numbers of hidden layers and neurons were examined by calculating the mean of RMS errors across the cross-validation folds. The lowest errors were usually obtained for estimation using spectral parameter sets SPS-5 (the MFCC set) and SPS-6 (the proposed parameter set comprising MSSC, NSASS, NHBE, NVLBE, PAE, and MLBE). For these two parameter sets, plots of the mean of RMS errors versus the number of neurons in the hidden layers for networks with one, two, and three hidden layers are shown in Figure 3.13.

For networks with one hidden layer, the number of neurons was varied from 1 to 30. Figure 3.13(a) shows plots of the mean of RMS errors versus the number of neurons in the hidden layer for the spectral parameter sets SPS-5 and SPS-6. The errors decrease with an increase in the number of neurons for both the spectral parameter sets, but no significant decrease occurs after a certain number of neurons in each case. The mean of RMS errors for other spectral parameter sets showed similar variations. The networks having 24, 29, 30, 20, 26, and 12 neurons in the hidden layer were found as optimal for the spectral parameter sets SPS-1, SPS-2, SPS-3, SPS-4, SPS-5, and SPS-6, respectively. The errors were highest for the spectral parameter set SPS-2 (SM1, SM2, SM3) and SPS-3 (SM1, SM2, SM3, SPF), and they were lowest for SPS-6 (proposed parameter set). Further, SPS-6 required a much lower number of neurons than the other spectral parameter sets.

The effect of two hidden layers was examined using networks with the number of neurons in the first hidden layer same as the optimal values for the network with one hidden layer. The number of neurons in the second hidden layer was varied from 1 to 20. Figure 3.13(b) shows plots of the mean of RMS errors versus the number of neurons in the second hidden layer for the spectral parameter sets SPS-5 and SPS-6. These errors are lower than the corresponding errors for networks with one hidden layer. The errors decrease with an increase in the number of neurons, but the decreases are lower than those in the case of networks with one hidden layer. Similar variations in mean RMS errors were observed for other spectral parameter sets. The networks having 3, 4, 4, 5, 5, and 2 neurons in the second hidden layer were found as optimal for the spectral parameter sets SPS-1, SPS-2, SPS-3, SPS-4, SPS-5, and SPS-6, respectively. The spectral parameter set SPS-6 (proposed parameter set) required the least number of neurons and had the lowest error.

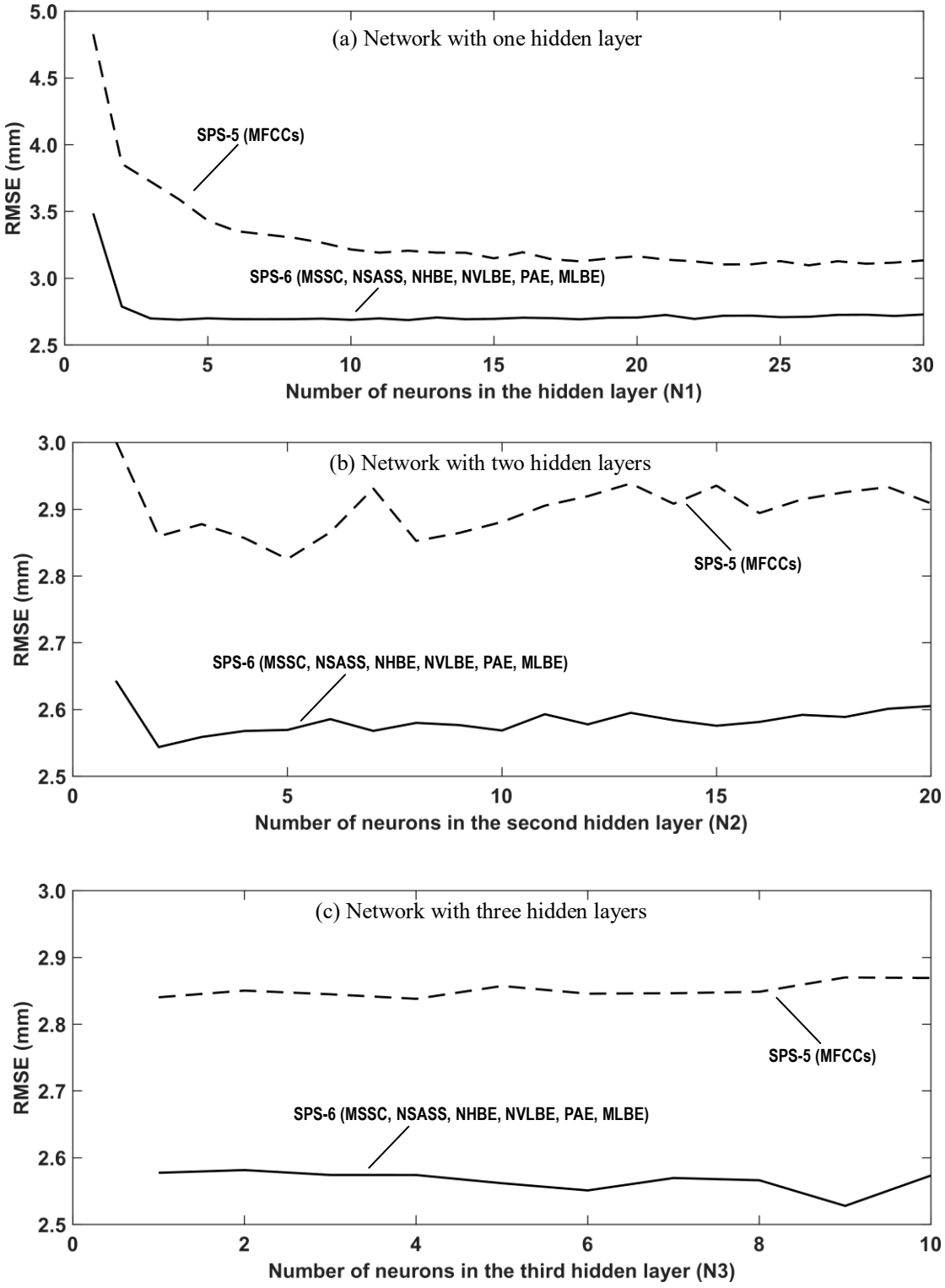



**Figure 3.13** Effect of number of hidden layers and number of neurons: mean of RMS errors (RMSE) versus the number of neurons for spectral parameter sets SPS-5 and SPS-6 and ANN with (a) one hidden layer, (b) two hidden layers (N1: 26 for SPS-5, 12 for SPS-6), and (c) three hidden layers (N1: 26 for SPS-5, 12 for SPS-6, N2: 5 for SPS-5, 2 for SPS-6).

**Table 3.3:** Effect of input spectral parameter set: mean of errors, standard deviation (S.D.) of errors, RMS of errors, and correlation coefficient using different spectral parameter sets and optimal networks with two hidden layers (standard deviation in parentheses).

| Spectral Parameter Set (N1, N2: Number of neurons in the first and second hidden layers) | Mean Error | S.D. Error | RMS Error | Corr. Coeff. |
|---|---|---|---|---|
| SPS-1: SM1, SM2, SM3, SM4 (N1 = 24, N2 = 3) | 0.06 (0.42) | 3.53 (0.35) | 3.55 (0.35) | 0.952 (0.008) |
| SPS-2: SM1, SM2, SM3 (N1 = 29, N2 = 4) | −0.02 (0.33) | 3.69 (0.45) | 3.71 (0.44) | 0.946 (0.011) |
| SPS-3: SM1, SM2, SM3, SPF (N1 = 30, N2 = 4) | −0.08 (0.41) | 3.64 (0.43) | 3.65 (0.42) | 0.950 (0.012) |
| SPS-4: SM1, SM2, SM3, SPF, n-Amp (N1 = 20, N2 = 5) | 0.08 (0.42) | 3.20 (0.26) | 3.22 (0.27) | 0.962 (0.004) |
| SPS-5: MFCCs (N1 = 26, N2 = 5) | −0.06 (0.26) | 2.82 (0.12) | 2.83 (0.13) | 0.970 (0.000) |
| SPS-6: MSSC, NSASS, NHBE, NVLBE, PAE, MLBE (N1=12, N2=2) | 0.02 (0.32) | 2.53 (0.13) | 2.54 (0.13) | 0.978 (0.004) |

The effect of three hidden layers was examined using networks with the number of neurons in the first two hidden layers same as the optimal numbers for the networks with two hidden layers. The number of neurons in the third hidden layer was varied from 1 to 10. Figure 3.13(c) shows plots of the mean of RMS errors versus the number of neurons in the third hidden layer for the spectral parameter sets SPS-5 and SPS-6. The errors do not change with an increase in the number of neurons, and the errors are higher than the corresponding errors for two hidden layers. The error variations for other spectral parameter sets were similar. A comparison of the errors for networks with one, two, and three hidden layers shows the networks with two hidden layers to provide the best performance for the dataset used in our study, indicating that two hidden layers are needed for modeling the complexity of the relation between the spectral parameters and the place of articulation. Several studies to estimate the articulatory parameters from the acoustic parameters have employed DNNs (Uria *et al.* 2012, Liu *et al.* 2015, Illa and Ghosh 2018). However, using three or more hidden layers involves determining a larger number of weights, which possibly leads to performance degradation due to limited data availability for fricative utterances in the XRMB database.

The errors and correlation coefficients for different spectral parameter sets using the optimal networks with two hidden layers are given in Table 3.3. The values of mean of errors are small for all the spectral parameter sets. The mean RMS of errors, subsequently referred to as error, is highest at 3.71 mm for SPS-2, the spectral parameter set comprising first three

**Table 3.4**: RMS errors for different fricatives for estimation using different spectral parameter sets and optimal networks with two hidden layers.

| Spectral Parameter Set | RMS errors with respect to PoA-art (mm) | | | | | | |
|---|---|---|---|---|---|---|---|
| | All | f | v | s | z | ʃ | ʒ |
| SPS-1 (SM1, SM2, SM3, SM4) | 3.6 | 3.9 | 2.4 | 2.9 | 4.2 | 3.8 | 6.2 |
| SPS-2 (SM1, SM2, SM3) | 3.7 | 4.4 | 2.4 | 2.9 | 4.4 | 4.1 | 6.3 |
| SPS-3 (SM1, SM2, SM3, SPF) | 3.7 | 4.3 | 2.6 | 2.8 | 4.3 | 4.0 | 5.5 |
| SPS-4 (SM1, SM2, SM3, SPF, n-Amp) | 3.2 | 3.7 | 2.3 | 2.6 | 3.7 | 3.5 | 5.1 |
| SPS-5 (MFCCs) | 2.8 | 3.2 | 2.3 | 2.5 | 2.9 | 3.2 | 4.0 |
| SPS-6 (MSSC, NSASS, NHBE, NVLBE, PAE, MLBE) | 2.5 | 2.5 | 1.8 | 2.5 | 2.6 | 2.9 | 3.8 |

spectral moments. It does not decrease with the inclusion of the fourth spectral moment in SPS-1 and the spectral peak frequency in SPS-3. The spectral parameter set SPS-4, obtained by inclusion of the normalized amplitude, results in a decrease in error to 3.22 mm. The error for SPS-5 (MFCCs) is 2.83 mm and lower than SPS-4. The error is lowest at 2.54 mm for SPS-6 (proposed parameter set). The correlation coefficients show a spread of 0.946−0.978 for the six spectral parameter sets. The performance order of the spectral parameter sets in terms of the correlation coefficient is the same as in terms of the errors. The standard deviation of the performance measures (errors and correlation coefficients) across the five folds are low for SPS-6 (proposed parameter set) and SPS-5 (MFCCs) than the other spectral parameter sets. Thus, these results show SPS-6 to be most effective in estimating the place of articulation.

The RMS errors for estimation using different spectral parameter sets are given in Table 3.4 for different fricatives. The errors using SPS-5 (MFCCs) for all the fricatives are lower than the corresponding errors using SPS-4 (three spectral moments, spectral peak frequency, normalized amplitude). The errors using SPS-6 (the proposed parameter set) are lowest for all the fricatives, indicating this set to be the most effective in estimating the place of articulation.

### *3.6.2 Effect of training data size*

Table 3.5 summarizes the training data size effect using 20%, 40%, 60%, and 80% of the dataset for training. The errors for the optimal ANN with two hidden layers are shown in Table 3.5 for SPS-5 (MFCCs) and SPS-6 (proposed parameter set) as the spectral parameter sets. The errors decrease with an increase in the training data size. The error for SPS-5

**Table 3.5:** Effect of training data size: RMS of errors using optimal ANN (two hidden layers)

| Spectral Parameter Set | RMS of errors (mm) as a function of training data size (standard deviation in parentheses) | | | |
|---|---|---|---|---|
| | 20% | 40% | 60% | 80% |
| SPS-5 (MFCCs) | 3.53 (0.30) | 3.10 (0.14) | 2.96 (0.21) | 2.83 (0.13) |
| SPS-6 (MSSC, NSASS, NHBE, NVLBE, PAE, MLBE) | 2.71 (0.11) | 2.71 (0.16) | 2.59 (0.14) | 2.54 (0.13) |

decreases from 3.53 mm at 20% to 2.83 mm at 80%. The corresponding decrease for SPS-6 is from 2.71 mm to 2.54 mm. Thus the results show that the proposed parameter set has lower errors than the MFCCs, and it also has a lower sensitivity to the training data size.

### *3.6.3 Place of articulation estimation using pellet locations as ANN output parameters*

An investigation was carried out using the *x-y* locations of the seven pellets on the articulators as the ANN output parameters for estimating the place of articulation and SPS-6 (proposed parameter set) as the input parameter set. Out of the various combinations of the number of layers and the number of neurons, an ANN with two hidden layers having six neurons in the first hidden layer and four neurons in the second hidden layer provided the best performance. The errors and the correlation coefficients were calculated considering all the output parameters. The pellets' *x-y* locations were estimated with an overall error of 3.10 mm. The place of articulation values were estimated from the estimated *x-y* locations using the graphical processing technique. The errors and correlation coefficients for the estimated values with reference to the PoA-art values are given in Table 3.6. The errors are significantly higher than the corresponding errors for the place of articulation estimated directly as the ANN output in Table 3.3. This increase in the errors may be due to the significant variability of the pellets' *x-y* locations due to the speaker's physiology and the pellets' placement on the articulators. The increase in errors may also be due to a larger number of neuronal weights in the network with multiple outputs.

### *3.6.4 Analysis of results for different fricative places*

The results of the optimal ANN-based estimation using the proposed spectral parameters as the input feature vector were analysed for different fricative places (labiodental, alveolar, palatal). The mean of errors, the standard deviation of errors, the RMS of errors, correlation coefficients, and significance level of correlation coefficients are given in Table 3.7. The table also provides the number of utterances and the mean and standard deviation of PoA-art

**Table 3.6:** Estimation using pellet locations as ANN output parameters: mean of errors, standard deviation (S.D.) of errors, RMS of errors, and correlation coefficient (standard deviation in parentheses).

| Spectral Parameter Set | Mean Error | S.D. Error | RMS Error | Corr. Coeff. |
|---|---|---|---|---|
| SPS-6 (MSSC, NSASS, NHBE, NVLBE, PAE, MLBE) | 0.29 (0.93) | 4.71 (0.54) | 4.79 (0.55) | 0.918 (0.018) |

**Table 3.7**: Estimation for different fricative places: mean of errors, standard deviation (S.D.) of errors, RMS of errors, and correlation coefficients for different fricatives for estimation using the spectral parameter set SPS-6, optimal ANN (two hidden layers) (N = number of utterances in the dataset).

| Fricatives | PoA-art | | | Mean Error | S.D. Error | RMS Error | Corr. Coeff. | Signi-ficance level |
|---|---|---|---|---|---|---|---|---|
| | N | Mean | S. D. | | | | | |
| Labiodentals | 3046 | 0.28 | 0.75 | −0.31 | 2.21 | 2.23 | 0.053 | 0.003 |
| Alveolars | 5279 | 23.84 | 2.63 | 0.08 | 2.55 | 2.55 | 0.467 | < 0.001 |
| Palatals | 1787 | 28.72 | 3.32 | 0.42 | 2.96 | 2.99 | 0.498 | < 0.001 |
| All | 10112 | 17.60 | 11.76 | 0.02 | 2.53 | 2.54 | 0.978 | < 0.001 |

values. The means of errors are small for all fricative places. The RMS of errors for labiodentals, alveolars, and palatals are 2.23, 2.55, and 2.99 mm, respectively. The errors for labiodentals may be attributed to the spectral shape of the low-energy unvoiced fricatives (/f/) getting affected by noise in the recordings. The errors for the alevolars may be attributed to the mixed excitation of the voiced alveolar (/z/) resulting in an occurrence of formants in the low-frequency region along with the frication noise. The sparsity of information on the tongue shape in its palatal segment may cause errors in obtaining the reference place of articulation from the pellet locations. Therefore, the higher errors for the palatals may be due to errors in the reference values of these fricatives. The error for the palatals may also be due to very few instances of the voiced palatal (/ʒ/) in the database. The errors for all three fricative places were smaller than the differences between the adjacent places, and the correlation coefficients were highly significant ($p < 0.001$) for all the fricatives except for the labiodentals. The low value of correlation coefficient for the labiodentals may be attributed to the small variation in the PoA-art values for these fricatives.

## 3.7 Discussion

The relationship between the place of articulation measured from the articulatory data and the spectral characteristics of the fricatives was investigated using the XRMB database, for

several earlier reported parameters and a set of proposed parameters. The reference place of articulation was estimated from the articulograms in the database using a graphical technique. An ANN-based technique for speaker-independent estimation of the place of articulation of the fricatives using the spectral parameters as the input feature vector and the place of articulation graphically estimated from the articulatory data as the reference was investigated. This investigation examined the effects of (i) set of input parameters, the number of hidden layers, and the number of neurons in the network, (ii) size of training data, and (iii) pellet locations as the network output parameters.

A technique was developed for automated estimation of the place of articulation from the upper and lower contours of the oral cavity image. In this technique, an axial curve of the oral cavity was iteratively estimated as an axis of symmetry, approximately bisecting the normals to it. Oral cavity opening was obtained as the distance between the contours along the normal to the axial curve, and the position of the smallest opening measured from lips was used as the place of articulation. The values estimated using the automated technique closely matched those obtained by the manual marking of the visually estimated place of maximum constriction for the oral cavity images of vowels, stops, and fricatives from the XRMB and MRI databases.

Investigation on the relationship between the place of articulation and the spectral parameters was carried out to identify the parameters suitable for speaker-independent estimation of the place of articulation. Out of the several earlier reported spectral parameters, spectral moments, spectral peak frequency, and normalized amplitude were selected for further examination. Based on a visual examination of the characteristics of the magnitude spectra of the utterances in the database, a set of six spectral parameters, including maximum-sum segment centroid (MSSC), normalized sum of absolute spectral slopes (NSASS), and four spectral energy parameters (NHBE, NVLBE, PAE, MLBE), were proposed. To examine the relation between the spectral parameters and the place of articulation obtained from the articulatory data, place of articulation values were quantized in 1-mm steps, and mean and standard deviation of the spectral parameters corresponding to each of the quantization steps were calculated. The first three spectral moments (SM1, SM2, SM3) appeared to be associated with the place of articulation, but with significant overlap. Spectral peak frequency (SPF) was associated to the place of articulation, but had a large spread for labiodentals. The fourth spectral moment (SM4) and normalized amplitude (n-Amp) showed a significant spread across the fricatives. The six proposed parameters were found to be associated with the place of articulation with a lesser spread. The spectral parameter MSSC, calculated as the centroid of contiguous large-value spectral samples of the magnitude spectrum (identified as

the maximum-sum segment), was effective in relating the energy concentration to the place of articulation. It was not affected by large extraneous samples and was particularly effective in the case of alveolars. The values of the spectral parameter NSASS, calculated from the slopes in the smoothed magnitude spectrum, for the labiodentals and palatals were well separated. Among the four spectral energy parameters, NHBE had large values for alveolars, NVLBE had large values for labiodentals, MLBE had large values for alveolars and palatals, and PAE had large values for alveolars and palatals. These observations indicate that these parameters collectively may be useful in estimating the place of articulation using machine learning.

