HSCR 2019, 2019, Sep. 13-14, Vienna, Austria.

EARLY INDIRECT TECHNIQUES FOR ESTIMATING THE VOCAL TRACT AREA FUNCTION

K. S. Nataraj, H. Dasgupta, P. C. Pandey

Indian Institute of Technology Bombay, Mumbai, India
natarajks@ee.iitb.ac.in

Abstract: Estimation of the vocal tract area function has been an interesting area of research for many decades due to its potential applications in speech coding, speech synthesis, speech recognition, and speech training. Earlier studies using acoustic tube models of the vocal tract showed that the speech spectrum could be estimated from the vocal tract area function. The inverse problem of obtaining the vocal tract area function from the speech spectrum was challenging due to the non-uniqueness of the solution as many area functions could produce the same spectrum. Some innovative and pioneering techniques were proposed to estimate the vocal tract area functions indirectly by using acoustic measurements. These techniques showed that reasonable vocal tract area functions could be obtained from the acoustic data alone and stimulated further research in this area. This paper provides a review of three early techniques for estimating the vocal tract area function: the technique using the acoustic impedance measurement at the lips, the technique using the impulse response measured at the lips, and the technique using linear prediction analysis of the speech signal, followed by a discussion of the applications and limitations of these techniques.

1 Introduction

Several studies have contributed to improving the understanding of speech production using the acoustic tube models of the vocal tract [1]–[5]. These studies model the vocal tract 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. Assuming plane wave propagation and appropriate boundary conditions, a speech signal can be synthesized for given vocal tract area function by using electric analogs of the vocal tract [3] or by digital simulation of the vocal tract [5]. 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 functions from the speech signal is difficult due to the non-uniqueness of the mapping between the acoustic parameters and the vocal tract area function, as different vocal tract area functions can produce similar spectra. Research for solution to this inverse problem was motivated by its potential applications in speech coding, speech synthesis, speech recognition, and speech training.

Chiba and Kajiyama [1] carried out 3D measurements of the vocal tract using X-ray photographs and solid palatograms made of plaster of paris, used the measurements to create mechanical models of the vocal tract using a plasticine material, and used these models to generate vowel sounds using a telephone receiver as the excitation source. Arai [7] replicated the models using acrylic resin and produced vowel sounds using an electrolarynx as the excitation source. Fant [4], [6] has reported the use of X-ray based direct imaging to study vocal tract shapes of vowel sounds.

The X-ray imaging provided good visualization of the vocal tract configuration during articulation for use in speech research, but were discontinued due to the risk of radiation hazards for the speakers. As an alternative to direct imaging, several innovative techniques were proposed to estimate the vocal tract area functions from the acoustic measurements or the speech waveform. In 1967, Schroeder [8] proposed a technique to estimate the vocal tract area function from the acoustic impedance measured at the lips. Sondhi and Gopinath [9] pointed out limitations of the impedance technique and proposed a technique to estimate the vocal tract area function from the measurement of the impulse response at the lips. These techniques using the acoustic measurements showed that reasonable vocal tract shape could be obtained from acoustic measurement alone, but they were not practical for applications other than collection of the articulatory data for speech research. Wakita [10] proposed a technique based on the linear prediction analysis of the speech signal for estimating the vocal tract area function from the speech waveform and showed that the technique provided reasonable estimates for the non-nasalized vowels. This paper provides a review of these three innovative techniques of estimating the vocal tract area function indirectly from acoustic measurements or the speech waveform. The second section reviews the technique using the acoustic impedance measured at the lips, the third section provides a review of the technique based on the impulse response measurement, and the linear prediction based technique using only the speech signal is reviewed in the fourth section. A discussion of the limitations and applications of the techniques along with a brief description of the current status is presented in the last section.