An ANN-based estimation of the place of articulation, using a feedforward network with hyperbolic-tangent activation function in the hidden layers and place of articulation values obtained from the articulatory data as reference, was investigated to examine the suitability of different spectral parameters and size of the training data. The evaluation was performed using a five-fold cross-validation, by dividing the fricative utterances extracted from the XRMB database into five sets with four sets for training and the fifth set for validation in each step. The investigation used networks with one, two, and three hidden layers. It was carried out with six sets of input spectral parameters: set SPS-1 with four spectral moments (SM1, SM2, SM3, SM4), set SPS-2 with three spectral moments (SM1, SM2, SM3), set SPS-3 with three spectral moments (SM1, SM2, SM3) and spectral peak frequency (SPF), set SPS-4 with four spectral moments, spectral peak frequency, and normalized amplitude, set SPS-5 with MFCC (12 coefficients), and set SPS-6 with the proposed spectral parameters (MSSC, NSASS, NHBE, NVLBE, PAE, and MLBE).

Evaluation results showed that networks with two hidden layers provided better performance than one and three hidden layers. The optimal numbers of the neurons in the hidden-layer varied across the input feature vectors. The input feature vector with proposed parameters required fewer neurons than the other vectors. The input feature vector with spectral moments resulted in the highest error, and the input feature vector with proposed spectral parameters resulted in the lowest error. The investigation using different sizes of the dataset for training showed proposed spectral parameters to have a lower sensitivity to the training data size than MFCCs. Further, proposed spectral parameters resulted in smallest error for all fricatives. The errors for the labiodental, alveolar, and palatal fricatives were smaller than the differences between the adjacent places, and the correlation coefficients were significant. These results indicate suitability of the set with proposed parameters for ANN-based speaker-independent estimation of the place of articulation.

An investigation was carried out using the pellet's *x*-*y* locations on the articulators as the network output parameters, with the place of articulation obtained from the estimated *x*-*y*

locations and proposed spectral parameters as the input feature vector. The method resulted in increased errors, which may be attributed to variability in the pellets' placements on the speaker's articulators and more neuronal weights in the network with multiple outputs. The results indicate the unsuitability of pellet locations as the network output parameters for estimating the place of articulation.

Another investigation, presented in Appendix B, was carried out to examine the effect of vocal tract length variation by normalizing each speaker's acoustic space to that of a target speaker, using a frequency warping technique. It also used proposed spectral parameters as input feature vector. In this investigation, the optimal warping factors were obtained using a GMM-based acoustic model, and the ANN-based estimation of the place of articulation was carried out using the parameters obtained from the frequency-warped spectrum using the optimal warping factor. Results showed that the use of normalized spectral parameters increased the errors, indicating the unsuitability of this processing step for ANN-based estimation of the place of articulation.

Estimation of the place of articulation from spectral parameters suffers from the problem of non-uniqueness as different values of the place of articulation can produce the speech signal with similar spectral parameters. The non-uniqueness in the mapping from the spectral parameters to the place of articulation was investigated using a GMM-based model for speaker-independent estimation of the place of articulation, as presented in Appendix C. The presence of multiple peaks in the conditional probability density function for a combination of place of articulation and input parameters was considered as a non-uniqueness instance. The error related to non-uniqueness was calculated as the difference between the estimated place of articulation and the location of the peak in the conditional probability density function nearest to the reference value. These errors were found to be large for labiodentals and small for palatals. The results showed that approximately 61% of the errors in the estimation of place of articulation of all the fricatives could be attributed to the non-uniqueness.

In summary, the investigations showed the feasibility of the speaker-independent estimation of the place of articulation of fricatives from the set of proposed spectral parameters. The dataset for the ANN-based estimation had a total of 10,112 utterances. The estimation showed the RMS error using the proposed set of parameters decreased from 2.71 mm with the training set corresponding to 20% of the dataset to 2.54 mm with the training set corresponding to 80% of the dataset, indicating scope for improving the estimation by increasing the training set size. The RMS errors were lowest for the alveolars and largest for the labiodentals. The errors for palatals (2.99 mm) and alveolars (2.55 mm) were comparable to the standard deviations of the reference values. The errors for labiodentals (2.23 mm) were

smaller than their distance from the alveolars. Thus the results indicate the feasibility of ANN-based speaker-independent estimation for feedback of place of articulation of the fricatives. Normalization of the speaker's acoustic space and use of articulatory locations as the network outputs did not help reduce the errors. Another investigation showed that the estimation errors were contributed mainly by non-uniqueness in the mapping from spectral parameters to the place of articulation. An investigation for using spectral parameters during the vowel-fricative and fricative-vowel transitions to reduce the estimation errors is presented in the following chapter.

# Chapter 4
# PLACE OF ARTICULATION OF FRICATIVES FROM SPECTRAL PARAMETERS DURING FRICATION AND VOCALIC TRANSITION SEGMENTS

## 4.1 Introduction

The investigation presented in the previous chapter showed the feasibility of ANN-based estimation of the place of articulation of fricatives from the spectral parameters during the frication segment. A significant part of the estimation error could be attributed to the non-uniqueness in the acoustic-to-articulatory mapping, indicating a need for examining additional acoustic characteristics related to the place. Generally, the alveolars /s, z/ have high-frequency energy concentration, and the labiodentals /f, v/ have predominantly low-frequency energy. However, the spectral cues for different places have significant overlap and utterance-to-utterance variation. An analysis of several fricatives in the XRMB database (Westbury 1994) showed considerable variation in the relations between spectral characteristics and place. Figure 4.1(a) shows an example of /s/ with energy distributed instead of being concentrated in the high-frequency region, and Figure 4.1(b) shows an example of /f/ with energy distributed instead of being concentrated in the low-frequency region. These utterances with overlapping frication spectra are perceived distinctly, and they indicate involvement of additional cues in the perception of fricative place.

It has been established that the formant transitions in the vocalic segments preceding or following the stop closure provide an important cue for the perception of the place of stops (Stevens and Blumstein 1978). Harris (1958) investigated the importance of vocalic segment on perception of the fricative place. CV utterances with the unvoiced fricatives /f, θ, s, ʃ/ and the vowels /i, e, o, u/, from a male speaker, were used to generate a set of 64 CV sequences by interchanging the frication and vocalic segments in the utterances having the same vowel. Listening test for fricative identification by 22 listeners showed the vocalic segments to be important for perception of the non-sibilants /f, θ/ and perception of the sibilants /s, ʃ/ to be dependent primarily on the frication segment. Jongman *et al.* (2000) analyzed F2 frequency at the vowel onset as a linear regression of F2 frequency at the vowel midpoint in CV utterances with the vowels /i, e, ae, a, o, u/. The regression equations were obtained for the fricatives /f, v, θ, ð, s, z, ʃ, ʒ/ for 20 speakers. It was reported that the F2 frequency at the vowel onset and the regression parameters did not discriminate all four places of articulation, but their values for /f, v/ were significantly different from other fricatives. Wagner *et al.* (2006) used pseudo

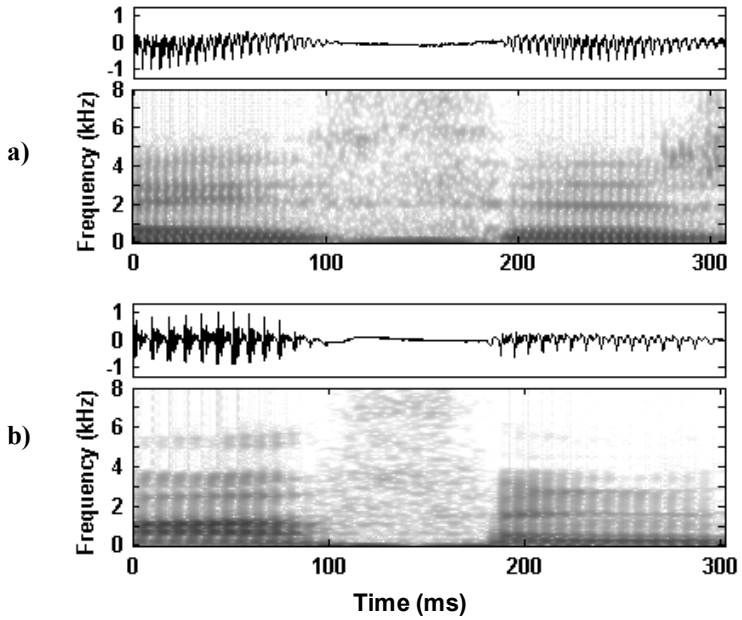



**Figure 4.1** Examples, from the XRMB database, of fricatives with frication spectra deviating from the commonly reported characteristics: Waveform and spectrogram of (a) /ɛsʌ/ (speaker JW62, task 78), (b) /afʌ/ (speaker JW44, task 24).

words involving misleading and coherent formant transitions for /s/ and /f/ preceded and followed by /a, i, u/ spoken by Dutch and Spanish speakers for fricative identification tests. Listening tests were carried out with native Dutch, English, German, Polish, and Spanish speakers as listeners. The Dutch and German listeners were not affected by the misleading formant transitions, as both the languages do not have spectrally similar fricatives. The other listeners were confused by the misleading formant transitions. The confusion was attributed to the spectral similarity of /f/ and /θ/ in English and Spanish and the similarity of /s/ with /ʃ, ɕ/ in Polish. It was concluded that the formant transitions provided important cues for the English, Polish, and Spanish listeners. Mean response times for the stimuli with the misleading formant transitions were longer than those with the coherent formant transitions for all listeners, indicating that the listeners were sensitive to the transitions.

The fricative perception is also affected by duration (Baum and Blumstein 1987, Jongman 1989). For studying the importance of the frication duration, Jongman (1989) conducted listening tests on 12 subjects using CV utterances comprising the fricatives /f, v, θ, ð , s, z, ʃ, ʒ/ and the vowels /a, i, u/ spoken by a male speaker, with the utterances edited to include 20–70 ms frication in 10-ms steps and the entire frication. The results showed that a 30–50 ms frication was required to identify /f, v, s, z, ʃ, ʒ/, while the entire frication was required to identify /θ, ð/. A decrease in the frication duration affected the place identification more than the manner and voicing identifications.

Earlier studies have examined effects of the vocalic transition and the frication on perception of the place of articulation of fricatives separately. As the two effects may interact, it may be interesting to investigate the vocalic transition's effect on the perception of fricatives with different frication durations. A perceptual study is carried out to investigate the relative importance of the transition and the frication on perception of the unvoiced fricatives /f, s, ʃ/. For quantifying the potential of the vocalic transition in improving the ANN-based estimation of the place of articulation, the place of articulation is estimated using a combination of spectral parameters during the frication and vocalic transition segments as the input feature vector. The perpetual study is presented in the second section, followed by the ANN-based estimation of the place of articulation in the third section and the discussion in the last section.

## 4.2 Effects of vocalic transition and frication on the perception of fricatives in VCV utterances

A perceptual study is conducted to investigate the relative importance of the vocalic transition and the frication on perception of the unvoiced fricatives /f, s, ʃ/. It involves listening tests for consonant identification using vowel-fricative-vowel sequences, generated using the vocalic transition segments extracted from natural utterances and the frication segments synthesized with a duration of 50–300 ms. The following subsections present the speech material, experimental method, test results, and analysis of the results.

### *4.2.1 Speech material*

A set of VCV utterances with the unvoiced fricatives /f, s, ʃ/ and the vowel /a/ was recorded at a sampling frequency of 16 kHz from a male speaker. The utterances were verified for acceptable quality and intelligibility. The fricative /θ/ was not included as some listeners confused it with a stop having nearly the same place in their first language. For studying the effects of vocalic transition and frication on perception of the place of fricatives, the test stimuli with different frication durations were generated. Multiple utterances with the same vowel-fricative-vowel have significant utterance-to-utterance and duration-related variations in the frication spectrum. For avoiding the effects of these variations on the test results, synthesized frication segments were used in the test stimuli.

The recorded VCV utterances were edited, using a visual examination of the waveform and its spectrogram, for extracting the VC and CV segments. The frication segment for each fricative was synthesized by filtering white noise using an FIR filter, with its magnitude response approximating the averaged magnitude spectrum of the frication segment in the natural utterance calculated using 30-ms windows with 5-ms window shift. The synthesized

frication was scaled to have the same RMS value as the natural frication. Each VCV sequence for the test material was generated by concatenating the extracted VC segment, the synthesized frication segment, and the extracted CV segment. The frication amplitude envelope was multiplied with a trapezoidal window with 20-ms rising and falling sub-segments. A faster rise-fall resulted in a perceptible discontinuity in the concatenated sequence, and a slower rise-fall decreased the frication's steady part.

For studying the effect of frication duration, the VCV sequences were generated with the frication duration varying from 50 ms to 300 ms. Sequences with combinations of seven frication durations (50, 70, 90, 120, 150, 200, 300 ms), three vocalic transitions, and three fricatives, resulting in 63 VCV sequences, were used as the test stimuli in the listening test.

### *4.2.2 Experimental method*

Ten normal-hearing adults (eight male and two female students, 22–35 years) served as the listening test subjects. They had studied English as their first or second language and could perceive the differences between the fricatives used in the study. The test was administered using a graphical user interface (GUI) on a PC-based automated experimental setup for consonant identification. The stimuli comprising the 63 VCV sequences, as described in the preceding subsection, were presented in a randomized order using a Sennheiser HD 202 headphone at the most comfortable level for the listener. The GUI had four buttons marked as 'start', 'play', 'next', and 'pause' to control the stimulus presentation and seven response buttons. A pilot listening test showed that some of the stimuli, particularly those with short frication segments, were perceived as the stops /p, t/ and affricate /t͡ʃ/. Therefore, the response buttons included three buttons for the fricatives /f, s, ʃ/, two buttons for the stops /p, t/, a button for the affricate /t͡ʃ/, and a button for 'none of above'. The task involved listening to the stimulus by clicking the 'play' button and identifying the consonant by clicking one of the response buttons. The subject could listen to a sound more than once before responding. The 'next' button was clicked to proceed to the next stimulus. The 'pause' and 'start' buttons could be used to pause and resume the test, respectively. The test continued until all the test stimuli were presented and the responses were recorded.

### *4.2.3 Listening test results*

The listening test had 63 stimuli and ten subjects, resulting in 630 responses. A plot of the identification scores is shown in Figure 4.2, with each bar representing the scores for a stimulus.

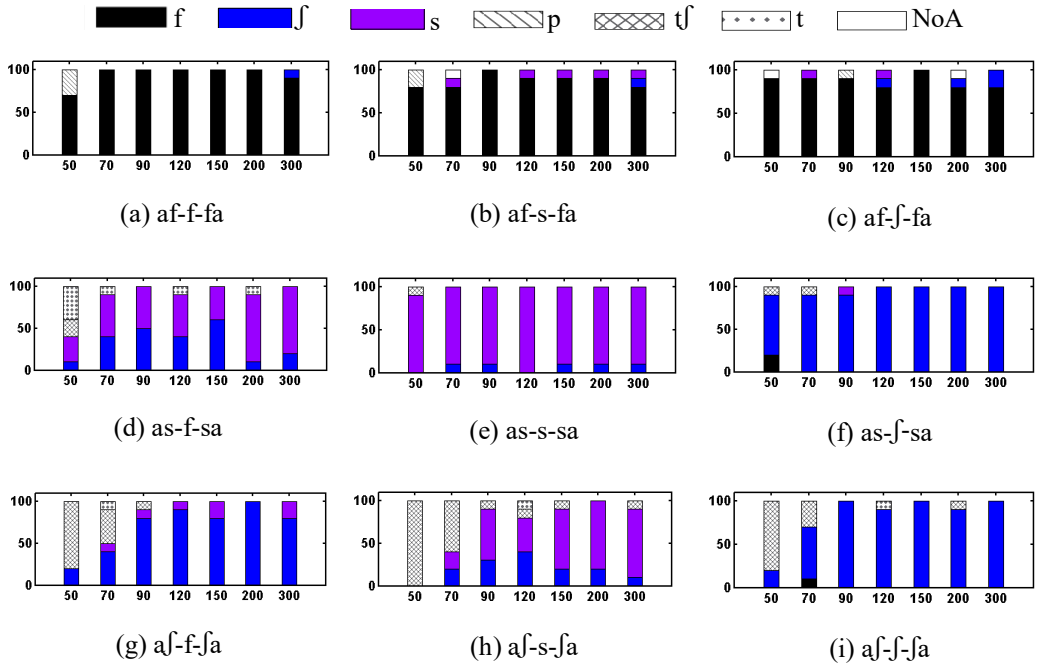



**Figure 4.2** Consonant identification scores for the test sequences with different transition and frication combinations (e.g. 'af-s-fa' being the test sequence with the VC and CV transition segments from the recorded utterance /afa/ and the synthesized frication corresponding to /s/) and different frication durations: seven response scores (%) as shaded bars (response 'none of above' abbreviated as NoA) along the y-axis and duration in ms along the x-axis.

The first row (a, b, c) of Figure 4.2 shows the results for test sequences generated using the VC and CV segments corresponding to the utterance /afa/. For the frication durations of 50 ms and longer, the consonants in most of the test sequences are perceived as /f/ and the responses are 70% or higher irrespective of the frication segments' spectral characteristics. These results indicate that the perception of the test sequences having the transition corresponding to /f/ is not affected by the frication's spectral characteristics. The transition's effect in this case is not reduced by an increase in the frication duration.

The second row (d, e, f) of Figure 4.2 shows the results for test sequences generated using the VC and CV segments corresponding to the utterance /asa/. For the frication durations of 50 ms and longer and the frication segment corresponding to /s/, the consonants in the test sequences are perceived as /s/ with 70% or higher responses. For the frication durations of 70–300 ms and the frication segment corresponding to /f/, the consonants are perceived as /s/ or /ʃ/, indicating that transition dominates the perception. The consonants in the test sequences with the frication segments corresponding to /ʃ/ are perceived as /ʃ/, indicating that frication dominates the perception in this case.

The third row (g, h, i) of Figure 4.2 shows the results for test sequences generated using the VC and CV segments corresponding to the utterance /aʃa/. For the frication duration of 50

ms, the consonants are perceived as affricate /t͡ʃ/, which has the place corresponding to the transition. The responses for many of these sequences are 70% or higher, indicating that the transition dominates the perception. For the frication durations of 70–300 ms and the frication segments corresponding to /ʃ/, the consonants are perceived as /ʃ/ with 70% or higher responses. For the frication durations of 90 and 120 ms and the frication segments corresponding to /s/, the consonants are perceived as /s/ or /ʃ/ with 70% or lower responses, indicating a similarity of the transitions for /s/ and /ʃ/. For the frication durations of 150 ms and longer and the frication segment corresponding to /s/, the consonants are perceived as /s/ with 70% or higher responses, indicating that the longer frication duration helped discrimination between /s/ and /ʃ/. These results show that the frication dominates the perception when the transition corresponds to /ʃ/ and the frication corresponds to /s/. However, when the frication corresponding to /f/ is combined with the transition corresponding to /ʃ/, the consonants are perceived as /ʃ/ with 70% or higher responses, indicating that the transition dominates the perception.