2 Estimation of the vocal tract area function using measurement of the acoustic impedance at the lips

In this technique reported by Schroeder [8], the vocal tract is modeled as a lossless and rigid-walled tube with the transverse dimension of less than 5 cm, resulting in plane wave propagation in the vocal tract for frequencies less than 4 kHz. Based on these assumptions, the pressure $p(x)$ along the distance x from the glottis end in the tract is given as

$$\frac{d}{dx}\left[A(x)\frac{dp}{dx}\right]+\frac{\omega^2}{c^2}A(x)p=0 \tag{1}$$

where $A(x)$ is the cross-sectional area as a function of the distance along the vocal tract, ω is the frequency, and c is the sound velocity in air. This technique is based on the earlier finding by Borg [11] that the vocal tract area function can be uniquely estimated from the eigenvalues of two boundary value problems associated with plane wave propagation. One set of boundary conditions corresponds to the vocal tract closed at the glottis and open at the lips and the other one corresponds to the vocal tract closed at both ends. The resonant frequencies with the lips open correspond to the formant frequencies of the speech signal and can be obtained from its spectral analysis. However, measuring the resonant frequencies when the lips are closed is not feasible. The problem was solved by Schroeder by using the acoustic impedance measured at the lips to obtain the required resonant frequencies under the two boundary conditions. The resonant frequencies with closed-lip boundary condition correspond to the zeros in the acoustic impedance at the lips while the resonant frequencies during the open-lip boundary correspond to the poles of the acoustic impedance. An important result of this technique was that the first $2n$ lowest-order 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. Although the result was not analytically proved, it was empirically verified.

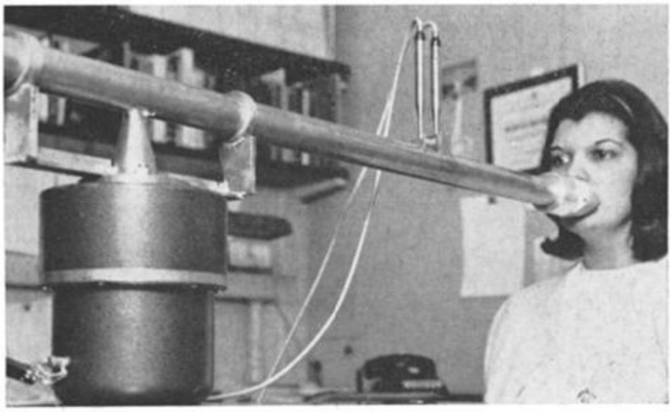

Figure 1. Experimental setup for measuring the acoustic impedance at the lips as used in Schroeder's technique [8])

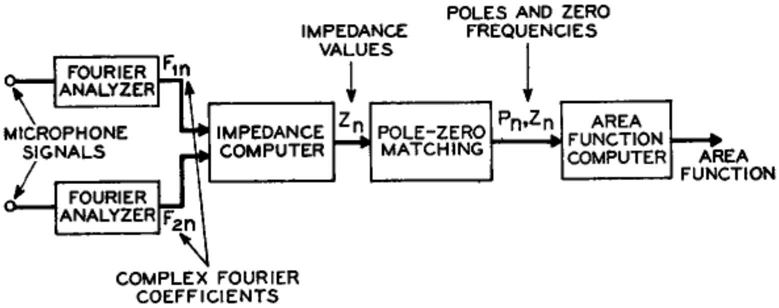

Figure 2. Block diagram of the computation of the area function from the microphone signals of the impedance measuring setup in Schroeder's technique [8])

The impedance measurement setup as used in Schroeder's technique is shown in Figure 1. It consists of an impedance tube, an electrodynamic driver unit, and two microphones. The electrodynamic driver unit, coupled to the left end of the tube provides the periodic acoustic pulses with a bandwidth of 4 kHz and a repetition period of 10 ms. The right end of the tube was connected to the mouth of the subject using a specially designed seal. The subject was asked to articulate silently with vocal cords closed. Two closely spaced condenser microphones were used to pick up the sound pressure of the incoming and reflected sound waves. The inner diameter of the tube was 3.48 cm resulting in plane wave propagation in the tube for frequencies less than 5.7 kHz. The microphone signals were acquired for digital processing. The input impedance z_n was computed using the complex Fourier coefficients F_{1n} and F_{2n} of the two signals as

$$z_n = z_0 \frac{F_{1n}\sin(2\pi f_0 \tau_1) - F_{2n}\sin(2\pi f_0 \tau_2)}{F_{1n}\cos(2\pi f_0 \tau_1) - F_{2n}\cos(2\pi f_0 \tau_2)} \quad (2)$$

where z_0 is the characteristic impedance of the tube, and τ_1 and τ_2 are the acoustic wave propagation times from microphone 1 and 2, respectively, to the lips.