*4.2.4 Analysis of the test results*

Investigations similar to the present one have been reported by Harris (1958), Wagner *et al.* (2006), and Jongman (1989), as described in the first section. Harris (1958) concluded that perception of the fricatives /s, ʃ/ depends on the frication part alone, irrespective of the transition segment. The results of the present study show that the perception of frication corresponding to /s, ʃ/ paired with the transitions corresponding to /f/ is dominated by the transition segment. It may be noted that the present study uses /a/, while this vowel was not used by Harris (1958). Results of the study by Wagner *et al.* (2006) are similar to the present study for /f, s/. The present study also has /ʃ/, and Wagner *et al.* (2006) did not investigate the effects of the changes in frication duration and transition segment together. In the study by Jongman (1989), the frication duration effect was studied without interchanging the transition segments. Thus it did not investigate the effect of transition segments on the fricative perception as in the present study.

Information transmission analysis has been used in many studies to analyze the stimulus-response confusion matrices of the listening tests for quantifying the contributions of different input features for consonant identification (Miller and Nicely 1955, Jongman 1989, Hornsby and Ricketts 2001, Xu *et al.* 2005, Zhou *et al.* 2010). As described by Miller and Nicely (1955), the relative information transmitted from the stimulus set **x** to the response set **y** is given as

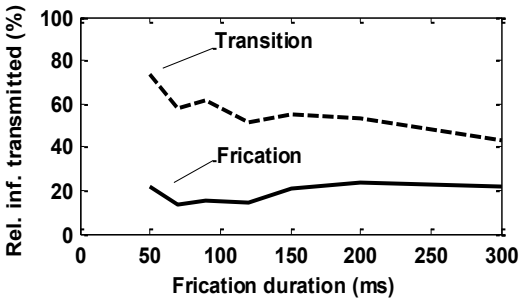



**Figure 4.3** Relative information transmitted (%) for transition and frication features as a function of frication duration.

$$I_{rel}(\mathbf{x},\mathbf{y}) = \frac{-\sum_{i,j} p(x_i, y_j)\ \log\left[ p(x_i)p(y_j) / p(x_i, y_j) \right]}{-\sum_{i} p(x_i)\ \log\left[ p(x_i) \right]} \tag{4.1}$$

where $p(x_i)$ is the probability of the stimulus $x_i$, $p(y_j)$ is the probability of the response $y_j$, and $p(x_i, y_j)$ is the joint probability of the stimulus $x_i$ and the response $y_j$.

For each frication duration in our listening test, the stimulus-response scores were used to obtain confusion matrices for the frication and transition features. The stimulus set for the frication feature comprised the three frications (/f/, /s/, /ʃ/), and that for the transition feature comprised the three transitions (/af/, /as/, /aʃ/). The response set for both features comprised the seven responses (/f/, /s/, /ʃ/, /p/, /t/, /t̲ʃ/, none of above). These confusion matrices were used to calculate the relative information transmitted for the two features.

Figure 4.3 shows the relative information transmitted for the transition and frication features as functions of the duration. The transition feature provides more information than the frication feature for all durations as transition dominates the perception in most cases. It may be partly due to the experimental conditions, as there were more conditions when the frication and transition did not match. For the frication durations above 50 ms, the information transmitted by transition decreases and that by frication increases. This variation may be attributed to the cases in which frication dominates the perception, i.e., the cases with vocalic transition and frication corresponding to either /s/ or /ʃ/.

Spectrograms during the VC and CV transitions of multiple VCV utterances with the vowel /a/ and fricatives /f, s, ʃ/, from a speaker in our recording, were examined for spectral cues for the place during misleading transitions. The variation of the second formant frequency during the transitions, as observed in the spectrograms, was consistently related to the place. During the VC transition for /f/, the second formant decreased by about 200 Hz. During the VC transition for /s/ and /ʃ/, the second formant increased by about 100 Hz and 200 Hz, respectively. Thus, the second formant transitions for the labiodentals differed from the alveolars and palatals. The similarity in the CV transitions for the alveolars and palatals may be the reason for domination of perception of the test sequences with these fricatives by

the frication segment. Formant transitions during the CV transitions were complementary to those during the VC transitions.

## 4.3 Estimation of the place of articulation using spectral parameters during frication and vocalic transition segments

Investigation on perception of the unvoiced fricatives /f, s, ʃ/, presented in the previous section, showed a significant effect of the vocalic transition on the fricative perception, particularly for short frication duration. Based on this result, it may be assumed that the vocalic transition's spectral characteristics may help improve the estimation of place of articulation of the fricatives. An investigation is carried out to extend the ANN-based estimation of the place of articulation of the fricatives using the input feature vector comprising the spectral parameters during the frication and vocalic transition segments.

### *4.3.1 Material and method*

Utterances with the voiced fricatives /v, z, ʒ/ and the unvoiced fricatives /f, s, ʃ/ along with a preceding or following vowel segment extracted from the VCV utterances, words, and sentences in the XRMB database spoken by 47 speakers (21 male, 26 female) in 24 tasks were used as the speech material. The dataset for the investigation using the frication segment, as described in the third chapter, had 10,112 utterances. Some of those utterances did not have vocalic transition segments and hence were excluded. The dataset for the current investigation had 8,855 utterances.

The earlier investigation for the fricative place estimation using the frication segment's spectral parameters showed that the proposed set of spectral parameters comprising MSSC, NSASS, NHBE, NVLBE, PAE, and MLBE, referred to as SPS-6, resulted in the lowest estimation error. The investigation is extended by using the same set of parameters calculated during the vocalic transition segment. The parameters during transition segments were calculated from the magnitude spectrum obtained using a 512-point FFT with Hanning window of the 20-ms transition segment immediately preceding or following the frication segment. In the utterances with the vocalic transitions on both sides of the frication, the transition following the frication was used for the spectral parameters. The proposed parameter set calculated for the frication segment is referred to as PS-F, and the spectral parameter set calculated for the transition segment is referred to as PS-T. The MFCC parameter set was also calculated using the transition segment and is referred to as MFCC-T. The vocalic transitions between fricative and vowel, for the fricatives with the same place of articulation, vary based on the adjacent vowel. Therefore, including the spectral parameters

during the adjacent vowel segments along with the spectral parameters during the transition segments may improve the estimation of the place of articulation. The parameters were calculated from the magnitude spectrum obtained using a 512-point FFT with Hanning window of a 20-ms segment in the adjacent vowel, and the resulting parameter set is referred to as PS-V. The MFCC parameter set was also calculated for this segment, and it is referred to as MFFC-V.

The investigation for ANN-based fricative place estimation was carried out using the networks similar to those in the earlier investigation, different input parameter sets, the place of articulation as the output parameter, and the articulatory place PoA-art as the reference.

The utterances extracted from the XRMB database were divided, maintaining a balance in terms of gender and speakers, into five utterance sets: set-1 with four male and five female speakers and 1752 utterances, set-2 with four male and five female speakers and 1746 utterances, set-3 with four male and five female speakers and 1745 utterances, set-4 with four male and six female speakers and 1789 utterances, and set-5 with five male and five female speakers and 1823 utterances. The evaluation was performed using these sets in a five-fold cross-validation manner, with four sets for training and the fifth set for validation in each step, as in the earlier investigation.

### *4.3.2 Investigation*

The investigation was carried out using the following nine sets of spectral parameters:

(i) PS-F,
(ii) union of PS-F and PS-T (PS-F + PS-T),
(iii) union of PS-F and PS-V (PS-F + PS-V),
(iv) union of PS-F, PS-T, and PS-V (PS-F + PS-T + PS-V),
(v) MFCC-F (first 12 MFCC coefficients calculated from the frication segment),
(vi) union of MFCC-F and MFCC-T (MFCC-F + MFCC-T),
(vii) union of MFCC-F and MFCC-V (MFCC-F + MFCC-V),
(viii) union of MFCC-F, MFCC-T, and MFCC-V (MFCC-F + MFCC-T + MFCC-V), and
(ix) union of PS-F, MFCC-T, and MFCC-V (PS-F + MFCC-T + MFCC-V).

For each spectral parameter set, several networks, differing in the number of hidden layers and the number of neurons, were used to find the optimal one. The effect of the number of hidden layers was examined for one, two, and three hidden layers. The effect of the number of neurons was examined using the networks differing in the number of neurons in the hidden layers.

### *4.3.3 Results of ANN-based place estimation*

The fricative place estimation was evaluated using mean and standard deviation of the errors with reference to the PoA-art values, for the five-fold cross validation. The RMS of errors was calculated as a composite error measure, and the correlation coefficient was calculated as a measure of association of the estimated values with the reference values. The performance measures were examined as in Subsection 3.6.1 of the third chapter. For each of the nine spectral parameter sets, the optimal number of hidden layers and the number of neurons in each hidden layer were obtained for the lowest error.

For all the spectral parameter sets, a network with two hidden layers was found to be optimal. The optimal numbers of neurons in the two hidden layers for networks with different spectral parameter sets and the corresponding performance measures are given in Table 4.1. The mean errors are negligible in all the cases. The standard deviations of the errors and the RMS of errors are comparable. A decrease in the RMS of errors is associated with an increase in the correlation coefficient.

The spectral parameter set PS-F + PS-T has slightly higher error than PS-F. It also requires a larger number of neurons. The spectral parameter set PS-F + PS-V has slightly lower error than PS-F, and it requires a larger number of neurons. The spectral parameter set PS-F + PS-T + PS-V results in lower error and smaller number of neurons. Thus, the results show that a combination of the transition and vowel-segment parameters improved the place estimation, although only one of them did not.

The errors and the numbers of neurons for MFCC-F are higher than those for PS-F. The spectral parameter sets MFCC-F + MFCC-T, MFCC-F + MFCC-V, and MFCC-F + MFCC-T + MFCC-V result in decreased errors. However, the errors and the numbers of neurons for them are much higher than those for PS-F. The spectral parameter set PS-F + MFCC-T + MFCC-V results in a much lower error than PS-F and requires a lower number of neurons, providing the best performance across all the spectral parameter sets.

The RMS errors for the different fricatives for the place estimation using the different spectral parameter sets are given in Table 4.2. The spectral parameter set PS-F + PS-T has a lower error than PS-F for /f/ and higher errors for /v/ and /ʃ/. The spectral parameter set PS-F + PS-V has lower errors than PS-F for /f/, /ʃ/, and /ʒ/ and a higher error for /v/. The spectral parameter set PS-F + PS-T + PS-V has lower errors than PS-F for /f/ and /ʃ/ and a higher error for /v/. The spectral parameter set MFCC-F has the highest errors for all the fricatives. Its combinations with MFCC-T and MFCC-V result in decreased errors for some fricatives and increased errors for others. The spectral parameter set PS-F + PS-T + PS-V has

**Table 4.1**: Mean error, standard deviation (S.D.) of errors, RMS of errors. and correlation coefficients across 5-fold validation sets using different spectral parameter sets and networks with two hidden layers, along with the number of neurons in the hidden layers (N1: number of neurons in the first hidden layer, N2: number of neurons in the second hidden layer).

| Spectral Parameter Set | No. of neurons | | Errors with respect to PoA-art (mm) | | | | | | Corr. Coeff. | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Mean Error | | S.D. Error | | RMS Error | | | |
| | N1 | N2 | Mean | S.D. | Mean | S.D. | Mean | S.D. | Mean | S.D. |
| **PS-F** | **12** | **2** | **0.04** | **0.16** | **2.61** | **0.34** | **2.62** | **0.33** | **0.976** | **0.005** |
| PS-F + PS-T | 10 | 7 | 0.00 | 0.23 | 2.64 | 0.29 | 2.65 | 0.29 | 0.976 | 0.005 |
| PS-F + PS-V | 3 | 13 | 0.04 | 0.30 | 2.60 | 0.26 | 2.61 | 0.26 | 0.976 | 0.005 |
| PS-F + PS-T + PS-V | 4 | 8 | 0.11 | 0.30 | 2.57 | 0.20 | 2.59 | 0.20 | 0.978 | 0.004 |
| MFCC-F | 26 | 5 | −0.04 | 0.30 | 2.87 | 0.22 | 2.88 | 0.22 | 0.972 | 0.008 |
| MFCC-F + MFCC-T | 28 | 4 | −0.01 | 0.34 | 2.77 | 0.15 | 2.79 | 0.16 | 0.974 | 0.005 |
| MFCC-F + MFCC-V | 25 | 2 | 0.06 | 0.35 | 2.81 | 0.41 | 2.82 | 0.42 | 0.972 | 0.008 |
| MFCC-F + MFCC-T + MFCC-V | 18 | 20 | 0.05 | 0.25 | 2.78 | 0.12 | 2.79 | 0.12 | 0.972 | 0.004 |
| **PS-F + MFCC-T + MFCC-V** | **3** | **3** | **−0.01** | **0.21** | **2.55** | **0.12** | **2.55** | **0.12** | **0.978** | **0.004** |

similar errors as PS-F + MFCC-T + MFCC-V for /s/, /z/, and /ʃ/, a higher error for /ʒ/, and a lower error for /f/.

Considering all fricatives, the spectral parameter set PS-F has an error of 2.62 mm and requires 14 neurons, while the spectral parameter set PS-F + PS-T + PS-V has an error of 2.57 mm and requires 12 neurons. The spectral parameter set PS-F + MFCC-T + MFCC-V has an error of 2.55 mm and requires six neurons. Thus, the results show that the estimation of place of articulation of the fricatives is improved by the inclusion of the vocalic information represented by the transition and vowel-segment parameters. The results further indicate that PS-F is better than MFCC-F for representing the frication, and MFCC-T + MFCC-V is slightly better than PS-T + PS-V for representing the vocalic transition.

## 4.4 Discussion

Two investigations have been presented in this chapter. The first investigation involved a perceptual study of the relative importance of vocalic transition and frication on the perception of the unvoiced fricatives /f, s, ʃ/. The second investigation involved ANN-based

**Table 4.2**: RMS errors for different fricatives for estimation using different spectral parameter sets.

| Spectral Parameter Set | RMS errors with respect to PoA-art (mm) | | | | | | |
|---|---|---|---|---|---|---|---|
| | All | f | v | s | z | ʃ | ʒ |
| PS-F | 2.62 | 2.48 | 1.81 | 2.44 | 2.77 | 3.02 | 4.09 |
| PS-F + PS-T | 2.65 | 2.34 | 1.92 | 2.46 | 2.76 | 3.13 | 4.06 |
| PS-F + PS-V | 2.61 | 2.36 | 1.96 | 2.48 | 2.77 | 2.96 | 3.98 |
| PS-F + PS-T + PS-V | 2.59 | 2.25 | 2.20 | 2.44 | 2.77 | 2.85 | 4.06 |
| MFCC-F | 2.88 | 2.65 | 2.14 | 2.65 | 3.13 | 3.16 | 4.93 |
| MFCC-F + MFCC-T | 2.79 | 2.49 | 2.25 | 2.52 | 2.83 | 3.15 | 4.87 |
| MFCC-F + MFCC-V | 2.82 | 2.52 | 1.79 | 2.69 | 3.02 | 3.47 | 3.79 |
| MFCC-F + MFCC-T + MFCC-V | 2.79 | 2.16 | 2.11 | 2.62 | 3.14 | 3.33 | 3.87 |
| PS-F + MFCC-T + MFCC-V | 2.55 | 2.47 | 1.98 | 2.39 | 2.80 | 2.90 | 3.24 |

estimation of place of articulation of the fricatives using spectral parameters during frication and vocalic transition as the input feature vector.

In the first investigation, a listening test for consonant identification was conducted using vowel-fricative-vowel sequences involving the vowel /a/ and the unvoiced fricatives /f, s, ʃ/ and generated using the vocalic segments extracted from natural utterances and the frication segments synthesized with the durations of 50–300 ms. Some of the sequences with durations of 50 ms and 70 ms were perceived as stops with the place determined by the transition. The sequences with 90 ms and longer frication durations were perceived as fricatives, with the place determined by a combination of the transition and frication segments. For sequences with transition corresponding to /f/, the perception was not affected by the frication segments. For sequences with the transition corresponding to /s/ or /ʃ/ and the frication corresponding to /f/, the perception was dominated by the transition. For sequences with transition and frication both corresponding to /s/ or /ʃ/, the perception was dominated by the frication. Thus, it may be concluded that the transition dominates the perception in the case of a mismatch between the transition and frication segments. The investigation needs to be extended for test sequences with other fricatives, different vowel contexts, and speech material from male and female speakers.

Based on the first investigation results, it may be assumed that the fricative place estimation may be improved by supplementing the frication parameters with the vocalic transition parameters. The second investigation involved an ANN-based estimation of the

fricative place using different combinations of the spectral parameters during the frication segments and those during the vocalic transitions as the input feature vector. The spectral parameters during the frication as proposed in the investigation in the third chapter formed the spectral parameter set PS-F. The parameters for a 20-ms segment immediately preceding or following the frication segment formed the spectral parameter set PS-T, and the parameters for a 20-ms segment in the adjacent vowel formed the spectral parameter set PS-V. The spectral parameter sets PS-T, PS-V, and PS-T + PS-V were examined for providing the vocalic transition information to supplement the spectral parameter set PS-F. The MFCC parameters were also calculated similarly, forming the spectral parameter sets MFCC-F, MFCC-T, and MFCC-V. Investigation for estimation of place of articulation were carried out using the spectral parameter sets PS-F, PS-F + PS-T, PS-F + PS-V, PS-F + PS-T + PS-V, MFCC-F, MFCC-F + MFCC-T, MFCC-F + MFCC-V, MFCC-F + MFCC-T + MFCC-V, and PS-F + MFCC-T + MFCC-V as the input feature vectors.

The results of the place estimation showed no reduction in errors by increasing the number of hidden layers beyond two for all the input parameter sets. The spectral parameter set MFCC-F resulted in the highest errors. The spectral parameter sets MFCC-F + MFCC-T, MFCC-F + MFCC-V, and MFCC-F + MFCC-T + MFCC-V resulted in lower errors than MFCC-F, indicating that MFCCs are suitable to capture the place-related information in the vocalic transitions. A comparison of the performance of PS-F, PS-F + PS-T, and PS-F + PS-V showed that the inclusion of PS-T or PS-V alone did not improve the place estimation. The errors for different fricatives did not show any specific pattern. The spectral parameter set PS-F + MFCC-T + MFCC-V had the lowest errors and number of neurons. The spectral parameter set PS-F + PS-T + PS-V had a comparable error but required a larger number of neurons. Thus, the inclusion of spectral parameters during the transition and the vowel segment improved the place estimation.

The investigation for place estimation from the spectral parameters during the frication segments as presented in the previous chapter used 10,112 utterances. Excluding the utterances without a vocal transition segment, the dataset for the current investigation had 8,855 utterances. The results show that the place estimation is improved by supplementing the frication information with the vocalic information represented by the transition-segment and vowel-segment parameters. The proposed parameter set for the frication segment resulted in an error of 2.62 mm, and it required 14 neurons. The inclusion of the proposed parameter set for the transition segment or vowel segment alone did not decrease error. However, the inclusion of the proposed parameter set for the transition and vowel segments resulted in an improved error of 2.57 mm, and the combined spectral parameter set required 12 neurons.

The combination of the proposed set of parameters calculated for the frication segment and the MFCC parameters for the transition and vowel segments resulted in the lowest error of 2.55 mm, and it required six neurons. The results further show that the proposed parameter set is better than MFCCs for representing the frication and slightly inferior to MFCCs for representing the vocalic transition. The use of these spectral parameter sets avoids the difficulties associated with formant tracking during the transition and frication segments. Therefore, it may be considered suitable for ANN-based speaker-independent estimation for feedback of place of articulation of the fricatives.

# Chapter 5
# SUMMARY AND CONCLUSION

## 5.1 Introduction

Lack of auditory feedback becomes an impediment to speech acquisition for children with hearing impairment. Speech-training aids providing visual feedback of articulatory efforts not visible on the speaker's face are found to be useful in improving speech articulation. This feedback can be conveniently provided by estimating the articulatory parameters from the speech signal. Most of the speech training aids use LP-based speech analysis to estimate the vocal tract shape, but the LP-based analysis is not suited for fricatives. Methods for estimating articulatory parameters during the fricatives may help improve the effectiveness of the speech-training aids.

The research objective was to develop a method for estimating the place of articulation of the fricatives. The XRMB acoustic-articulatory database was used to study the relationship between the place of articulation and the spectral parameters and to evaluate the estimation method. The reference place of articulation was estimated from the articulograms of the database using a graphical technique to estimate an axial curve of the oral cavity contours. The relationship between the spectral characteristics and the place of articulation of the fricative segments was investigated for several earlier reported spectral parameters and a set of additional proposed parameters. An investigation was carried out for ANN-based speaker-independent mapping from spectral parameters of the frication segments to the place of articulation. A perceptual study examined the relative importance of vocalic transition and the frication on place perception. An investigation was carried out to improve the ANN-based estimation of the place of articulation, by including the spectral parameters during the vocalic transition segments in the input parameter set. Summary of the investigations, conclusions, and some suggestions for further research are presented in the following sections.

## 5.2 Summary of the investigations

The investigations reported in the preceding chapters are summarized in the following paragraphs.

*1) Estimation of place of articulation from articulograms:* Direct imaging provides the oral cavity's upper and lower contours. Irregular shapes of these contours make it difficult to locate the maximum constriction consistently and find its distance from the lips. As a solution to this problem, an automated technique was developed for estimating the place of articulation

by graphical processing of the upper and lower contours of the oral cavity image. In this technique, an axial curve is iteratively estimated as an axis of symmetry of the oral cavity. Distance between the contours along the normal to the axial curve gives the oral cavity opening, and the position of the smallest opening provides the place of articulation. The technique and its evaluation have been presented in Section 3.3. The evaluation results showed that the values estimated using the automated technique closely matched those obtained by manual marking for the vowels, stops, and fricatives in the XRMB and MRI databases.

*2) Relationship of place of articulation with spectral parameters during frication segments:* The frication segments of the utterances with the voiced fricatives /v/, /z/, /ʒ/ and the unvoiced fricatives /f/, /s/, /ʃ/ in the XRMB database were used to study the relationship between the place of articulation and the spectral parameters. From the earlier reported spectral parameters, four spectral moments (SM1, SM2, SM3, SM4), spectral peak frequency (SPF), and normalized amplitude (n-Amp) were selected for investigation. A set of six spectral parameters were proposed after a visual examination of the average magnitude spectra corresponding to different values of the place of articulation. The proposed parameters are maximum-sum segment centroid (MSSC), normalized sum of absolute spectral slopes (NSASS), and four spectral energy parameters (NHBE, NVLBE, PAE, MLBE). The calculations and the plots of mean and standard deviation of the spectral parameters versus the place have been presented in Subsections 3.4.1 and 3.4.2, respectively. The plots showed SM1, SM2, SM3, and SPF to be associated with the place of articulation but with significant overlap across the fricatives. The SM4 and n-Amp values also had a significant overlap across the fricatives. The proposed parameters showed a lesser overlap across the fricatives. The MSSC values, calculated as the centroid of contiguous spectral samples in the segment with maximum-sum, were well separated for the alveolar and labiodental fricatives. The NSASS values, calculated as the sum of absolute values of the first difference of the spectrum, were well separated for the palatal and labiodental fricatives. The spectral energy parameters were able to resolve the confusion by providing additional spectral shape information. The spectral parameters' usefulness to estimate the place of articulation could not be quantified due to the overlap of values across fricatives.

*3) ANN-based estimation of place of articulation from spectral parameters during frication segments:* A feedforward ANN with multiple hidden layers and hyperbolic tangent activation function was used to obtain a speaker-independent mapping from the spectral parameters to the place of articulation. The investigation used different sets of spectral parameters as the input feature vectors and the place of articulation estimated from the

articulograms as the reference. Effects of the input parameter set, the number of hidden layers, the number of neurons in the hidden layers, and *x-y* positions of pellet locations as the network output parameters were examined to find an optimal combination. The dataset had a total of 10,112 utterances. The effect of the size of the training set on the place of articulation estimation was also examined using 20%, 40%, 60%, and 80% of the dataset for training and 20% of the dataset for validation. The details of this investigation and the results have been presented in Sections 3.5 and 3.6, respectively. Networks with two hidden layers were found to be adequate for all input parameter sets. Evaluation using five-fold cross-validation showed that the spectral moments resulted in the largest errors, and the proposed set of parameters resulted in the lowest errors. The RMS error using the proposed set of parameters decreased from 2.71 mm with the training set corresponding to 20% of the dataset to 2.54 mm with the training set corresponding to 80% of the dataset, indicating that the estimation can be improved by increasing the training set size. The errors for the alveolars and palatals were comparable to the standard deviation of the reference values estimated from articulograms. The errors for labiodentals were larger than the standard deviation of the reference values, but much smaller than the distance between the mean place of articulation values of labiodentals and alveolars. Thus, the results indicated that the input feature vector with the proposed spectral parameters is suitable for estimating the place of articulation using the ANN-based speaker-independent mapping. Estimation using *x-y* positions of the pellets on the articulators as the network output parameters resulted in larger error than that using the place of articulation values as the output parameter, which could be attributed to the variability of *x-y* positions across speakers due to differences in the oral cavity size and the pellets' placement on the speaker's articulators. An investigation involving normalization of the speaker's acoustic space, as presented in Appendix B, indicated this processing step's unsuitability for fricative place estimation. Another investigation, presented in Appendix C, showed that 61% of the estimation error was contributed by non-uniqueness in the mapping from spectral parameters to the place of articulation.

4) *Effect of frication and vocalic transition on the fricative perception:* A significant contribution of non-uniqueness to the RMS error in place estimation indicated that the spectral characteristics during frication provide only partial information related to the place. A perceptual study was conducted to investigate the relative importance of the vocalic transition and the frication on the perception of the unvoiced fricatives /f, s, ʃ/. Listening tests for consonant identification were conducted using VCV test stimuli, generated using the vowel and transition segments extracted from natural utterances and the frication segments synthesized with durations of 50–300 ms. The experimental method and the test results have

been presented in Section 4.2. The results showed that the sequences with frication durations of 90 ms and longer were perceived as fricatives, with the perception determined by a combination of the vocalic transition and frication. The perception for the sequences with transition corresponding to /f/ was determined by the transition segments alone. For sequences with transitions corresponding to /s/ and /ʃ/, the perception was determined by a combination of the transition and frication segments. Therefore, it may be assumed that estimation of the place of articulation of fricatives may be improved by combining the information from the vocalic transition and frication segments.

5) *ANN-based estimation of place of articulation from spectral parameters during frication and vocalic transition segments:* An ANN-based estimation of the place of articulation was carried out using a combination of the spectral parameters during the frication and vocalic transition segments as the input feature vector. Out of the 10,112 utterances in the dataset for the earlier investigation, the utterances without a vocal transition segment were excluded, and the dataset for the current investigation had 8,855 utterances. The proposed set of spectral parameters calculated using a 20-ms segment adjacent to the fricative were used as the transition-segment parameters, and the same set of spectral parameters calculated using a 20-ms segment in the adjacent vowel were used as the vowel-segment parameters. The MFCC parameters were also calculated similarly. The effect of different combinations of the spectral parameters during the vocalic transition and frication segments was examined. The investigation method and results have been presented in Subsections 4.3.2 and 4.3.3, respectively. The results showed that networks with two hidden layers were adequate for all input parameter sets. The proposed parameter set for the frication segments resulted in an RMS error of 2.62 mm, and it required 14 neurons. The inclusion of the proposed parameter set for the transition segment or the vowel segment alone did not improve the place estimation. The RMS error decreased to 2.57 mm using the combination of the proposed parameter set for the frication, transition, and vowel segments. This combined parameter set required 12 neurons. The estimation using the combination of the proposed set of parameters during the frication segment and the MFCC parameters during the transition and vowel segments resulted in the smallest RMS error of 2.55 mm, and it required six neurons. The results further show that the proposed parameter set is better than MFCCs for representing the frication and slightly inferior to MFCCs for representing the vocalic transition.

## 5.3 Conclusions

The conclusions from the investigations using the spectral parameters during the frication segment may be summarized as follows.

(i) The proposed set of six spectral parameters, comprising maximum-sum segment centroid, normalized sum of absolute spectral slopes, and four spectral energy parameters, showed a better association with the place of articulation than the earlier reported spectral parameters.

(ii) The errors using the ANN-based speaker-independent estimation of the place of articulation from the proposed set of spectral parameters obtained from the frication segments for the alveolars and the palatals were comparable to the standard deviation of the reference values. The errors for the labiodentals were much smaller than their distance from the alveolars. Therefore, this technique may be usable in speech-training aids for visual feedback to improve articulation of the fricatives. The estimation error decreased with the increase in training data size, indicating scope for improving the estimation by increasing the training data size.

The conclusions from the investigations on the effect of spectral parameters during the vocalic transitions may be summarized as follows.

(i) Perceptual study on the relative importance of the vocalic transition and the frication indicated the fricative place perception to be determined by a combination of the vocalic transition and frication segments, with the labiodental perception dominated by the transition.

(ii) Inclusion of the spectral parameters obtained from the vocalic transition in the input parameter set for the ANN-based place estimation can be used to reduce the estimation errors.

(iii) The proposed parameter set is better than MFCCs for representing the frication and slightly inferior to MFCCs for representing the vocalic transition.

The use of the set of proposed spectral parameters avoids the difficulties associated with formant tracking during the transition and frication segments. Therefore, it may be considered suitable for ANN-based speaker-independent estimation for feedback of place of articulation of the fricatives.

## 5.4 Suggestions for further research

The reference place of articulation for the ANN-based estimation was obtained using the four pellet points on the tongue in the XRMB database. This sparsity of information of the tongue

shape affected the estimation of the reference place of articulation, and it may have significantly contributed to the errors in the ANN-based estimation of the place of articulation, particularly for the palatals. The estimation errors may be reduced by obtaining the reference place from the tongue outline acquired using MRI or ultrasound imaging. Large errors for the voiced palatal /ʒ/ may have been due to its small number of utterances in the database. An investigation may be carried out using a database with a better balance of the fricatives. An investigation may also be carried out for the linguadentals /θ/ and /ð/ after removing the instances of their utterances with stop-like characteristics. The proposed method needs to be evaluated by applying it to children's speech for place estimation. In the absence of an acoustic-articulatory database, this evaluation may be carried out using a subjective assessment of the utterances and spectrograms.

# Appendix A
# VISUAL SPEECH-TRAINING AIDS

## A.1 Introduction

Speech development in children with hearing impairment is disrupted due to the lack of auditory feedback. They experience difficulty in acquiring the ability to control the movement of various articulators involved in speech production. Thus there is a need for speech-training aids providing appropriate non-auditory feedback to help the speech correction process in children with hearing impairment. Speech therapists commonly use a mirror to provide visual feedback about the lip movements for speech training of children with hearing impairment. However, the mirror does not provide feedback for the actions inside the oral cavity that are not visible from the outside. Computer-based speech-training aids providing a dynamic display of important acoustic parameters (such as speech level, voicing, pitch, and spectral features) and articulatory parameters (such as position and movements of lips, jaw, tongue) have been found to be useful for speech training of children with hearing impairment (Nickerson and Stevens 1973, Mahshie *et al.* 1988, Adams *et al.* 1989, Zahorian and Venkat 1990, Tiger DRS 1999, Carey 2004, Olson 2014, Micro Video Corp. 2017, Purr Programming 2017, Speechtools Ltd. 2019, DevExtras 2020, Crichton and Fallside 1974, Fletcher 1982, Pardo 1982, Black 1988, Dagenais *et al.* 1994, Massaro and Light 2004, Engwall *et al.* 2006, Martin *et al.* 2007, Bernhardt *et al.* 2008, Mahdi 2008, Bacsfalvi and Bernhardt 2011, Pickett 2013, Wilson 2014). Several speech-training systems have been developed using automatic speech recognition for feedback about speech production (Kewley-Port *et al.* 1991, Vicsi *et al.* 2000, Ahmed *et al.* 2018). A review of speech-training systems providing visual feedback using the acoustic parameters, automatic speech recognition, and articulatory parameters is presented in the following sections.

## A.2 Visual feedback of acoustic parameters

Speech signal production involves filtering a broad-band excitation signal by the vocal tract with the transfer function determined by its shape. The time-varying vocal tract shape is controlled by movement of the articulators. The acoustic characteristics of the speech signal may be useful in getting information about the articulatory movements and providing visual feedback about the articulation.

Nickerson and Stevens (1973) developed a computer-based system for speech training of persons with hearing impairment, providing visual feedback of level, pitch, voicing, and

spectral distribution. The speech signal was acquired using a head-mounted microphone, and an accelerometer attached to the throat was used to estimate the pitch. Spectral distribution was obtained using a bank of 19 bandpass filters over 80−6500 Hz. The level and pitch were displayed by varying diameter and vertical position of a circle, respectively, and plotted as time functions. For voicing feedback, a display with pitch trace as the center line and level as an envelope around it was used. For unvoiced sounds, the pitch trace was dropped to a baseline with a level envelope around it. The spectral distribution was displayed vertically as a symmetrical object by reflecting the distribution along the vertical axis. It was used for training of articulation of sustained vowels and fricatives. They reported that the system could be used for faster progress in some aspects of acquiring speech by children with hearing impairment.

Mahshie *et al.* (1988) reported a computer-based system with interactive games to train sustained voicing, repeated vocalization, level, and pitch control. The level information was estimated from the speech signal, and the pitch was extracted from the electroglottograph (EGG) signal. The initiation and termination of the phonation were detected using average oral airflow measured using a pneumotachograph. The system was used for speech training of 14 children with pre-lingual profound hearing loss for 15 months. It was reported that the system was easy to use, and it reduced the required training time.

Zahorian and Venkat (1990) developed a computer-based speech-training system for vowel articulation by hearing-impaired persons. An ANN was trained to transform a 16-channel filter bank outputs to the first and second formants. The system provided a real-time display showing a filled rectangle with its size proportional to the level and its position corresponding to the formants, with an unfilled circle marking the target vowel location. They reported that informal testing with hearing-impaired children of 4−10 years showed the display to improve vowel articulation.

Several commercially available speech-training systems displaying several acoustic parameters, such as pitch, level, and spectral features, have been developed for improving the articulation of persons with hearing impairment (Adams *et al.* 1989, Tiger DRS 1999, Micro Video Corp. 2017, Purr Programming 2017, Speechtools Ltd. 2019, DevExtras 2020). The system ‘Speech Viewer’, developed by IBM, has been widely used for speech training in many institutions across several countries (Adams *et al.* 1989). In this system, level, pitch, zero-crossings, and LPC-spectrum are calculated and used to display the features grouped into three models, named as awareness, skill-building, and pattern matching. The awareness module displays level as the size of a balloon, pitch as the mercury level of a thermometer, and voicing onset as a moving train. The skill-building module uses game-like exercises to

help in acquiring precise control of acoustic parameters. The pattern module provides a dynamic display of acoustic parameters to help in matching a target sequence. The system 'Video Voice Training System' from Micro Video Corporation (2017) provides a display of pitch, level, and formants, using interactive games to motivate the children to continue practicing the control of acoustic parameters. The pitch and level parameters are displayed as a function of time. The formants can be displayed as F1-F2 plot or as functions of time. The formant transitions during an utterance can be viewed as a connected graph on F1-F2 plot. The display can be used for vowel and duration training. It has a provision for overlapping target plots on the parameter plots.

The system 'Dr. Speech', developed by Tiger DRC Inc. (1999), provides voice-activated interactive games with a display of acoustic parameters, with modules for improving speech, language, hearing, and vocabulary skills. The speech therapy module provides games with real-time display of the level, voicing, pitch, and voice onset parameters and feedback on insufficient respiration during speech production. The real speech module provides a real-time display of pitch, jitter, shimmer, spectrogram, and formants, with the pitch, jitter, shimmer calculated from the EGG signal. It can be used to track formant locations on the F1-F2 plot with a marking of the target locations. The pitch master module provides a real-time pitch display with a target pitch template for prosody training. The vocal assessment module displays the acoustic parameters and the parameters calculated from the EGG signal for voice quality assessment, including jitter, shimmer, harmonic-to-noise ratio, normalized noise energy, contact quotient, contact index, open quotient, closed quotient, contact quotient perturbation, and contact index perturbation. The scope view module displays the glottis video using a laryngoscope, micro-laryngoscope, or a video camera, with a simultaneous display of the EGG signal. The nasal view module provides a real-time display of the nasal flow for detecting hyper-nasality or hypo-nasality.

With the increase in the mobile handsets' processing power, many mobile-based apps have been developed for speech therapy applications. The 'Voice Analyst' app from Speechtools Ltd. (2019) provides a real-time level and pitch display and statistical measures, with the facility of displaying a target. Similar features are available in the 'Voice Tools' app from DevExtras (2020). The 'Voice Pitch Analyzer' app from Purr Programming (2017) provides a real-time plot of the pitch with its average, minimum, and maximum values, and a pitch-based classification of the utterance.

## A.3 Speech-training aids for visual feedback using speech recognition

Automatic speech recognition is the process of deriving the word sequence from the speech signal. Several approaches, including spectral distance measure, template matching, hidden Markov model, etc. are used for speech recognition. This technology can be used for measuring articulation quality or correctness. Several speech-training systems have been developed using speech recognition for feedback for improving speech production (Kewley-Port *et al.* 1991, Vicsi *et al.* 2000, Ahmed *et al.* 2018). A review of some of these systems is presented in this section.

The system 'Indiana Speech Training Aid (ISTRA)' developed by Kewley-Port *et al.* (1991) used a speaker-dependent speech recognizer to provide non-auditory feedback for speech training. In this system, the utterances by each speaker are used to form templates. A 0−99 goodness-of-fit score is computed for the speaker's utterance using the stored templates. The basis of the speech training was that the children with hearing impairment produce utterances with good quality 5−10 % of the time but produce the same utterances with poor quality for the remaining time. Thus, using the goodness-of-fit score, the children can practice on their own. The system was used to train using words, phrases, and isolated phonemes to correct the errors related to duration, omission, insertion, and substitution. Case studies involving speech training of a person with hearing impairment, a person with profound hearing loss, and a person with normal hearing but misarticulation showed significant improvements.

The system 'Box of Tricks', developed by Vicsi *et al.* (2000), provides a real-time display of acoustic parameters to assist children in learning to speak. In this system, the phonemes are associated with symbolic pictures, and amusing drawings are used to display level, pitch, spectrum, and voicing. The system uses acoustic parameters from sounds spoken by a child with a clear and good pronunciation as reference and a database of sounds spoken by 72 children in the 5−10 years age group to obtain the maximum and minimum limits of the acoustic parameters. The spectral distances are used for feedback on the pronunciation quality. The system can be used for training isolated phonemes and words involving vowels and fricatives. A picture of a face with articulators (lips, tongue, teeth, etc.) at appropriate positions for producing a phoneme is also displayed before the articulation training. It was reported that the system resulted in a consistent reduction in the time required for articulation training of children of 5−10 years and with different hearing-impairment levels.