The computation of the vocal tract area function is shown as a block diagram in Figure 2, with the impedance calculated from the microphone signals using (2). The poles and zeros of the measured impedance were compared with those obtained from the perturbed vocal tract area function. The area function was iteratively adjusted to match the poles and zeros with

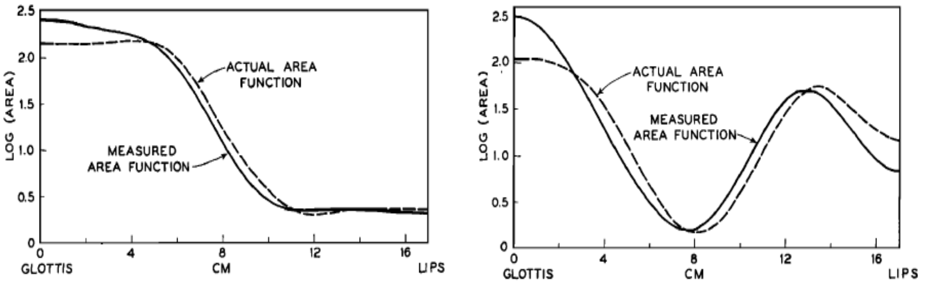

Figure 3. Vocal tract area function estimated from impedance technique (continuous), actual area function (dotted) (from Schroeder [8])

those obtained from the impedance measurement, using the root mean square error criterion. From the results obtained for some known fixed shapes as shown in Figure 3, it can be observed that the correspondence between measured and actual area function was remarkably good. This technique showed that reasonable vocal tract area function could be estimated from the acoustic measurement alone. It stimulated further research in this area. Mermelstein [12] showed that only the first few poles and zeros of the impedance were needed to obtain a good approximation of the area function. The impedance technique suffers from unnatural articulation due to the sealing between the acoustic tube and the mouth. Further, it was pointed out by Sondhi and Gopinath [8] that the convergence of the iterative procedure in the impedance technique was not proved and the assumptions of fixed vocal tract length and closed glottis resulted in errors in the estimation.

3 Estimation of the vocal tract area function by measurement of the impulse response at the lips

To avoid the limitations of the acoustic impedance technique, Sondhi and Gopinath [9] proposed a technique to determine the vocal tract area function from the impulse response at the lips, measured as the pressure developed at the lips as a function of time in response to a unit impulse of volume velocity applied at the lips. 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, and showed that the vocal tract area function can be uniquely obtained from the impulse response without involving the vocal tract length and boundary conditions at the lips.

The experimental setup for impulse response measurement, shown in Figure 4, consisted of a tube with a sound absorbing wedge at the left end and connected to the mouth at the other end, a microphone, and a sound source. The impulse generated by the source travels in both directions. The impulse moving towards the left is absorbed by the wedge. The impulse moving towards the right is picked-up by the microphone. After reaching the lips, it is reflected towards the source and picked-up by the microphone. The impulse response at the lips was obtained from the pressure and volume measurements.

For a vocal tract length of 17 cm, the impulse response measured for a duration less than 1 ms was found to be sufficient to estimate the area function. For this short interval, the area function can be assumed to be stationary even during conversational speech. The technique allowed for estimation of the dynamically varying vocal tract area functions by periodic measurements of the impulse response at the interval of 20–30 ms. From the result obtained

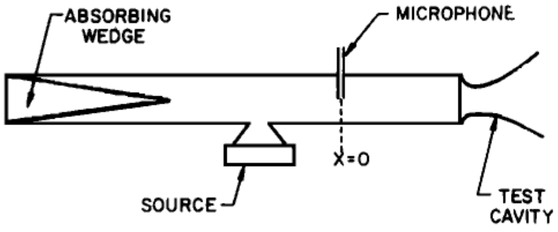

Figure 4. Experimental setup used by Sondhi and Gopinath [9] for measuring the impulse response at lips

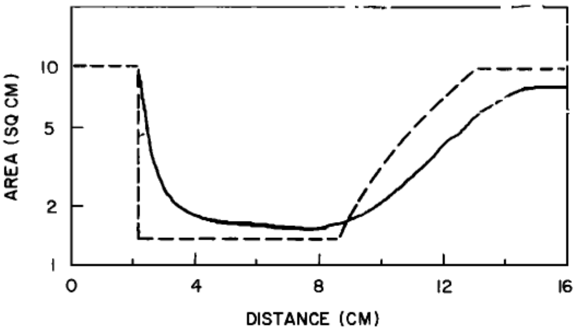

Figure 5. Vocal tract area function using the impulse response technique for a metal tube (Dotted: metal tube shape, continuous: estimated shape, from Sondhi and Gopinath [9].

for a metal tube of known area function, shown in Figure 5, it is observed that technique estimated a reasonable vocal tract shape with a smoothening at the sharp discontinuities.

4 Estimation of the vocal tract area function using linear prediction analysis of the speech signal

Atal [13] reported a technique for estimating the vocal tract area function directly from the speech signal using linear prediction (LP) analysis. He showed that modelling of the vocal tract as an acoustic tube with a specific number of cylindrical sections can be used to uniquely determine the vocal tract area function from the formant frequencies and their bandwidths and that the vocal tract area function can be obtained directly from the autocorrelation function of the speech signal. Wakita [10] showed that the acoustic tube model of the vocal tract is equivalent to the inverse filter obtained from LP analysis of the speech signal. He also demonstrated that a reasonable vocal tract area function can be estimated from the autocorrelation coefficients of the speech signal after pre-emphasis, and with appropriate boundary conditions.

In Wakita's technique [10], speech is modeled as being generated by an excitation source followed by a filter which takes into consideration the effects of the glottis, vocal tract, and radiation at the lips as shown in Figure 6. For non-nasalized voiced sounds, excitation is assumed to be an impulse train. The inverse filter coefficients are obtained by least mean square error technique. An acoustic tube filter that is equivalent to the inverse filter is obtained to relate the vocal tract configuration to the frequency domain behavior of the speech signal. The vocal tract configuration is estimated by relating the two inverse filter models.

In this technique, the vocal tract is modeled as a lossless acoustic tube with *M* sections of equal length and varying cross-sectional area to obtain an acoustic inverse transfer function,

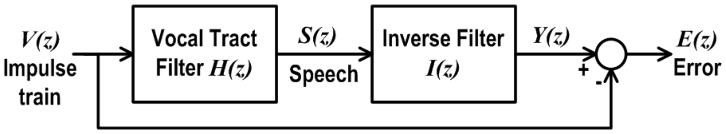

Figure 6. Speech analysis model for estimating vocal tract area function by the LP analysis in Wakita's technique [10]