Ahmed *et al.* (2018) integrated the 'PocketSphinx' based speech recognition, with a 150-word dictionary, into five popular mobile games ('WordPop' and 'SpeechWorm' on iOS; 'Asteroids', 'Whack-A-Mole', and 'Memory' on Android) with attractive graphics interfaces

and examined their effectiveness for speech therapy of children with apraxia and typically developing children. In WordPop, the child touches a word that appears on the screen and utters it. One point is added for each letter in the word if the word is produced correctly, and the word can be uttered again if the pronunciation is incorrect. SpeechWorm displays a grid of letters with a set of words embedded into it. The child finds the word, swipes the fingers across the letters in it, and utters the word. Points are added for every correctly produced word, and the child can attempt again if recognition fails. Whack-A-Mole displays two rows of moles that randomly flip over one at a time and show a picture corresponding to a word. The child taps the flipped mole to stop it from flipping back and utters the word. A star is added at the bottom if the word is correctly recognized, and the mole flips back otherwise. In Asteroids game, the goal is to save lives on a spaceship by breaking asteroids before asteroids hit the spaceship. The child has to touch the asteroid and utter the word displayed on the screen to break the asteroid. Recording and speech recognition engine are started when the child touches an asteroid on the screen. If the word is correctly produced, the asteroid breaks into pieces. Memory has ten images behind bubbles. The child touches a bubble to uncover the image and utters the word corresponding to it. Good and fair productions are acknowledged by golden and silver stars, respectively. Evaluation involved ten children with mild to severe apraxia of speech and the age of 6–11 years and six typically developing children with the age of 7–11 years. It indicated that the children preferred games over traditional speech therapy exercises, were interested in playing games with challenging tasks and multiple difficulty levels, and expressed frustration when the uttered words were not recognized. The performance of automatic speech recognition, assessed by listeners, was 56% for children with apraxia and 52% for typically developing children, indicating a need for improvement for it to be effective for speech training.

The speech-training systems based on speech recognition have been shown to be useful for children with hearing and speech difficulties by encouraging practice at home. As the children become uncomfortable when the system fails to recognize a correctly produced sound, the system's usefulness depends on the reliability and accuracy of its speech recognition engine. Another limitation of these systems is that they can only provide feedback on the articulation correctness and not on how to correct it.

## A.4 Visual feedback of vocal tract shape

Studies on the articulation errors in the hearing-impaired children’s speech have reported the bilabial consonants to be more intelligible than the lingual consonants and the vowels. The findings indicate that these children move their articulators appropriately for visible actions

but fail to coordinate them for actions inside the oral cavity (Huntington *et al.* 1968, Smith 1975). Speech therapists instruct the children to correctly position the articulators (lips, tongue, and jaw) by showing the articulatory actions for producing various sounds. They use a mirror for visual feedback of the child's actions during sound production. However, it is difficult to show and provide feedback for the actions inside the oral cavity that are not visible from the outside. Several speech-training systems have been developed for visual feedback on the actions inside the oral cavity. These systems estimate the articulators' positions and movements, either directly using imaging techniques or indirectly from the speech signal.

Fletcher (1982) developed a speech-training system called dynamic orometer to provide visual feedback of articulatory movements and acoustic parameters. The orometer consisted of a glossometer for the tongue's position and movements, a palatometer for the tongue's pattern of contact with upper palate and teeth, a gnathometer for the position and movements of lips and jaw, and a vocometer for the spectrum, fundamental frequency, and level of the speech. The glossometer consists of eight pairs of narrow-beam LEDs and phototransistors along the midline of a pseudopalate. The LED light reflected from the tongue surface and reaching the phototransistor is converted to a voltage inversely related to the distance of the tongue surface. The palatometer includes 96 metal electrodes embedded on the pseudopalate, and gnathometer has a camera detecting the LEDs attached on the lips and jaw. The vocometer obtains the spectrum using a 32-channel filter bank over 200–5000 Hz and obtains the level and pitch using a piezoelectric accelerometer positioned against the neck below the larynx. The training system provided a side-by-side display of the tongue positions and the linguapalatal contact patterns of the child and the teacher. Results from training of young children indicated that the system helped improve vowel and consonant articulation. It was inferred that glossometric training showing the tongue's shape and position was essential for long-lasting improvement of vowel articulation.

Dagenais *et al.* (1994) studied the effect of electropalatograph (EPG) based speech training on 18 children with hearing impairment and age of 10–15 years, by dividing them into two groups with one group taught using traditional techniques and the other group using EPG. The training consisted of articulation training of /t, d, k, g, s, z, ʃ/ in CV syllables. The training effectiveness was evaluated by examining the linguapalatal contact patterns and perceptual evaluation, before training, after training completion, and 6 months later. Both groups' articulation showed improvement for all the consonants, with more improvement for the EPG group. For alveolar stops, the EPG group had better listener identification scores. For velar stops, the scores six months after training were similar for both groups. The alveolar and palatal fricatives were confused during all the testing conditions. Other studies have also

reported EPG-based training to improve articulation (Martin *et al.* 2007, Bacsfalvi and Bernhardt 2011, Pickett 2013).

With advances in ultrasound technology, portable ultrasound devices have been used for imaging the tongue surface (Gick 2002). To evaluate the effectiveness of visual feedback using ultrasound imaging in speech therapy, Bernhardt *et al.* (2008) studied the progress of 13 children of 7−15 years and having difficulty articulating /r/. Speech therapy was carried out in three phases, with 7−8 non-ultrasound sessions, 1−3 ultrasound sessions, and 7−8 non-ultrasound sessions. Listening tests showed a significant improvement in the /r/ articulation for 11 children after the ultrasound sessions. In a study by Bacsfalvi and Bernhardt (2011) involving ultrasound and EPG-based speech therapy on seven hearing-impaired adolescents, improvements in the articulation of vowels, fricatives, and /r/ were maintained even after 2−4 years after the therapy. A limitation of the ultrasound imaging is that it cannot be used to obtain the position of the tongue tip, tongue root, and lips. Further, it is not suitable for speech therapy on a wide scale as it is time-consuming, tedious, and requires specialist training to obtain the tongue contour from the images (Eshky *et al.* 2018).

Massaro and Light (2004) developed a talking head animation, named 'Baldi', as a speech-training aid. The animation had a tongue, hard palate, and 3D teeth, with the articulatory movements based on the data obtained from the EPG and ultrasound imaging during speech production. The display had multiple head views, including a sagittal view with tongue and teeth alone, a side view with transparent skin, and a front view of the face with transparent skin. Virtual vocal cord vibration indicated voicing, and turbulent airflow from the mouth indicated frication. For training, Baldi provided slow-motion animation with audio and instructions for positioning the tongue with respect to the teeth, tongue, and lips. The evaluation involved 21-week training of seven hearing-impaired children of 8–13 years for voicing, consonant production, and affricate-fricative distinction. A 21% improvement in recognition was reported. A significant limitation of this aid is that it does not provide the corrective feedback relating the speaker's production with the target production.

Crichton and Fallside (1974) developed a computer-based speech-training aid for improving the articulation of sustained vowels. In this aid, the vocal tract area function is obtained from the windowed speech signal using LP-based inverse filtering (Wakita 1973), and it is smoothed using parabolic interpolation. The area values are displayed as a function of distance from the glottis, with a target function superimposed for correcting the articulation. Evaluation of the aid involved its use by hearing-impaired children of 6–9 years for nine months, and it was found useful for improving vowel articulation. In the speech-training aid developed by Pardo (1982), the area ratios obtained from Wakita's method are

normalized by the vocal tract volume calculated as the sum of the area ratios. This normalization reduced the dispersion of the area values as compared with that assuming unity area at the glottis. It was reported that the aid helped in improving vowel articulation, and the errors with respect to the target shapes decreased progressively with training.

Black (1988) reported a speech-training aid displaying the tongue shape in the head's mid-sagittal view during vowel production in real-time. Filter-bank analysis of the speech signal followed by a peak-picking algorithm was used to obtain the formant frequencies. From these frequencies, the tongue shape was obtained using an empirical relationship between the formant frequencies and the degree of back and front raising of the tongue, as proposed by Ladefoged *et al.* (1978). The display also showed a target tongue shape. It was reported that the articulatory information displayed by the aid helped improve the accuracy and consistency of vowel production.

Park *et al.* (1994) developed a speech-training aid displaying vocal tract shape and acoustic parameters in real-time. The vocal tract shape estimated using Wakita's LP-based method was displayed on a mid-sagittal view, with the upper part of the oral cavity fixed and the lower part movable. For reducing the error due to the assumption of a fixed vocal tract length, the heights of two sections from lips to teeth were obtained from the first three formant frequencies using the empirical relation by Ladefoged *et al.*(1978). The pitch and nasality were estimated using vibration sensors at the throat and nose, respectively. The speech level was calculated as the log energy of the speech signal, and the spectrum was obtained using the autoregressive spectral estimation method. The aid also displayed the target shapes. The evaluation involving two hearing-impaired children of 12–13 years showed that they could master articulation of the syllables /ja/ and /pa/ in 5-6 days.

Engwall *et al.* (2006) reported a speech-training aid, named 'ARticulation TUtoR (ARTUR)', to provide 3D animation of the face and the articulators for corrective feedback on the speaker's articulation with reference to a target articulation. The 3D animations were based on a vocal tract model consisting of 3D-mesh structures of tongue, jaw, palate, and vocal tract walls. The articulatory parameters controlling the tongue model were obtained from the MRI data of a reference Swedish speaker holding the articulatory configuration during 13 Swedish vowels and 10 consonants in VCV context. Information on articulatory kinematics was obtained using the same speaker's real-time MRI, EMA, and EPG measurements. The model could be adapted to a new user using the mid-sagittal MRI images. A Wizard of Oz study (an experiment with the subject interacting with a human through a computer interface without knowing that the responses are from a human rather than a computer) study was conducted to test and revise the interface for speech training. It involved

three children of 9–14 years and three children of 6 years, with language disorders without hearing difficulties. A phonetically trained person, sitting in a room not visible to the child under training, detected the mispronunciation and performed the articulatory inversion to provide the corrective feedback. The results showed the feedback to be helpful in correcting the articulation, easy to use, and helpful in practicing alone. However, the younger children found it difficult to imitate the pronunciation.

Mahdi (2008) reported a speech-training aid displaying the vocal tract in the mid-sagittal view, three formants marked on the spectrum, pitch, and level. The vocal tract area function was estimated using LP analysis, with the vocal tract modeled as a 17-cm acoustic tube with 18 sections. For reducing the error due to fixed vocal tract length assumption, the first two section areas were computed using the first three formant frequencies using the method proposed by Ladefoged et al.(1978). The upper jaw was assumed to be fixed, and the lower jaw was movable. The area values were mapped to the graphics using a reference grid with 18 sections obtained from the X-ray data. It was reported that the vocal tract area functions for 10 American English vowels correlated well with reference data from X-ray imaging. The mean squared error between the estimated and reference shapes was lowest for the front vowels and relatively larger for the back vowels.

Jain *et al.* (2016) developed a speech-training aid with a dynamic display of vocal tract shape obtained from the speech signal using Wakita's LP-based inverse filtering method. The display had separate panels for displaying the vocal tract shapes of the learner and a reference speaker for facilitating a visual comparison of the articulatory efforts. A variable-rate animation of the vocal tract shape was used to provide the articulatory feedback. The system also displayed pitch and level for speech training. The speech waveform, spectrogram, and areagrams were displayed for validation by the speech therapist or the teacher. For emphasizing the position of maximum constriction, the tongue was displayed as a triangle with the endpoints of the tongue forming the left and right vertices and the position of maximum constriction indicated as the middle vertex of the triangle. Vocal tract shapes displayed for three vowel utterances from a male speaker showed a good match with the MRI scan images for the corresponding vowels obtained from the MRI database (Narayanan *et al.* 2014).

A review of speech-training aids has shown that the aids providing visual feedback of the articulatory efforts not visible from outside are useful for improving articulation of hearing-impaired persons. The aids using direct imaging to obtain positions and movements of the articulators are time-consuming and expensive, and they interfere with speech production. The aids estimating the articulatory information from the speech signal do not interfere with

speech production and are convenient for wide-scale use in speech training. Most of these aids are based on visual feedback of the vocal tract shape estimated using LP analysis and are suitable for vowel articulation. For improving the effectiveness of speech-training aids, it is important to investigate techniques for vocal-tract shape estimation during consonants.

# Appendix B
# VOCAL TRACT LENGTH NORMALIZATION FOR ESTIMATION OF PLACE OF ARTICULATION

## B.1 Introduction

Machine-learning methods for acoustic-to-articulatory mapping are reported to work well for speaker-dependent mapping (Hiroya and Honda 2004, Toda *et al.* 2008, Richmond 2006, Ji *et al.* 2014b, Liu *et al.* 2015, Ji *et al.* 2016, Illa and Ghosh 2018). However, these methods perform poorly for acoustic data from an unseen speaker. The speaker's vocal tract shape and vocal tract length affect the spectral characteristics and may introduce errors in the mapping. Vocal tract length normalization (VTLN) based on frequency warping is often employed to reduce the inter-speaker variability in ASR (Lee and Rose 1998, Jaitly and Hinton 2013, Serizel and Giuliani 2014). VTLN based on a linear frequency warping function is known as conventional VTLN (Lee and Rose 1998, Jaitly and Hinton 2013, Serizel and Giuliani 2014). The warping factor (slope of the linear warping function) is estimated using a grid search based on maximum-likelihood estimation. In VTLN using a nonlinear warping function, the warping factor varies with frequency to obtain an improved model of the inter-speaker variability (Eide and Gish 1996, Kumar and Umesh 2008, Arsikere *et al.* 2013). Even though nonlinear warping is considered a better approximation for VTLN, linear warping is more commonly employed (Serizel and Giuliani 2017, Shahnawazuddin *et al.* 2018).

Sivaraman *et al.* (2019) investigated the VTLN technique based on frequency warping to improve the speaker-independent acoustic-to-articulatory mapping using acoustic and articulatory data from the XRMB database. In this investigation, multiple speakers' acoustic data were normalized towards a target speaker's acoustic space using a three-segment piecewise linear frequency warping function as in the HTK toolkit (Young *et al.* 2006). The optimal warping factors were obtained by a maximum likelihood estimation using a 64-component GMM fitted on a MFCC feature set comprising 13 coefficients, 13 slopes, and 13 accelerations for all the analysis frames (frame length = 20 ms, shift = 10 ms) of all the utterances without any contextual information. For the ANN-based estimation of articulatory parameters, 13 MFCCs of a frame concatenated with the coefficients for eight frames on either side of the frame were used as the input feature vector. The articulatory parameters were lip aperture, lip protrusion, tongue body constriction location, tongue body constriction degree, tongue tip constriction location, and tongue tip constriction degree. These six parameters were obtained from the pellets' *x-y* locations using geometric transformations

(McGowan 1994, Mitra *et al.* 2012, Nam *et al.* 2012). The hard palate was approximated as a circular arc through the points on the palate trace, and the tongue body was approximated as a circular arc passing through the pellets T2, T3, and T4. The tongue tip was approximated as a line segment between the pellets T1 and T2. The lip aperture was measured as the distance between the pellets UL and LL. The lip protrusion was measured as the horizontal displacement of the pellet LL along the horizontal axis measured from the median of its position across utterances. The tongue body and tip constriction locations were measured as the angular displacements from 0º at the chin to the point of maximum constriction on the tongue body and tip, respectively, with respect to a reference line connecting the floor of the mouth to the center of the hard palate. The tongue body constriction degree and the tongue tip constriction degree were measured as the distance between the palate trace and the point of maximum constriction on the tongue body and the tongue tip, respectively. Speaker-independent mapping was evaluated using data from 46 speakers, with the data from 35, 5, and 5 speakers for training, validation, and testing, respectively. Results showed that the averaged correlation coefficient between the estimated and actual vocal tract parameters improved from 0.782 for the non-normalized acoustic parameters to 0.791 for the normalized acoustic parameters. The evaluation was also carried out using data from randomly selected ten speakers, with one mapping obtained for each of the ten speakers using data from nine speakers for training and 10% of the data from the target speaker for testing. Results showed that the averaged correlation coefficient between the estimated and actual vocal tract parameters improved from 0.692 to 0.793 by using the normalized acoustic parameters, thus indicating that the normalization provides significant improvement for the training data with a small number of speakers. It may be noted that the tongue body and tip constriction locations as used in this investigation do not provide the place of articulation as used in our investigation reported in Chapter 3.

In Chapter 3, an ANN-based speaker-independent mapping was used to estimate the place of articulation from the input spectral parameters. The utterances with the same place of articulation from two speakers may have different spectral characteristics due to the difference in vocal tract lengths. To compensate for such differences, a vocal tract length normalization based on frequency warping and similar to that used by Sivaraman *et al.* (2019) is investigated, by normalizing multiple speakers' acoustic data towards a target speaker's acoustic space. The normalized spectral parameters are used for the ANN-based estimation of the place of articulation. The normalization method is described in the next section, followed by the results of the ANN-based estimation in the third section.

## B.2 Method

For VTLN, a two-segment piecewise linear frequency warping function as shown in Figure B.1, was used. The relationship between the frequency $f(k)$ at frequency index $k$ and the corresponding warped frequency $f_W(k,\alpha)$ with the warping factor $\alpha$ is given as

$$f_W(k,\alpha)=\begin{cases}\alpha f(k), & f(k)\le(2f_u/(1+\alpha)\\ \frac{f_{\max}-\alpha f_u}{f_{\max}-f_u}(f(k)-f_u)+2\alpha f_u/(1+\alpha), & f(k)>(2f_u/(1+\alpha)\end{cases} \quad \text{(B.1)}$$

where $f_{\max}$ is the highest frequency and $f_u$ is the break-point of the warping function. The function uses $f_{\max}$ set as $f_s/2$ for sampling frequency of $f_s$ and $f_u$ set as 3.20 kHz. This warping function is similar to the three-segment warping function used by Sivaraman *et al.* (2019) and defined using a lower frequency $f_l$ and an upper frequency $f_u$, with $f_l$ set as 0 Hz instead of 60 Hz. A single warping factor $\alpha$ is used for all utterances from a speaker. The warped spectrum $S_W(k)$ is obtained from the unwarped spectrum $S_{UW}(k)$ using the relation

$$S_W(k)=S_{UW}(g(k,\alpha)) \quad \text{(B.2)}$$

where $g(k,\alpha)$ is the warped frequency index. It is obtained from the warped frequency $f_W(k,\alpha)$ as

$$g(k,\alpha)=\lfloor f_W(k,\alpha)L/f_s+0.5\rfloor \quad \text{(B.3)}$$

where $L$ is the FFT size.

The warping factor for normalizing the acoustic data of a speaker towards the target speaker is determined using a maximum likelihood approach. The probability density of the acoustic parameters obtained from a target speaker is modeled using the GMMs with 64 Gaussian components and diagonal covariance matrices. The warped acoustic features are obtained for all the acoustic frames without segmentation for each of the speakers for $\alpha$ varying from 0.8 to 1.2 in steps of 0.025. The range and step size for grid search were selected based on the default values of the HTK's implementation and as used by Sivaraman *et al.* (2019). The warping factor $\alpha_r$ for a speaker $r$ can be obtained using the GMM acoustic model for the target speaker ( $\gamma_{ts}$ ), as

$$\alpha_r=\arg\max_{\alpha}\sum_{n=1}^{N}\log(p(\mathbf{x}_r^n|\gamma_{ts},\alpha) \quad \text{(B.4)}$$

where $\mathbf{x}_r^n$ are the warped acoustic features of the acoustic frame $n$ of the speaker $r$ obtained using the warping factor $\alpha$.

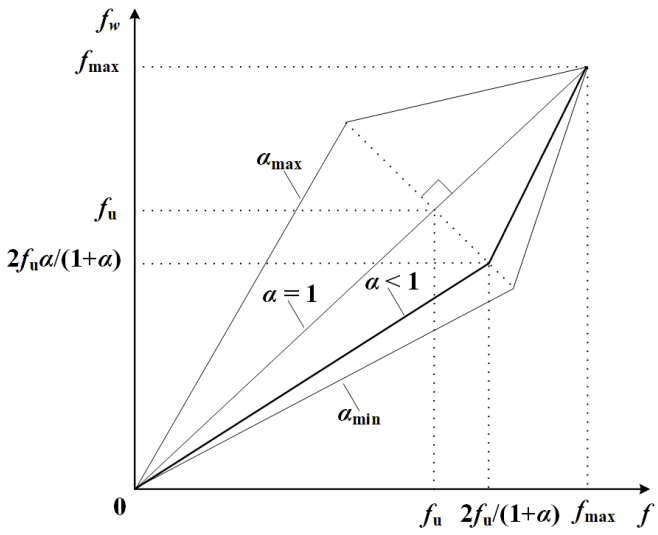



**Figure B.1** Piecewise linear frequency warping function, adapted from HTK toolkit (Young *et al.* 2006), for VTLN.

The investigation used the acoustic data of 46 speakers in the XRMB database. All the audio recordings were down-sampled to 8 kHz. A 39-dimension MFCC feature set, including 13 coefficients, 13 slopes, and 13 accelerations, calculated using a 1024-point FFT of and Hanning window for the 20-ms segments with 10-ms shift, was used as the acoustic parameter set. A male speaker JW12 of the XRMB database was used as the target speaker, and its acoustic model based on GMM was used to obtain the optimum warping factors for each of the other 45 speakers in the database.

The estimated optimal warping factors for male speakers were between 0.900 to 1.075 with the median value of 1.000, and those for female speakers were between 1.025 to 1.200 with the median value of 1.125. The optimum warping factor $\alpha_r$ for each speaker was used to obtain warped magnitude spectra of the fricative segments, with the sampling frequency of 16 kHz, $f_{max}$ of 8 kHz and $f_u$ of 6.8 kHz. The warped spectra were used for calculating the spectral parameters for the ANN-based estimation of the place of articulation. The estimation was carried out using the SPS-6 parameter set (MSSC, NSASS, NHBE, NVLBE, PAE, and MLBE) calculated as described in Section 3.4.1 of Chapter 3. The investigation used a two-layer network with the optimal number of hidden layers and neurons obtained in the investigation presented in Section 3.5 and Section 3.6 of Chapter 3.

**Table B.1**: Mean and standard deviation (S.D.) of correlation coefficients and errors across 5-fold validation sets using networks with two hidden layers for estimation using spectral parameters calculated after VTLN.

| Spectral Parameter Set | Errors with respect to PoA-art (mm) | | | | | | Corr. Coeff. | |
|---|---|---|---|---|---|---|---|---|
| | Mean Error | | S.D. Error | | RMS Error | | | |
| | Mean | S.D. | Mean | S.D. | Mean | S.D. | Mean | S.D. |
| SPS-6 (MSSC, NSASS, NHBE, NVLBE, PAE, | 0.1 | 0.2 | 2.7 | 0.2 | 2.7 | 0.1 | 0.974 | 0.005 |

**Table B.2**: Mean and standard deviation (S.D.) of errors and RMS errors for different fricatives for estimation using spectral parameters calculated after VTLN.

| Spectral Parameter Set | Fricative | Errors with respect to PoA-art (mm) | | | | | |
|---|---|---|---|---|---|---|---|
| | | Mean Error | | S.D. Error | | RMS Error | |
| | | Mean | S.D. | Mean | S.D. | Mean | S.D. |
| SPS-6 (MSSC, NSASS, NHBE, NVLBE, PAE, MLBE) | f | -0.5 | 0.3 | 2.7 | 0.7 | 2.7 | 0.8 |
| | v | 0.0 | 0.2 | 1.9 | 0.7 | 1.9 | 0.7 |
| | s | 0.0 | 0.2 | 2.6 | 0.2 | 2.6 | 0.2 |
| | z | 0.2 | 0.3 | 2.7 | 0.3 | 2.7 | 0.4 |
| | ʃ | 0.4 | 1.0 | 3.0 | 0.4 | 3.2 | 0.6 |
| | ʒ | 1.2 | 1.5 | 3.6 | 1.0 | 3.8 | 1.4 |

## B.3 Results

The mean and the standard deviation of correlation coefficients and errors for the five-fold validation of the ANN-based estimation of the place of articulation using the spectral parameter after VTLN are given in Table B.1. The errors after VTLN are 0.2 mm higher than the corresponding errors in Table 3.3 in Chapter 3. The fricative-wise errors after VTLN, given in Table B.2, are higher for all fricatives than the corresponding errors in Table 3.4 in Chapter 3. Thus, the VTLN did not help in the fricative place estimation.

# Appendix C
# NON-UNIQUENESS IN ESTIMATION OF PLACE OF ARTICULATION FROM THE SPECTRAL PARAMETERS

## C.1 Introduction

The inverse problem of obtaining the articulatory configuration from the acoustic parameters is difficult as different articulatory configurations can produce similar acoustic parameters. Qin and Carreira-Perpiñán (2007) quantized the acoustic and articulatory parameters into clusters, and they proposed that the mapping from acoustic parameters to the articulatory configuration can be identified as non-unique if the articulatory parameters corresponding to the acoustic parameters in a cluster are mapped to multiple clusters. Investigation using LPCs and *x-y* locations of pellets from a male speaker in the XRMB database showed that the acoustic clusters of phonemes /θ, ɹ, w, l/ result in non-unique mapping and those of /ae, u, y/ result in unique mapping. Ananthakrishnan *et al.* (2009) proposed that multiple peaks in GMM-based conditional probability density function of the articulatory parameters for the input acoustic parameter can be used as the non-uniqueness instances. They quantified non-uniqueness as the average spread of the peak locations in the conditional probability density function. Investigations using MFCCs and *x-y* locations of coils from two speakers in the MOCHA-TIMIT database showed that the stops, fricatives, and nasals result in higher non-uniqueness compared to the vowels, liquids, and diphthongs.

The ANN-based estimate of the place of articulation from the input spectral parameters presented in Chapter 3 suffers from non-uniqueness as different values of the place of articulation can result in the speech signal with similar spectral parameters. For quantifying this non-uniqueness, a speaker-independent mapping from the proposed set of spectral parameters to the place of articulation based on a Gaussian mixture model (GMM) was used to estimate the place of articulation (Toda *et al.* 2008). The non-uniqueness was quantified using the conditional probability density function of place of articulation for the input spectral parameter set. The basics of GMM-based estimation of the place of articulation is presented in the next section, the method and results of quantifying the non-uniqueness in the estimation of the place of articulation are presented in Section C.3 and Section C.4, respectively.

## C.2 GMM-based estimation of place of articulation from spectral parameters

The method based on the Gaussian mixture model (GMM) estimates the place of articulation by maximizing the likelihood of conditional probability density of place of articulation for the

input spectral parameter set. This section presents the basics of this method, adapted from Toda *et al.* (2008). In this method, the multivariate joint probability density function of spectral and articulatory parameters is modeled as a weighted sum of Gaussian probability density functions. Let $\mathbf{x}_t = [x_t(1), x_t(2), x_t(3), ..., x_t(L_x)]^T$ be the $L_x$-dimensional spectral feature vector and $\mathbf{y}_t = [y_t(1), y_t(2), y_t(3), ..., y_t(L_y)]^T$ be the $L_y$-dimensional articulatory feature vector at time *t*. For estimating the place of articulation, the articulatory feature vector consists of the place of articulation as the single parameter, i.e., $L_y$ = 1. We represent $\mathbf{q}_t = [\mathbf{x}_t^T, \mathbf{y}_t^T]^T$ as the joint feature vector. The GMM approximation of the joint probability density of spectral and articulatory parameters is expressed as,

$$p(\mathbf{q}_t | \boldsymbol{\theta}^{(q)}) = \sum_{m=1}^{N} \beta_m N(\mathbf{q}_t; \boldsymbol{\eta}_m^{(q)}, \boldsymbol{\Sigma}_m^{(q)}) \tag{C.1}$$

where $\beta_m$, $\boldsymbol{\eta}_m^{(q)}$, and $\boldsymbol{\Sigma}_m^{(q)}$ are the weight, the mean vector, and the covariance matrix of the *m*th component, respectively. The GMM parameter set $\boldsymbol{\theta}^{(q)}$ consists of mean vectors, weights, and covariance matrices for component density functions. The set $\boldsymbol{\theta}^{(q)}$ is obtained by training the GMM model using the expectation-maximization (EM) algorithm on the training set of the joint vectors. The initial mean vector for the EM algorithm is obtained using the k-means algorithm. The initial covariance matrices are obtained using covariance of the points associated with the corresponding mean vectors. The mean vector $\boldsymbol{\eta}_m^{(q)}$ and covariance matrix $\boldsymbol{\Sigma}_m^{(q)}$ obtained after training can be split using the properties of the Gaussian multivariate distribution as the following:

$$\boldsymbol{\eta}_m^{(q)} = \begin{bmatrix} \boldsymbol{\eta}_m^{(x)} \\ \boldsymbol{\eta}_m^{(y)} \end{bmatrix} \tag{C.2}$$

$$\boldsymbol{\Sigma}_m^{(q)} = \begin{bmatrix} \boldsymbol{\Sigma}_m^{(xx)} & \boldsymbol{\Sigma}_m^{(xy)} \\ \boldsymbol{\Sigma}_m^{(yx)} & \boldsymbol{\Sigma}_m^{(yy)} \end{bmatrix} \tag{C.3}$$

In the above equations, the vectors $\boldsymbol{\eta}_m^{(x)}$ and $\boldsymbol{\eta}_m^{(y)}$ are the mean vectors for the spectral and articulatory density functions, respectively. The matrices $\boldsymbol{\Sigma}_m^{(xx)}$ and $\boldsymbol{\Sigma}_m^{(yy)}$ are the covariance matrices for the spectral and articulatory density functions, respectively. The matrix $\boldsymbol{\Sigma}_m^{(xy)}$ is the cross-covariance matrix between the acoustic and articulatory parameters. The matrices $\boldsymbol{\Sigma}_m^{(xy)}$ and $\boldsymbol{\Sigma}_m^{(yx)}$ are equal. All the covariance matrices are modeled as full covariance matrices.

The conditional density function of $\mathbf{y}_t$, for given $\mathbf{x}_t$, can also be expressed as a mixture of Gaussian density functions as

$$p(\mathbf{y}_t|\mathbf{x}_t,\boldsymbol{\theta}^{(q)}) = \sum_{m=1}^{N} p(m|\mathbf{x}_t,\boldsymbol{\theta}^{(q)}) p(\mathbf{y}_t|\mathbf{x}_t,m,\boldsymbol{\theta}^{(q)}) \tag{C.4}$$

where

$$p(m|\mathbf{x}_t,\boldsymbol{\theta}^{(q)}) = \frac{\beta_m N(\mathbf{x}_t;\boldsymbol{\eta}_m^{(x)},\boldsymbol{\Sigma}_m^{(xx)})}{\sum_{n=1}^{N}\beta_m N(\mathbf{x}_t;\boldsymbol{\eta}_n^{(x)},\boldsymbol{\Sigma}_n^{(xx)})}, \quad p(\mathbf{y}_t|\mathbf{x}_t,m,\boldsymbol{\theta}^{(q)}) = N(\mathbf{y}_t;\mathbf{E}_{m,t}^{(y)},\mathbf{D}_m^{(y)})$$

The mean vector $\mathbf{E}_{m,t}^{(q)}$ and the covariance matrix $\mathbf{D}_{m,t}^{(q)}$ of the conditional density function can be obtained using the parameters of the joint density functions as the following:

$$\mathbf{E}_{m,t}^{(y)} = \boldsymbol{\eta}_m^{(y)} + \boldsymbol{\Sigma}_m^{(yx)}\boldsymbol{\Sigma}_m^{(xx)^{-1}}(\mathbf{x}_t - \boldsymbol{\eta}_m^{(x)}) \tag{C.5}$$

$$\mathbf{D}_m^{(y)} = \boldsymbol{\Sigma}_m^{(yy)} - \boldsymbol{\Sigma}_m^{(yx)}\boldsymbol{\Sigma}_m^{(xx)^{-1}}\boldsymbol{\Sigma}_m^{(xy)} \tag{C.6}$$

The maximum likelihood estimator of the articulatory parameters for the spectral parameters $\mathbf{x}_t$ and the GMM parameter set $\boldsymbol{\theta}^{(q)}$ is expressed as

$$\hat{\mathbf{y}}_t = \arg\max_{\mathbf{y}_t} p(\mathbf{y}_t|\mathbf{x}_t,\boldsymbol{\theta}^{(q)}) \tag{C.7}$$

The EM algorithm is used to maximize the above likelihood function. The estimated articulatory feature vector that maximizes the likelihood is given as

$$\hat{\mathbf{y}}_t = \overline{(\mathbf{D}_t^{(y)^{-1}})}^{-1}\,\overline{(\mathbf{D}_t^{(y)^{-1}}\mathbf{E}_t^{(y)})} \tag{C.8}$$

where

$$\overline{(\mathbf{D}_t^{(y)^{-1}})} = \sum_{m=1}^{M}\lambda_{m,t}^{q}\mathbf{D}_m^{(y)^{-1}}, \quad \overline{(\mathbf{D}_t^{(y)^{-1}}\mathbf{E}_t^{(y)})} = \sum_{m=1}^{M}\lambda_{m,t}^{q}\mathbf{D}_m^{(y)^{-1}}\mathbf{E}_{m,t}^{(y)}, \quad \lambda_{m,t}^{q} = p(m|\mathbf{x}_t,\mathbf{y}_t,\boldsymbol{\theta}^{(q)})$$

The articulatory feature vector obtained by the minimum mean square error estimate is used as the initial vector for the articulatory feature vector $\mathbf{y}_t$ for the EM algorithm.

## C.3 Non-uniqueness in estimating place of articulation from the spectral parameters

The non-uniqueness in estimating place of articulation was studied using the five-fold cross-validation as in the investigation presented in Sections 3.4 and 3.5 of the third chapter. The estimation errors were calculated for parameter set SPS-6 (Parameters MSSC, NSASS, NHBE, NVLBE, PAE, and MLBE) and the reference place of articulation (PoA-art). A joint probability density of the spectral parameters and PoA-art was modeled using a mixture of 20

Gaussians with full covariance matrices. A joint vector $q$ for each utterance in the training data was obtained by concatenating the corresponding spectral parameters and the reference place of articulation PoA-art. The GMM parameter set $\boldsymbol{\theta}^{(q)}$ consisted of mean vector, covariance matrices, and Gaussian component weights of the joint probability density function of these joint vectors. The parameters were estimated using the expectation-maximization algorithm. The place of articulation $a$ was estimated for the utterance with a given spectral parameter set **r,** by maximizing the likelihood of the 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

$$\hat{a} = \arg\max_{a} p(a|\mathbf{r},\boldsymbol{\theta}^{(q)}) \tag{C.9}$$

This likelihood function was maximized using the expectation-maximization algorithm.

The occurrence of more than one peak in the conditional probability density function $p(a|\mathbf{r},\boldsymbol{\theta}^{(q)})$ indicates that there are more than one place of articulation values that result in the given input feature vector and may be considered non-uniqueness instances. The maximum likelihood estimation estimates the place of articulation as the value corresponding to the peak location with maximum probability in the conditional probability density function. More than one peak in the conditional probability density function does not always lead to error as the peak location with the maximum probability may correspond to the PoA-art value. However, if the PoA-art value corresponds to the peak with lower probability, then the error is largely contributed by the non-uniqueness.

For estimating the proportion of total errors caused due to the non-uniqueness, peaks in the conditional probability density function were obtained using a mode-finding algorithm for mixtures of Gaussian distributions proposed by Carreira-Perpiñán (2000) using the MATLAB function provided by Carreira-Perpiñán (2006). The error related to non-uniqueness is calculated as the difference between the estimated place of articulation value and the peak location nearest to the PoA-art value. For only one peak in the probability density function, the non-uniqueness error is minimum as the peak location obtained by the maximum likelihood estimation and the peak location nearest to the PoA-art value correspond to the same value. If there is more than one peak in the probability density function and peak location obtained by the maximum likelihood estimation and the peak location nearest to PoA-art are different, the non-uniqueness error computed forms a significant portion of the total error.

**Table C.1**: Mean and standard deviation (S.D.) of correlation coefficients and errors across 5-fold validation sets using maximum likelihood estimation based on GMM.

| Spectral Parameter Set | Errors with respect to PoA-art (mm) | | | | | | Corr. Coeff. | |
|---|---|---|---|---|---|---|---|---|
| | Mean Error | | S.D. Error | | RMS Error | | | |
| | Mean | S.D. | Mean | S.D. | Mean | S.D. | Mean | S.D. |
| SPS-6 (MSSC, NSASS, NHBE, NVLBE, PAE, MLBE) | -0.1 | 0.3 | 2.8 | 0.2 | 2.8 | 0.2 | 0.972 | 0.004 |

**Table C.2**: Mean and standard deviation (S.D.) of errors and RMS errors for different fricatives for estimation using maximum likelihood estimation based on GMM.

| Spectral Parameter Set | Fricative | Errors with respect to PoA-art (mm) | | | | | |
|---|---|---|---|---|---|---|---|
| | | Mean Error | | S.D. Error | | RMS Error | |
| | | Mean | S.D. | Mean | S.D. | Mean | S.D. |
| SPS-6 (MSSC, NSASS, NHBE, NVLBE, PAE, MLBE ) | f | -0.6 | 0.4 | 2.8 | 0.4 | 2.9 | 0.4 |
| | v | -0.7 | 0.2 | 2.7 | 0.8 | 2.7 | 0.8 |
| | s | 0.0 | 0.3 | 2.6 | 0.2 | 2.6 | 0.2 |
| | z | 0.1 | 0.3 | 2.5 | 0.4 | 2.5 | 0.3 |
| | ʃ | 0.1 | 1.1 | 2.9 | 0.3 | 3.0 | 0.4 |
| | ʒ | 0.4 | 1.6 | 3.8 | 1.2 | 4.0 | 1.3 |

**Table C.3**: Mean and standard deviation (S.D.) of errors and RMS errors due to non-uniqueness for different fricatives for estimation using maximum likelihood estimation based on GMM.

| Spectral Parameter Set | Fricative | Errors with respect to PoA-art (mm) | | | | | |
|---|---|---|---|---|---|---|---|
| | | Mean Error | | S.D. Error | | RMS Error | |
| | | Mean | S.D. | Mean | S.D. | Mean | S.D. |
| SPS-6 (MSSC, NSASS, NHBE, NVLBE, PAE, MLBE ) | f | 0.4 | 0.4 | 2.3 | 0.7 | 2.4 | 0.7 |
| | v | 0.6 | 0.2 | 2.3 | 0.7 | 2.4 | 0.7 |
| | s | -0.1 | 0.1 | 1.2 | 0.4 | 1.2 | 0.4 |
| | z | -0.1 | 0.1 | 0.9 | 0.9 | 0.9 | 0.9 |
| | ʃ | 0.0 | 0.0 | 0.6 | 0.6 | 0.6 | 0.6 |
| | ʒ | -0.2 | 0.4 | 1.3 | 2.0 | 1.3 | 2.0 |
| | All | 0.1 | 0.1 | 1.7 | 0.3 | 1.7 | 0.3 |

## C.4 Results

The errors and the correlation coefficients for the five-fold validation for the set of the proposed parameters using the GMM-based maximum likelihood estimation are given in Table C.1. The RMS errors are about 0.3 mm more than those using the ANN with two hidden layers in Table 3.3 in the third chapter. The fricative-wise errors in Table C.2 show an increase for labiodentals and small changes for alveolars and palatals compared to those in Table 3.4 in the third chapter. The lower error using the ANN with two hidden layers may be

attributed to better modeling of the nonlinear mapping. Table C.3 gives the errors due to non-uniqueness for different fricatives using the GMM-based estimation. Comparison of these errors with those in Table C.2 shows that the error contributions of the non-uniqueness for the unvoiced and voiced labiodentals to be approximately 81% and 88%, respectively. The corresponding values are 48% and 37% for the unvoiced and voiced alveolars, and they are 19% and 32% for the unvoiced and voiced palatals. Therefore, it may be inferred that the non-uniqueness is a major error contributor for labiodentals and alveolars. Comparison of the errors for all fricatives in Tables C.1 and C.3 shows approximately 61% of the errors contributed by the non-uniqueness.