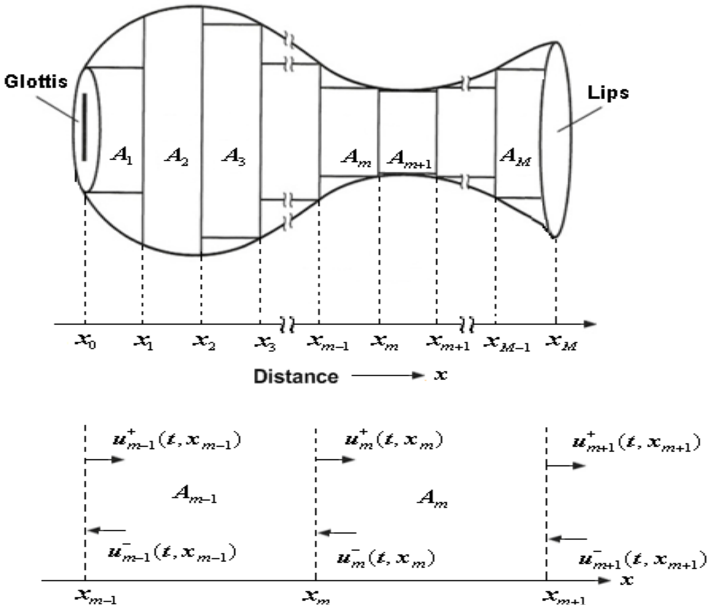

Figure 7. Acoustic tube model of the vocal tract used in Wakita's technique [10]

as shown in Figure 7. The volume velocity in section *m* is represented by $u_m(t, x_m)$, with *t* as the time and x as the distance from the glottis. For plane wave propagation through the acoustic tube, reflections occur at the section interfaces due to different areas on the two sides. The inverse transfer function for the acoustic tube model is given as

$$D_M(z) = \frac{\text{forward-going volume velocity component at the glottis end}}{\text{volume velocity component at the lip end}} \tag{3}$$

It is obtained using the following recursive relation

$$\begin{bmatrix} D^+_{m+1}(z) \\ D^-_{m+1}(z) \end{bmatrix} = \begin{bmatrix} 1 & -\mu_{m+1} \\ -\mu_{m+1} z^{-k} & z^{-1} \end{bmatrix} \begin{bmatrix} D^+_m(z) \\ D^-_m(z) \end{bmatrix} \tag{4}$$

where μ_m is the reflection coefficient at the junction between sections *m* and *m+1*. To obtain the inverse filter transfer function using LP analysis, the envelope of the power spectrum is approximated by poles only. The corresponding inverse filter has only zeros in its transfer function and is given as

$$I(z) = 1 - \sum_{k=1}^{p} \alpha_k z^{-k} \tag{5}$$

where α_k's are predictor coefficients for minimizing the sum of the squared approximation error between the output of the inverse filter and the input impulse train. The optimum inverse filter coefficients are obtained using the recursive relation given as

$$\begin{bmatrix} I_{m+1}(z) \\ J_{m+1}(z) \end{bmatrix} = \begin{bmatrix} 1 & -k_m \\ -k_m z^{-k} & z^{-1} \end{bmatrix} \begin{bmatrix} I_m(z) \\ J_m(z) \end{bmatrix}, \quad m=0,1,\ldots,M\text{-}1 \tag{6}$$

The coefficients k_m's are the partial correlation coefficients calculated using the auto-correlation coefficients with $I_0(z)=1$ and $J_0(z)=-z^{-1}$. Comparing (4) and (6), Wakita showed that the two transfer functions are equivalent, and that the reflection coefficients μ_m's are identical to the partial correlation coefficients k_m's. In other words, the wave propagation in the lossless acoustic tube is equivalent to the optimum inverse filtering of the speech wave-form. Thus the reflection coefficients are obtained as

$$\mu_m = k_{m-1} \tag{7}$$

The area ratio A_m of the mth section is calculated from the reflection coefficient μ_m as

$$A_m = \frac{1+\mu_m}{1-\mu_m} A_{m+1} \tag{8}$$

These area ratios, directly obtained from the acoustic speech waveform, are converted into areas by assuming a constant normalized area of unity at the glottis end.

Wakita used +6 dB/octave pre-emphasis to account for the –12 dB/octave slope of the glottal spectral envelope and +6 dB/octave slope for the radiation impedance. The spectrum of the resulting speech was assumed to be the spectrum of the transfer function of a lossless tract, with a zero load at the lips and a resistive load at the glottis. The speech signal was digitized with a sampling frequency of 7 kHz. The technique can be used for estimating fixed as well as transitional vocal tract configurations during speech segments with glottal excitation. The results for the five vowels uttered by a male American speaker, in Figure 8, show that reasonable area function for voiced sounds could be extracted.