## REFERENCES


Adams FR, Crepy H, Jameson D, and Thatcher J (1989) IBM products for persons with disabilities. *Proc IEEE Global Telecommun. Conf. and Exhibition 'Communications Technology for the 1990s and Beyond'*, Dallas, Texas, USA, 980–984.

Afshan A and Ghosh PK (2015) Improved subject-independent acoustic-to-articulatory inversion. *Speech Commun.*, 66, 1–16.

Ahmed B, Monroe P, Hair A, Tan CT, Gutierrez-Osuna R, and Ballard KJ (2018) Speech-driven mobile games for speech therapy: User experiences and feasibility. *Int. J. Speech Lang. Pathol.*, 20, 644–658.

Ananthakrishnan G, Neiberg D, and Engwall O (2009) In search of non-uniqueness in the acoustic-to-articulatory mapping. *Proc. Interspeech 2009*, Brighton, UK, 2799–2802.

Anjos I, Eskenazi M, Marques N, Grilo M, Guimarães I, Magalhães J, and Cavaco S (2020) Detection of voicing and place of articulation of fricatives with deep learning in a virtual speech and language therapy tutor. *Proc. Interspeech 2020*, Shanghai, China, 3156–3160.

Arsikere H, Lulich S M, and Alwan A (2013) Non-linear frequency warping for VTLN using subglottal resonances and the third formant frequency. *Proc. IEEE Int. Conf. Acoust. Speech and Signal Process.*, Vancouver, Canada, 7922–7926.

Atal BS, Chang JJ, Mathews MV, and Tukey JW (1978) Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique. *J. Acoust. Soc. Am.*, 63, 1535–1555.

Bacsfalvi P and Bernhardt BM (2011) Long-term outcomes of speech therapy for seven adolescents with visual feedback technologies: Ultrasound and electropalatography. *Clinical Linguistics Phonetics*, 25, 1034–1043.

Baer T, Gore JC, Gracco LC, and Nye PW (1991) Analysis of vocal tract shape and dimensions using magnetic resonance imaging: Vowels. *J. Acoust. Soc. Am.*, 90, 799–828.

Baum SR and Blumstein SE (1987) Preliminary observations on the use of duration as a cue to syllable-initial fricative consonant voicing in English. *J. Acoust. Soc. Am.*, 82, 1073–1077.

Bentley J (1984) Programming pearls. *Commun. ACM*, 27, 865–871.

Bernhardt BM, Bacsfalvi P, Adler-Bock M, Shimizu R, Cheney A, Giesbrecht N, O'connell M, Sirianni J, and Radanov B (2008) Ultrasound as visual feedback in speech habilitation: Exploring consultative use in rural British Columbia Canada. *Clinical Linguistics Phonetics*, 22, 149–162.

Bigham JP, Kushalnagar R, Huang TK, Flores JP, and Savage S (2017) On how deaf people might use speech to control devices. *Proc. Int. ACM SIGACCESS Conf. Computers Accessibility*, Baltimore, MD, USA, 383–384.

Black ND (1988) Application of vocal tract shapes to vowel production. *Proc. Int. Conf. IEEE Engg. Med. Biol. Soc.*, New Orleans, LA, USA, 1535–1536.

Borg G (1946) Eine Umkehrung der Sturm-Liouvilleschen Eigenwertaufgabe (An inversion of the Sturm-Liouville eigenvalue problem). *Acta Mathematica*, 78, 1–96.

Bresch E, Adams J, Pouzet A, Lee S, Byrd D, and Narayanan S (2006a) Semi-automatic processing of real-time MR image sequences for speech production studies. *Proc. Int. Seminar Speech Prod.*, Ubatuba, Brazil, 427-434.

Bresch E, Nielsen J, Nayak K, and Narayanan S (2006b) Synchronized and noise-robust audio recordings during realtime magnetic resonance imaging scans. *J. Acoust. Soc. Am.*, 120, 1791–1794.

Busset J and Laprie Y (2013) Acoustic-to-articulatory inversion by analysis-by-synthesis using cepstral coefficients. *Proc. Meetings on Acoustics*, Montreal, Canada, 1–9.

Carey M (2004) Visual feedback for pronunciation of vowels: Kay Sona-Match. *CALICO J.*, 21, 571–601.

Carreira-Perpinan MA (2000) Mode-finding for mixtures of Gaussian distributions. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22, 1318-1323.

Carreira-Perpinan MA (2006) How many modes can a Gaussian mixture have ?. [online] Available: cs.toronto.edu/~miguel/research/GMmodes.html

Chiba T and Kajiyama M (1941) *The vowel: Its nature and structure* (Tokyo-Kaiseikan Pub. Co., Ltd, Tokyo, Japan).

Coker CH (1976) A model of articulatory dynamics and control. *Proc. IEEE*, 64, 452–460.

Crichton RG and Fallside F (1974) Linear prediction model of speech production with applications to deaf speech training. *Proc. IEE Control Sci.*, 121, 865–873.

Dagenais PA, Citz-Crosby P, Fletcher SG, and McCutcheon MJ (1994) Comparing abilities of children with profound hearing impairments to learn consonants using electropalatography or traditional aural-oral techniques. *J. Speech Hear. Res.*, 37, 687–699.

Deng H, Ward RK, Beddoes MP, and Hodgson M (2004) Estimating vocal-tract area functions from vowel sound signals over closed glottal phases. *Proc. IEEE Int. Conf. Acoust. Speech and Signal Process.*, Montreal, Canada, 589–592.

DevExtras (2020) Voice tools: Pitch, tone, and volume. Version 1.01.72, Stafford, UK. [Online] Available: play.google.com/store/apps/details?id=com.DevExtras.VoiceTools

Eide E and Gish H (1996) A parametric approach to vocal tract length normalization. *Proc. IEEE Int. Conf. Acoust. Speech and Signal Process.*, Atlanta, GA, USA, 346–348.

Elfenbein JL, Hardin-Jones MA, and Davis JM (1994) Oral communication skills of children who are hard of hearing. *J. Speech Hear. Res.*, 37, 216–226.

Engwall O (2012) Analysis of and feedback on phonetic features in pronunciation training with a virtual teacher. *Computer Assisted Lang. Learn.*, 25, 37–64.

Engwall O, Balter O, Öster A, and Kjellstrom H (2006) Designing the user interface of the computer-based speech training system ARTUR based on early user tests. *Behaviour Info. Technol.*, 25, 353–365.

Eshky A, Ribeiro MS, Cleland J, Richmond K, Roxburgh Z, Scobbie JM, and Wrench AA (2018) Ultrasuite: a repository of ultrasound and acoustic data from child speech therapy sessions. *Proc. Interspeech 2018*, Hyderabad, India, 1888–1892.

Fant G (1960) *Acoustic theory of speech production: With calculations based on X-ray studies of Russian articulations* (De Gruyter Mouton, Mouton, The Hague, Netherlands).

Flanagan JL (1975) *Speech Analysis, Synthesis, and Perception*, 2nd ed. (Springer-Verlag, New York, USA).

Fletcher SG (1982) Seeing speech in real time: The deaf can now view tongue jaw and other vocal-tract movements on a CRT display. *IEEE Spectrum*, 19, 42–45.

Forrest K, Weismer G, Milenkovic P, and Dougall RN (1988) Statistical analysis of word-initial voiceless obstruents: Preliminary data. *J. Acoust. Soc. Am.*, 84, 115-123.

Ghosh PK and Narayanan S (2010) A generalized smoothness criterion for acoustic-to-articulatory inversion. *J. Acoust. Soc. Am.*, 128, 2162–2172.

Gick B (2002) The use of ultrasound for linguistic phonetic fieldwork. *J. Int. Phonetic Assoc.*, 32, 113–121.

Glasser A, Kushalnagar K, and Kushalnagar R (2017) Feasibility of using automatic speech recognition with voices of deaf and hard-of-hearing individuals. *Proc. Int. ACM SIGACCESS Conf. Computers Accessibility*, Baltimore, MD, USA, 334–336.

Grossinho A, Cavaco S, and Magalhaes J (2014) An interactive toolset for speech therapy. *Proc. Adv. in Computer Entertainment Technolo. Conf.*, Funchal, Portugal, 1–4.

Hardcastle WJ, Gibbon FE, and Jones W (1991) Visual display of tongue palate contact: Electropalatography in the assessment and remediation of speech disorders. *British J. Disorders of Commun.*, 26, 41–74.

Harris KS (1958) Cues for the discrimination of American English fricatives in spoken syllables. *Lang. and Speech*, 1, 1–7.

Heinz JM and Stevens KN (1961) On the properties of voiceless fricative constants. *J. Acoust. Soc. Am.*, 33, 589–596.

Hiroya S and Honda M (2004) Estimation of articulatory movements from speech acoustics using an HMM-based speech production model. *IEEE Trans. Speech Audio Process.*, 12, 175–185.

Hogden J, Lofqvist A, Gracco V, Zlokarnik I, Rubin P, and Saltzman E (1996) Accurate recovery of articulator positions from acoustics: New conclusions based on human data. *J. Acoust. Soc. Am.*, 100, 1819–1834.

Hornsby BWY and Ricketts TA (2001) The effects of compression ratio signal-to-noise ratio and level on speech recognition in normal-hearing listeners. *J. Acoust. Soc. Am.*, 109, 2964–2973.

Hughes GW and Halle M (1956) Spectral properties of fricative consonants. *J. Acoust. Soc. Am.*, 28, 303–310.

Huntington DA, Harris KS, and Sholes GN (1968) An electromyographic study of consonant articulation in hearing-impaired and normal speakers. *J. Speech Hear. Res.*, 11, 147–158.

Illa A and Ghosh PK (2018) Low resource acoustic-to-articulatory inversion using bi-directional long short term memory. *Proc. Interspeech 2018*, Hyderabad, India, 3122–3126.

International Telecommunications Union (1993) *Objective Measurement of Active Speech Level*. Rec. ITU-T P.56, Geneva, Switzerland. [online] Available: itu.int/rec/T-REC-P.56

Jagabandhu (2012) A visual feedback of vocal tract shape for speech training. *M.Tech. dissertation*, EE Dept., IIT Bombay, Mumbai, India.

Jain R, Nataraj KS, and Pandey PC (2016) Dynamic display of vocal tract shape for speech training. *Proc. National Conf. on Commun.*, Guwahati, India, paper no. 1570220186.

Jaitly N and Hinton GE (2013) Vocal tract length perturbation (VTLP) improves speech recognition. *Proc. of Int. Conf. Machine Learning (ICML)*, Atlanta, Georgia, USA, 1–5.

Ji A, Berry JJ, and Johnson MT (2014a) The electromagnetic articulography Mandarin accented English (EMA-MAE) corpus of acoustic and 3D articulatory kinematic data. *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Florence, Italy, 7719–7723.

Ji A, Johnson MT, and Berry JJ (2014b) Palate-referenced articulatory features for acoustic-to-articulator inversion. *Proc. Interspeech 2014*, Singapore, 721–725.

Ji A, Johnson MT, and Berry JJ (2016) Parallel reference speaker weighting for kinematic-independent acoustic-to-articulatory inversion. *IEEE Trans. Audio Speech Lang. Process.*, 24, 1865–1875.

Johnson K (2003) Fricatives. *Acoustic and Auditory Phonetics*, 2nd ed. (Backwell, Oxford, UK), 124–127.

Jongman A (1989) Duration of fricative noise required for identification of English fricatives. *J. Acoust. Soc. Am.*, 85, 1718–1725.

Jongman A, Wayland R, and Wong S (2000) Acoustic characteristics of English fricatives. *J. Acoust. Soc. Am.*, 108, 1252–1263.

Katz WF and Mehta S (2015) Visual feedback of tongue movement for novel speech sound learning. *Frontiers Human Neurosci.*, 9, 1–13.

Kelly JL and Lochbaum C (1962) Speech synthesis. *Proc. 4th Int. Congo Acoust.*, Copenhagen, Denmark, 1–4.

Kewley-Port D, Watson CS, Elbert M, Maki D, and Reed D (1991) The Indiana Speech Training Aid (ISTRA) II: Training curriculum and selected case studies. *Clinical Linguistics Phonetics*, 5, 13–38.

Kiritani S, Ito K, and Fujimura O (1975) Tongue-pellet tracking by a computer-controlled X-ray micro beam system. *J. Acoust. Soc. Am.*, 57, 1516–1520.

Kubichek R (1993) Mel-cepstral distance measure for objective speech quality assessment. *Proc. IEEE Pacific Rim Conf. Commun. Computers Signal Proc.*, Victoria, Canada, 125–128.

Kuhl PK (2000) A new view of language acquisition. *Proc. Nat. Acad. Sci United States Am.*, 97, 11850–11857.

Kumar SVB and Umesh S (2008) Nonuniform speaker normalization using affine transformation. *J. Acoust. Soc. Am.*, 124, 1727–1738.

Ladefoged P (1982) *A Course in Phonetics*, 2nd ed. (HBJ, New York, USA).

Ladefoged P, Harshman R, and Goldstien L (1978) Generating vocal tract shapes from formant frequencies. *J. Acoust. Soc. Am.*, 64, 1027–1035.

Lee L and Rose R (1998) A frequency warping approach to speaker normalization. *IEEE Trans. Speech Audio Process.*, 6, 49–60.

Levinson SE and Schmidt CE (1983) Adaptive computation of articulatory parameters from the speech signal. *J. Acoust. Soc. Am.*, 74, 1145–1154.

Li F, Menon A, and Allen JB (2012) A psychoacoustic method for studying the necessary and sufficient perceptual cues of American English fricative consonants in noise. *J. Acoust. Soc. Am.*, 132, 2663–2675.

Liljencrants J and Fant G (1975) Computer program for VT-resonance frequency calculations. *STL-QPSR*, 16, 15–20.

Liu P, Yu Q, Wu Z, Kang S, Meng H, and Cai L (2015) A deep recurrent approach for acoustic-to-articulatory inversion. *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Brisbane, Australia, 4450–4454.

Loizou PC (2017) *Speech Enhancement: Theory and Practice*, 2nd ed. (CRC, New York, USA). [online] Available: crcpress.com/Speech-Enhancement-Theory-and-Practice-Second-Edition/Loizou/p/book/9781138075573

Maeda S (1990) Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal tract shapes using an articulatory model. *Speech production and speech modeling*, edited by Hardcastle WJ and Marchal A (Kluwer Academic, Dordrecht, Netherlands), 131–149.

Mahdi AE (2008) Visualisation of the vocal-tract shape for a computer-based speech training system for the hearing-impaired. *Open Elect. Electron. Eng. J.*, 2, 27–32.

Mahshie JJ, Vari-Alquist D, Waddy-Smith B, and Bernstein LE (1988) Speech training aids for hearing-impaired individuals: III. Preliminary observations in the clinic and children's homes. *J. Rehab. Res. Development*, 25, 69–82.

Mahshie JJ (1996) Feedback considerations for speech training systems. *Proc. Int. Conf. Spoken Lang. Process.*, Philadelphia, PA, USA, 153–156.

Martin KL, Hirson A, Herman R, Thomas J, and Pring T (2007) The efficacy of speech intervention using electropalatography with an 18-year-old deaf client: A single case study. *Adv. Speech Lang. Pathol.*, 9, 46–56.

Massaro DW and Light J (2004) Using visible speech to train perception and production of speech for individuals with hearing loss. *J. Speech Lang. Hear. Res.*, 47, 304–320.

McGowan RS (1994) Recovering articulatory movement from formant frequency trajectories using task dynamics and a genetic algorithm: Preliminary model tests. *Speech Commun.*, 14, 19–48.

McGowan RS, Nittrouer S, and Chenausky K (2008) Speech production in 12-month-old children with and without hearing loss. *J. Speech Lang. Hear. Res.*, 51, 879–888.

Micro Video Corp. (2017) Video voice speech training system. Version 3.0, Ann Arbor, Michigan, USA, [Online] Available: videovoice.com

Miller GA and Nicely PE (1955) An analysis of perceptual confusions among some English consonants. *J. Acoust. Soc. Am.*, 27, 338–352.

Mitra V, Nam H, Espy-Wilson CY, Saltzman E, and Goldstein L (2010) Retrieving tract variables from acoustics: A comparison of different machine learning strategies. *IEEE J. Sel. Topics Signal Process.*, 4, 1027–1045.

Mitra V, Nam H, Espy-Wilson, CY, Saltzman E, and Goldstein L (2012) Recognizing articulatory gestures from speech for robust speech recognition. *J. Acoust. Soc. Am.*, 131, 2270–2287.

Moeller MP, Hoover B, Putman C, Arbataitis K, Bohnenkamp G, Peterson B, Wood S, Lewis D, Pittman A, and Stelmachowicz P (2007) Vocalizations of infants with hearing loss compared to infants with normal hearing. Part I: Phonetic development. *Ear Hear.*, 28, 605–627.

Munhall KG, Vatikiotis-Bateson E, and Tohkura Y (1995) X-ray film database for speech research. *J. Acoust. Soc. Am.*, 98, 1222–1224.

Nam H, Mitra V, Tiede M, Hasegawa-Johnson M, Espy-Wilson CY, Saltzman E, and Goldstein L (2012) A procedure for estimating gestural scores from speech acoustics. *J. Acoust. Soc. Am.*, 132, 3980–3989.

Narayanan S, Alwan A, and Haker K (1995) An articulatory study of fricative consonants using magnetic resonance imaging. *J. Acoust. Soc. Am.*, 98, 1325–1347.

Narayanan S, Nayak K, Lee S, Sethy A, and Byrd D (2004) An approach to real-time magnetic resonance imaging for speech production. *J. Acoust. Soc. Am.*, 115, 1771–1776.

Narayanan S, Toutios A, Ramanarayanan V, Lammart A, Kim J, Lee S, Nayak KS, Kim Y, Zhu Y, Goldstein L, Byrd D, Bresch E, Ghosh PK, Katsamanis A, and Proctor M (2014) Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research. *J. Acoust. Soc. Am.*, 136, 1307–1311.

Nataraj KS, Jagbandhu, Pandey PC, and Shah MS (2011) Improving the consistency of vocal tract shape estimation. *Proc. National Conf. Commun.*, Bangalore, India, paper SpPrII.4.

Nataraj KS, Pandey PC, and Dasgupta H (2017) Estimation of place of articulation of fricatives from spectral characteristics for speech training. *Proc. Interspeech 2017*, Stockholm, Sweden, 339–343.

Nayak NS, Velmurugan R, Pandey PC, and Saha S (2012) Estimation of lip opening for scaling of vocal tract area function for speech training aids. *Proc. National Conf. Commun.*, Kharagpur, India, 521–525.