Several limitations of the LP-based technique were described by Sondhi [14] and by Wakita [15]. Use of pre-emphasis does not eliminate the glottal and radiation characteristics, introducing uncertainty in the estimated area functions. One of the error sources in the estimated vocal tract shapes is the assumption of lossless acoustic tube model for the vocal tract. Use of unit area at the glottis introduces errors during dynamically varying vocal tract configuration as the area at the glottis end varies during these transitions. The technique fails during nasals and fricatives due to spectral zeros in the spectral envelope, as these can not be modelled by the all-pole LP analysis. The technique is not suitable during stop closures due to lack of spectral energy. Wakita [15] proposed some solutions to these problems. It has been suggested [16] that the area function estimation can be improved by analysis using a vocal tract model including the losses and estimation of the vocal tract length from the speech signal.

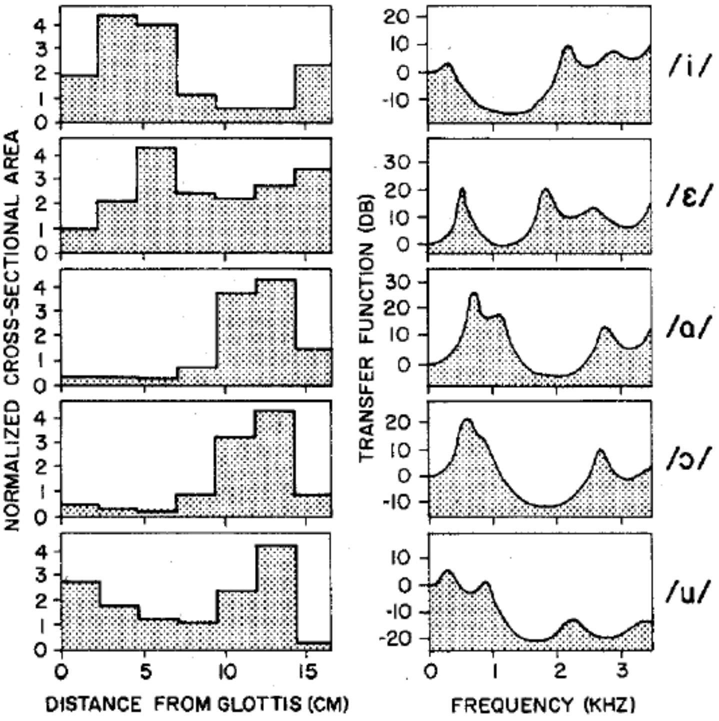

Figure 8. Vocal rtact area functions of five American vowels and their transfer functions estimated by the LP analysis (Wakita [10])

5 Discussion

The research for estimation of the vocal tract area function was motivated by its potential applications in speech coding, speech recognition, speech synthesis, and speech training. Another motivation was to improve the knowledge of speech production by creating acoustic and articulatory data for the articulatory-phonetic investigations. The X-ray techniques were able to obtain the articulatory data, but large datasets could not be collected due to the risk of radiation hazards for the speakers. Therefore, the research efforts were focused on developing techniques to estimate the vocal tract area functions from acoustic measurements or speech signal. Schroeder's technique [8] using the acoustic impedance measurement at the lips showed that reasonable vocal tract area functions could be estimated from the acoustic measurements. Sondhi and Gopinath [9] proposed the technique using measurement of the impulse response at the lips to overcome some of the limitations of the impedance based technique. These techniques based on acoustic measurements were not suited for practical applications due to the need for specialized instrumentation and unnatural speech production. As the estimates from these techniques were not verified with parallel imaging data, they could not be used for creating the acoustic-articulatory database. However, these techniques instilled the confidence that reasonable vocal tract area function can be estimated from the acoustic measurements, despite the non-uniqueness in the acoustic-articulatory mapping. A relatively recent investigation by Qin and Carreira-Perpinán [17] indicated that only 5% of the acoustic data were produced by multiple articulatory configurations, indicating that the speech sounds are mostly produced with unique vocal tract shapes.

In the 1970s, the research for estimating the vocal tract area function was motivated mainly by its potential application in speech coding. It was based on the assumption that the area values change slowly as compared to other parameters and are suitable for decimation and interpolation. They were also expected to be less sensitive to quantization errors. The partial correlation coefficients (PARCOR) proposed by Itakura and Saito [18] were shown to be suitable for transmission as they produce stable all-pole filters even after quantization. Wakita [10] showed that the PARCOR coefficients are negative of the reflection coefficients. The reflection coefficients when close to 1 result in large quantization errors, but the log area ratios obtained from the reflection coefficients using Wakita's method result in good quantization properties [19]. However, line spectral frequencies (LSF) reported by Itakura [20] provide better quantization capability and hence are better suited for speech coding [21]. The main application of Wakita's LP-based technique has been in obtaining vocal tract area functions of non-nasalized vowels for providing a visual feedback for speech training of persons with hearing-impairment [22]–[23].

Recently, several acoustic-articulatory databases have been developed for articulatory-phonetic research [24]–[26]. Availability of these databases has led to research in the techniques based on machine learning for estimating the vocal tract area function [27]–[29]. Magnetic resonance imaging (MRI) has been used to collect the articulatory data as it does not pose radiation hazards [25]. Inspired by the success of deep learning based methods in automatic speech recognition, similar methods are being investigated for estimation of vocal tract shapes [28]. Availability of good-quality parallel acoustic-articulatory databases with a large number of speakers and speech material is likely to facilitate further development of the techniques for estimation of the vocal tract area function for all classes of speech sounds and different articulatory disorders.

6 Acknowledgments

The research was supported by the projects 'National Programme on Perception Engineering,' and 'Visual Speech Training Systems Phase 2', sponsored by the Ministry of Electronics & Information Technology, Government of India.

7 References

[1] Chiba, T. and Kajiyama, M.: *The vowel: Its nature and structure.* Tokyo: Tokyo-Kaiseikan Pub. Co., Ltd., l941.

[2] Dunn, H. K.: *The calculation of vowel resonances, and an electrical vocal tract.* J. Acoust. Soc. Am., vol. 22, pp. 740-753, 1950.

[3] Stevens, K. N., Kasowski, S., and Fant, G. C. M.: *An electrical analog of the vocal tract.* J. Acoust. Soc. Am., vol. 25, pp. 734-742, 1953.

[4] Fant, G. C. M.: *Acoustic theory of speech production: With calculations based on X-ray studies of Russian articulations.* The Hague: Mouton, 1960.

[5] Kelly, J. L. and Lochbaum, C.: *Speech synthesis*. Proc. 4th Int. Congo Acoust., G42, pp. 1-4, 1962.

[6] Flanagan, J. L.: *Speech analysis, synthesis and perception*, New York: Springer-Verlag, 1972.

[7] Arai, T.: *The replication of Chiba and Kajiyama's mechanical models of the human vocal cavity.* J. Phonet. Soc. Jpn., vol. 5, pp. 31–38, 2001.

[8] Schroeder, M. R: *Determination of the geometry of the human vocal tract by acoustic measurements.* J. Acoust. Soc. Am., vol. 41, pp. 1002-1010, 1967.

[9] Sondhi, M. M. and Gopinath, B.: *Determination of vocal tract shape from impulse response at lips.* J. Acoust. Soc. Am., vol. 49, pp. 1867-1873, 1971.

[10] Wakita, H.: *Direct estimation of the vocal tract shape by inverse filtering of acoustic speech waveforms.* IEEE Trans. AU-21, pp. 417-427, 1973.

[11] BORG, G.: *Eine Umkehrung der Sturm-Liouvilleschen Eigenwertaufgabe, (An inversion of the Sturm-Liouville eigenvalue problem)*. Acta Mathematica, vol. 78, pp. 1-96, 1946.