Neri A, Cucchiarini C, Strik H, and Boves L (2002) The pedagogy–technology interface in computer assisted pronunciation training. *Computer Assisted Lang. Learning*, 15, 441–467.

Ngo T, Akagi M, and Birkholz P (2020) Effect of articulatory and acoustic features on the intelligibility of speech in noise: An articulatory synthesis study. *Speech Commun.*, 117, 13–20.

Nickerson RS and Stevens KN (1973) Teaching to a deaf: can a computer help?. *IEEE Trans. Audio Electroacoust.*, 21, 445–455.

Nissen SL and Fox RA (2005) Acoustic and spectral characteristics of young children's fricative productions: A developmental perspective. *J. Acoust. Soc. Am.*, 118, 2570–2578.

Nober EH (1967) Articulation of the deaf. *Exceptional Children*, 33, 611–621.

Olson DJ (2014) Benefits of visual feedback on segmental production in the L2 classroom. *Lang. Learning Technol.*, 18, 173–192.

O'Shaughnessy D (2000) *Speech Communication: Human and Machine*, 2nd ed. (IEEE Press, Piscataway, NJ, USA).

Öster A (2006) Computer-based speech therapy using visual feedback with focus on children with profound hearing impairments. *Ph.D. dissertation*, Dept. of Speech, Music and Hearing, KTH School of Computer Science and Communication, Stockholm, Sweden.

Panchapagesan S and Alwan A (2011) 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.*, 129, 2144–2162.

Pandey PC and Shah MS (2009) Estimation of place of articulation during stop closures of vowel-consonant-vowel utterances. *IEEE Trans. Audio Speech Lang. Process.*, 17, 277–286.

Pardo JM (1982) Vocal tract shape analysis for children. *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Paris, France, 763–766.

Park SH, Kim DJ, Lee JH, and Yoon TS (1994) Integrated speech training system for hearing impaired. *IEEE Trans. Rehab. Eng.*, 2, 189–196.

Pickett KL (2013) The effectiveness of using electropalatography to remediate a developmental speech sound disorder in a school-aged child with hearing impairment. *M.S. thesis*, Dept. Commun. Disorders, Brigham Young University, Provo, Utah, USA.

Purr Programming (2017) Voice pitch analyser. Version 1.3.0, Hamburg, Germany. [Online] Available: play.google.com/store/apps/details?id=de.lilithwittmann.voicepitchanalyzer&hl=en

Qin C and Carreira-Perpiñán MA (2007) An emperical investigation of the nonuniqueness in the acoustic-to-articulatory mapping. *Proc. Interspeech 2007*, Antwerp, Belgium, 74–77.

Qin C, Carreira-Perpiñán MA, Richmond K, Wrench A, and Renals S (2008) Predicting tongue shapes from a few landmark locations. *Proc. Interspeech 2008*, Brisbane, Australia, 2306-2309.

Rabiner LR, Sambur MR, and Schmidt CE (1975) Applications of a nonlinear smoothing algorithm to speech processing. *IEEE Trans. Acoust. Speech Signal Process.*, ASSP-23, 552- 557.

Richmond K (2006) A trajectory mixture density network for the acoustic-articulatory inversion mapping. *Proc. Intrespeech 2006*, Pittsburgh, Pennsylvania, USA, 577–580.

Richmond K, Hoole P, and King S (2011) Announcing the electromagnetic articulography (day 1) subset of the mngu0 articulatory corpus. *Proc. Interspeech 2011*, Florence, Italy, 1505–1508.

Riegelsberger EL (1997) The acoustic-to-articulatory mapping of voiced and fricated speech. *Ph.D. dissertation*, The Ohio State University, Columbus, Ohio, USA.

Roth FP and Worthington CK (2011) *Treatment Resource Manual for Speech-Language Pathology*, 4th ed. (Delmar, New York, USA).

Rouco A and Recasens D (1996) Reliability of electromagnetic midsagittal articulometry and electropalatography data acquired simultaneously. *J. Acoust. Soc. Am.*, 100, 3384–3389.

Schroeder MR (1967) Determination of the geometry of the human vocal tract by acoustic measurements. *J. Acoust. Soc. Am.*, 41, 1002–1010.

Schroeter J and Sondhi MM (1994) Techniques for estimating vocal-tract shapes from the speech signal. *IEEE Trans. Speech and Audio Process.*, 2, 133–150.

Serizel R and Giuliani D (2014) Vocal tract length normalisation approaches to DNN-based children's and adults' speech recognition. *Proc. IEEE Spoken Lang. Technology Workshop*, South Lake Tahoe, NV, USA, 135–140.

Serizel R and Giuliani D (2017) Deep-neural network approaches for speech recognition with heterogeneous groups of speakers including children. *Natural Lang. Engg.*, 23, 325–350.

Shadle CH and Mair SJ (1996) Quantifying spectral characteristics of fricatives. *Proc. Int. Conf. Spoken Lang. Process.*, Philadelphia, PA, USA, 1521–1524.

Shahnawazuddin S, Kathania HK, Dey A, Sinha R (2018). Improving children's mismatched ASR using structured low-rank feature projection. *Speech Commun.*, 105, 103–113.

Shiller DM and Rochon M (2014) Auditory-perceptual learning improves speech motor adaptation in children. *J. Exp. Psychol.: Human Perception Performance*, 40, 1308–1315.

Shirai K and Masaki S (1983) An estimation of the production process for fricative consonants. *Speech Commun.*, 2, 111–114.

Sivaraman G, Mitra V, Nam H, Tiede M, and Espy-Wilson C (2019) Unsupervised speaker adaptation for speaker independent acoustic to articulatory speech inversion. *J. Acoust. Soc. Am.*, 146, 316–329.

Smith CR (1975) Residual hearing and speech production in deaf children. *J. Speech Hear. Res.*, 18, 795–811.

Sock R, Hirsch F, Laprie Y, Perrier P, Vaxelaire B, Brock G, Bouarourou F, Fauth C, Hecker V, Ma L, and Busset J (2011) An X-ray database tools and procedures for the study of speech production. *Proc. Int. Seminar Speech Production*, Montréal, Canada, 41-48.

Sondhi MM and Gopinath B (1971) Determination of vocal-tract shape from impulse response at the lips. *J. Acoust. Soc. Am.*, 49, 1867–1873.

Sondhi MM (1979) Estimation of vocal-tract areas: The need for acoustical measurements. *IEEE Trans. Acoust. Speech and Signal Process.*, 27, 268–273.

Sondhi MM and Schroeter J (1987) A hybrid time-frequency domain articulatory speech synthesizer. *IEEE Trans. Acoust. Speech Signal Process.*, ASSP-35, 955–967.

Soquet A, Saerens M, and Jospa P (1990) Acoustic-articulatory inversion based on a neural controller of a vocal tract model. *Proc. The ESCA Workshop on Speech Synthesis*, Autrans, France, 71–74.

Speechtools Ltd. (2019) Voice analyst. Version 3.8.1, Bristol, UK. [Online] Available: play.google.com/store/apps/details?id=co.speechtools.voiceanalyst

Stevens KN (2000) Obstruent consonants. *Acoustic Phonetics*, (MIT Press, Cambridge, MA, USA), 389–391.

Stevens KN and Blumstein SE (1978) Invariant cues for place of articulation in stop consonants. *J. Acoust. Soc. Am.*, 64, 1358–1368.

Stevens KN, Kasowski S, and Fant G (1953) An electrical analog of the vocal tract. *J. Acoust. Soc. Am.*, 25, 734–742.

Stone M (1990) A three-dimensional model of tongue movement based on ultrasound and X-ray microbeam data. *J. Acoust. Soc. Am.*, 87, 2207–2217.

Story BH (2007) Time dependence of vocal tract modes during production of vowels and vowel sequences. *J. Acoust. Soc. Am.*, 121, 3770–3789.

Story BH, Titze IR, and Hoffman EA (1996) Vocal tract area functions from magnetic resonance imaging. *J. Acoust. Soc. Am.*, 100, 537–554.

Takaoka T (2002) Efficient algorithms for the maximum subarray problem by distance matrix multiplication. *Electron. Notes Theoretical Comput. Sci.*, 61, 191–200.

Tiger DRS (1999) Dr. Speech. Seatle, WA, USA, 1998. [Online] Available: drspeech.com

Toda T, Black AW, and Tokuda K (2004) Mapping from articulatory movements to vocal tract spectrum with Gaussian mixture model for articulatory speech synthesis. *Proc. 5th ISCA Speech Synth. Workshop*, Pittsburgh, PA, USA, 31–36.

Toda T, Black A, and Tokuda K (2008) Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model. *Speech Commun.*, 50, 215–227.

Todd AE, Edwards JR, and Litovsky RY (2011) Production of contrast between sibilant fricatives by children with cochlear implants. *J. Acoust. Soc. Am.*, 130, 3969–3979.

Uria B, Murray I, Renals S, and Richmond K (2012) Deep architectures for articulatory inversion. *Proc. Interspeech 2012*, Portland, Oregon, USA, 867–870.

Vicsi K, Roach P, Öster A, Kacic Z, Barczikay P, Tantos A, Csatári F, Bakcsi Zs, and Sfakianaki A (2000) A multimedia multilingual teaching and training system for children with speech disorders. *Int. J. Speech Technol.*, 3, 289–300.

Wagner A, Ernestus M, and Cutler A (2006) Formant transitions in fricative identification: The role of native fricative inventory. *J. Acoust. Soc. Am.*, 120, 2267–2277.

Wakita H (1973) Direct estimation of the vocal tract shape by inverse filtering of acoustic speech waveforms. *IEEE Trans. Audio Electroacoust.*, AE-21, 417–427.

Wakita H (1979) Estimation of vocal-tract shapes from acoustical analysis of the speech wave: The state of the art. *IEEE Trans. Acoust. Speech Signal Process.*, ASSP-27, 281–285.

Westbury JR (1994) X-ray microbeam speech production database user's handbook. Version 1.0, Waisman Center on Mental Retardation and Human Development, University of Wisconsin, Madison, Wisconsin, USA. [Online] Available: berkeley.app.box.com/v/xray-microbeam-database-data

Wilson I (2014) Using ultrasound for teaching and researching articulation. *Acoust. Sci. Technol.*, 35, 285–289.

Wilson I and Gick B (2006) Ultrasound technology and second language acquisition research. *Proc. Generative Approaches to Second Lang. Acquisition Conf.*, Somerville, MA, USA, 148–152.

Wrench AA and William HJ (2000) A multichannel articulatory database and its application for automatic speech recognition. *Proc. 5th Seminar on Speech Production: Models and Data*, Bavaria, Germany, 305–308.

Xu L, Thompson CS, and Pfingst BE (2005) Relative contributions of spectral and temporal cues for phoneme recognition. *J. Acoust. Soc. Am.*, 117, 3255–3267.

Young S, Evermann G, Gales M, Hain T, Kershaw D, Liu X, Moore G, Odell J, Ollason D, Povey D, Valtchev V, and Woodland P (2006) The HTK book, Version 3.4. [online] Available: htk.eng.cam.ac.uk/prot-docs/htk_book.shtml

Yuan J and Liberman M (2008) Speaker identification on the SCOTUS corpus. *J. Acoust. Soc. Am.*, 123, 3878.

Zahorian SA and Venkat S (1990) Vowel articulation training aid for the deaf. *Proc. lnt. Conf. on Acoust. Speech and Signal Process.*, New York, USA, 1121–1124.

Zhang J (1997) Articulograph AG100 electromagnetic articulation analyzer. Version 1.2, UCLA Phonetics Lab, Los Angeles, California, USA. [Online] Available: phonetics.linguistics.ucla.edu/facilities/physiology/Emamual.html

Zharkova N (2016) Ultrasound and acoustic analysis of sibilant fricatives in preadolescents and adults. *J. Acoust. Soc. Am.*, 139, 2342–2351.

Zhou N, Xu L, and Lee C (2010) The effects of frequency-place shift on consonant confusion in cochlear implant simulations. *J. Acoust. Soc. Am.*, 128, 401–409.

Zhou X, Zhang Z, and Espy-Wilson C.Y (2004) VTAR: A Matlab-based computer program for vocal tract acoustic modeling. *J. Acoust. Soc. Am.*, 115, 2543.

Zwicker E (1961) Subdivision of the audible frequency range into critical bands (Freqenzgruppen). *J. Acoust. Soc. Am.*, 33, 248.

# Thesis Related Publications

**Journal papers**

1. K. S. Nataraj, P. C. Pandey, and H. Dasgupta "Estimation of place of articulation of fricatives from spectral parameters using artificial neural network," under revision (September 2021).
2. K. S. Nataraj, P. C. Pandey, and H. Dasgupta "Estimation of place of articulation of fricatives using spectral parameters during frication and vocalic transition segments," under preparation (September 2021).

**Papers in conference proceedings**

1. K. S. Nataraj, P. C. Pandey, and H. Dasgupta, "Effect of frication duration and formant transitions on the perception of fricatives in VCV utterances," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. 2020 (ICASSP 2020)*, Barcelona, Spain, May 2020, pp. 6259–6263.
2. K. S. Nataraj, H. Dasgupta, and P. C. Pandey, "Early indirect techniques for estimating the vocal tract area function," in *Proc. 3 rd Int. Workshop History Speech Commun. Res. (HSCR 2019)*, Vienna, Austria, 2019.
3. K. S. Nataraj, P. C. Pandey, and H. Dasgupta, "Estimation of place of articulation of English fricatives using the modified dominant spectral centroid and slope as the spectral parameters," in *Proc. Int. Workshop Speech Process. Voice Speech Hear. Disorders 2018 (WSPD 2018)*, Mysore, India, 2018, paper no. 39.
4. 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 Conf. Int. Speech Commun. Association (Interspeech 2017)*, Stockholm, Sweden, August 20-24, 2017, pp. 339–343.
5. R. Jain, K. S. Nataraj, and P. C. Pandey, "Dynamic display of vocal tract shape for speech training," in *Proc. National Conf. Commun. 2016 (NCC 2016)*, Guwahati, India, 2016, paper no. 1570220186.
6. K. S. Nataraj and P. C. Pandey, "Place of articulation from direct imaging for validation of its estimation from speech analysis for use in speech training," in *Proc. 5th National Conf. Computer Vision, Pattern Recognition, Image Process., and Graphics 2015 (NCVPRIPG 2015)*, Patna, India, 2015, paper ID 88.

## Author's Resume

**K. S. Nataraj** received the B.E. degree in electronics and communication engineering in 2005 from Bapuji Institute of Engineering and Technology, Davangere, affiliated to Visvesvaraya Technological University, Belgaum, India and the M.Tech. degree in electrical engineering in 2012 from the Indian Institute of Technology Bombay,Mumbai, India, where he is currently a Ph.D. student. He worked as a software engineer at Robert Bosch Engineering and Business Solutions, Bangalore from September 2005 to August 2008 and as a digital signal processing engineer at Tensilica Technologies India, Pune from July 2012 to July 2014. His research interests are speech processing and embedded system design.

### Publications

***Papers***


1. K. S. Nataraj, P. C. Pandey, and H. Dasgupta, "Effect of frication duration and formant transitions on the perception of fricatives in VCV utterances," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. 2020 (ICASSP 2020)*, Barcelona, Spain, 2020, pp. 6259–6263.

2. K. S. Nataraj, H. Dasgupta, and P. C. Pandey, "Early indirect techniques for estimating the vocal tract area function," in *Proc. 3 rd Int. Workshop History Speech Commun. Res. (HSCR 2019)*, Vienna, Austria, 2019.

3. H. Dasgupta, P. C. Pandey, and K. S. Nataraj, Epoch detection using Hilbert envelope for glottal excitation enhancement and maximum-sum subarray for epoch marking, *IEEE J. Sel. Topics in Signal Process.*, vol. 14, no. 2, pp. 461–471, 2020.

4. K. S. Nataraj, P. C. Pandey, and H. Dasgupta, "Estimation of place of articulation of English fricatives using the modified dominant spectral centroid and slope as the spectral parameters," in Proc. Int. Workshop *Speech Process. Voice Speech Hear. Disorders 2018 (WSPD 2018)*, Mysore, India, 2018, paper no. 39.

5. H. Dasgupta, P. C. Pandey, and K. S. Nataraj, Detection of glottal excitation epochs in speech signal using Hilbert envelope, in *Proc. 19th Annual Conf. Int. Speech Commun. Association (Interspeech 2018)*, Hyderabad, India, 2018, pp. 2132–2136.

6. 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 Conf. Int. Speech Commun. Association (Interspeech 2017)*, Stockholm, Sweden, 2017, pp. 339–343.

7. R. Jain, K. S. Nataraj, and P. C. Pandey, "Dynamic display of vocal tract shape for speech training," in *Proc. National Conf. Commun. 2016 (NCC 2016)*, Guwahati, India, 2016, paper no. 1570220186.

8. K. S. Nataraj and P. C. Pandey, "Place of articulation from direct imaging for validation of its estimation from speech analysis for use in speech training," in *Proc. 5th National Conf. Computer Vision, Pattern Recognition, Image Process., and Graphics 2015 (NCVPRIPG 2015)*, Patna, India, 2015, paper ID 88.

9. Jagbandhu, K. S. Nataraj, and P. C. Pandey, Detection of transition segements in VCV utterences for estimation of the place of closure of oral stops for speech training, Proc. 13th Annual Conference of the International Speech Communication Association *(Interspeech 2012)*, Portland, Oregon, 2012, pp. 406–409.

10. K. S. Nataraj, Jagbandhu, P. C. Pandey, and M. S. Shah, Improving the consistency of vocal tract shape estimation, Proc. National Conference on Communications 2011 (NCC 2011), Bangalore, India.

***Patents and patent applications***

1. P. C. Pandey, H. Dasgupta, and K. S. Nataraj, "Real-time pitch tracking by detection of glottal excitation epochs in speech signal using Hilbert envelope," Indian patent application publication 201821032901 A, Sep 01, 2018, PCT application publication WO2020/044362 A2, Aug. 03 2019.

**Acknowledgments**

I would like to express my sincere and profound gratitude to my supervisor Prof. P. C. Pandey, for his invaluable guidance, vision, motivation, and unconditional support throughout my time as a Ph.D. student. I am grateful to him for being incredibly patient and correcting me on many important aspects of doing research, including critical thinking, systematic experimentation, clear writing, and clear presentation. I am grateful to Prof. Preeti Rao and Prof. V. Rajbabu, members of the research progress committee, for their valuable suggestions and insightful questions at various stages of my work.

I would like to thank Mr. Vidyadhar Kamble for helping me with lab-related issues and for helping me work remotely on the computers in the lab during COVID situation. I am grateful to Hirak, Nitya, Pramod, Uttam, Mohan, Saketh, Vikas, Hitesh, and Dr. M. S. Shah for sharing interesting technical discussions and teaching me many technical concepts which were useful to carry out my research. I would like to thank Hirak and Pramod for reading my thesis and giving valuable suggestions. I would like to thank Prachir, Rahul, Shreyansh, Yogesh, Shibam, Susmi, and Anand for interesting discussions and for making my time at IIT memorable. I would like to thank all the staff at the EE department office, IIT Bombay, for helping me with all academic-related queries.

I am indebted to my parents, parents-in-law, and other family members for being patient and supportive throughout. I am grateful to my brothers for helping me pursue higher studies by freeing me from the family responsibilities and providing me the financial support. I am indebted to my wife Nitya for calming my nerves and providing unconditional support whenever I was exhausted and anxious. I want to remember my grandmother, who left me with so many memories of unconditional love that provide me the strength to face all the challenges.