[12] MERMELSTEIN, P.: *Determination of the vocal-tract shape from measured formant frequencies.* J. Acoust. Soc. Am., vol. 41, pp. 1283-1294, 1967.

[13] ATAL, B. S.: *Determination of the vocal tract shape directly from the speech wave.* J. Acoust. Soc. Am., vol. 47, pp. 65(A), 1970.

[14] SONDHI, M. M.: *Estimation of vocal-tract areas: The need for acoustical measurements.* IEEE Trans. Acoust., Speech, Signal Process., vol. 27, pp. 268-273, 1979.

[15] WAKITA, H.: *Estimation of vocal-tract shapes from acoustical analysis of the speech wave: The state of the art.* IEEE Trans. Acoust., Speech, Signal Process., vol. 27, pp. 281-285, 1979.

[16] FANT, G. C. M.: *Speech acoustics and phonetics: Selected writings*. London: Kluwer Academic, 2004.

[17] QIN, C. and CARREIRA-PERPINÁN M. Á.: *An empirical investigation of the nonuniqueness in the acoustic-to-articulatory mapping*. Proc. Interspeech 2007, Antwerp, Belgium, pp. 74-77, 2007.

[18] ITAKURA, F. and SAITO, S.: *On the optimum quantization of feature parameters in the PARCOR speech synthesizer*. Proc. IEEE Conf. Speech Communication and Processing, pp. 434-437, 1972.

[19] MAKHOUL, J.: *Linear prediction: A tutorial review.* Proc. IEEE, vol. 63, pp. 561-580, 1975.

[20] ITAKURA, F.: *Line spectrum representation of linear predictor coefficients of speech signals.* J. Acoust. Soc. Am., vol. 57, S35, 1975.

[21] SOONG, F. AND JUANG, B.: *Line spectrum pair (LSP) and speech data compression*. Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., San Diego, CA, pp. 37-40, 1984.

[22] CRICHTON, R. G. and FALLSIDE, F.: *Linear prediction model of speech production with applications to deaf speech training*. Proc. Inst. Elect. Eng. Control Sci., vol. 121, pp. 865-873, 1974.

[23] PARK, S. H., KIM, D. J., LEE, J. H., and YOON, T. S.: *Integrated speech training system for hearing impaired*. IEEE Trans. Rehabil. Eng., vol. 2, pp. 189-196, 1994.

[24] WESTBURY, J. R.: *X-ray microbeam speech production database user's handbook (version 1.0)*. 1994. Accessed: April, 2019. [online]. Available: www.haskins.yale.edu/staff/gafos_downloads/ubdbman.pdf

[25] NARAYANAN, S., TOUTIOS, A., RAMANARAYANAN, V., LAMMERT, A., KIM, J., LEE, S., NAYAK, K., KIM, Y.-C., ZHU, Y., GOLDSTEIN, L. et al.: *Realtime magnetic resonance imaging and electromagnetic articulography database for speech production research (TC)*. J. Acoust. Soc. Am., vol. 136, pp. 1307-1311, 2014.

[26] WRENCH, A. A. and WILLIAM, H. J.: *A multichannel articulatory database and its application for automatic speech recognition.* Proc. 5th Seminar on Speech Production: Models and Data, Bavaria, pp. 305-308, 2000.

[27] TODA, T., BLACK, A., and TOKUDA, K.: *Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model.* Speech Commun., vol. 50, pp. 215-227, 2008.

[28] WU, Z., ZHAO, K., WU, X., LAN, X., and MENG, H.: *Acoustic to articulatory mapping with deep neural network.* Multimedia Tools Appl., vol. 74, pp. 9889-9907, 2015.

[29] ILLA, A. and GHOSH, P. K.: *Representation learning using convolution neural network for acoustic-to-articulatory inversion.* IEEE Int. Conf. Acoust., Speech and Signal Process. Brighton, United Kingdom, 2019, pp. 5931-5935.