Regional Resonance of the Lower Vocal Tract and its Contribution to Speaker Characteristics

Lin Zhang[1], *Kiyoshi Honda*[1], *Jianguo Wei*[1], *Seiji Adachi*[2]

[1] Tianjin Key Laboratory of Cognitive Computing and Application,
College of Intelligence and Computing, Tianjin University, Tianjin, China
[2] Fraunhofer Institute for Building Physics, Stuttgart, Germany

lzhang.as@gmail.com, khonda@sannet.ne.jp, seiji.adachi@ibp.fraunhofer.de

Abstract

This study attempts to describe a plausible causal mechanism of generating individual vocal characteristics in higher spectra. The lower vocal tract has been suggested to be such a causal region, but a question remains as to how this region modulates vowels' higher spectra. Based on existing data, this study predicts that resonance of the lower vocal tract modulates higher vowel spectra into a peak-dip-peak pattern. A preliminary acoustic simulation was made to confirm that complexity of lower vocal-tract cavities generates such a pattern with the second peak. This spectral modulation pattern was further examined to see to what extent it contributes to generating static speaker characteristics. To do so, a statistical analysis of male and female F-ratio curves was conducted based on a speech database. In the result, three frequency regions for the peak-dip-peak patterns correspond to three regions in the gender-specific F-ratio curves. Thus, this study suggests that, while the first peak may be the major determinant by the human ears, the whole frequency pattern facilitates speaker recognition by machines.

Index Terms: Lower vocal tract, laryngeal cavity, piriform fossa, F-ratio, male-female speaker characteristics

1. Introduction

How individual vocal characteristics emanate in vowel spectra has attracted researchers. Among various factors, static speaker characteristics in vowels have been known to appear in the higher frequency band [1][2]. As a causal anatomical structure, the hypopharynx in the lower vocal tract was pointed out because of its geometrical stability across vowels [3]. The hypopharynx divides into three portions called the laryngeal cavity (LC) and bilateral cavities of the piriform fossa (PF). In a conventional account, the LC generates a closed-tube resonance, while the PFs cause closed-tube anti-resonance(s) [4]. As the result, the LC-peak and PF-dip together shape vowels' higher spectra. According to the ISO equal-loudness contours [5], the human ears are most sensitive to the 3 – 4 kHz frequency range, which closely overlaps with the spectral slope modulated by the LC and PF in male. It is thus suggested that perception of speaker characteristics by humans is brought about in part from such individual spectral variations in the frequency range.

However, acoustic effects of those cavities do not appear constant. The cavities constitute the closed end of the whole vocal tract and open to an expanded area in the lower pharynx, where the cross-sectional area of the region varies from vowel to vowel. Further, the larynx and pharynx are known to demonstrate a marked gender difference in size [6], which must be reflected by differential patterns of the higher spectra across genders. Thus, it may be reasonable to assume that questions for describing static speaker characteristics arise from the complexity of the lower vocal-tract geometry and its acoustic response, as we describe in what follows. The purpose of this study is two-fold. One is to propose a duct model of the lower pharynx that may dissolve the questions in the previous acoustic account. The other is to support the acoustic model by reporting statistical variability of the corresponding frequency sub-bands obtained from a speech database.

2. Previous work and current problems

Our current knowledge on the acoustic effects of the hypopharynx is as follows. The laryngeal cavity (LC) forms a closed tube above the glottis and emanates quarter-wavelength resonance in male [7]. When considering the laryngeal ventricles, the LC resembles a Helmholtz resonator assuming the volume of the ventricles as the resonator cavity, thus slightly lowering its resonance frequency. The spectral peak frequency of the LC is reported at about 3 kHz in male [8]. This frequency region is sensitive to the human ears and contributes to the auditory perception of voice quality. Also, individual variation of LC resonance frequency was thought to signal audible individual characteristics. The piriform fossa (PF) is located above the entrance of the esophagus and forms bilateral closed side branches to the vocal tract [9]. The PF absorbs sound energy of quarter-wavelength resonance, generating one or two anti-resonances in the 4 – 5 kHz region in male. Naturally, the width and depth of the PF are thought to vary across speakers, altering the dip's frequency and depth in spectra. Although spectral dips themselves are less noticeable to the human ears, the PF-dip is found in vowel sounds of many speakers, possibly offering one of the spectral features contributing to automatic speaker recognition [10].

Besides the above seemingly acceptable accounts, recent experimental and simulation results give rise to further questions on the regional resonance of the hypopharynx as listed below.

(1) Second peak of the regional resonance

The resonance of the hypopharyngeal cavities does not only found in the two frequency regions noted above. Available experimental and simulation data suggest that the PF generates an extra resonant peak at a frequency higher than the dip(s) [9][11][12]. The dip-peak pattern is equally visible in the above reports by comparing two spectra obtained with and without the PF. A natural question arises as to how the peak is derived by the side branch of the PF. A simple side branch opening to a

simple duct generates a dip alone. The apparent effect of the dip-and-peak pattern by the PF is hardly explained by a transmission line model of a duct with a side branch.

(2) Gender-related spectral variations

As noted above, the LC-peak and PF-dip(s) determine the higher spectral pattern in vowels. In contrast, in available data from acoustic experiments using solid vocal-tract replicas, this typical peak-dip spectral pattern is not clearly observed in female. Certainly, the shift of the whole pattern to the higher frequencies is seen, but the dip-peak pattern is reduced in amplitude and smeared over frequencies [13][14]. The lack of the spectral salience in female may not be accounted for only by the shortness of the LC and PF.

(3) Vowel-to-vowel variations

The hypopharynx is relatively stable in shape during speech, and this observation supported the notion that the hypopharynx is the origin of static speaker characteristics [4][15]. However, the existing data demonstrate vowel-to-vowel variations both in geometry [16] and spectra [13][14]. The shape of the lower vocal tract immediately above the LC and PF may be reflected by such higher spectral variations.

3. Model of the lower vocal tract

3.1. Anatomy

The hypopharynx is not a simple duct but a multi-compartment section of the vocal tract. Figure 1(a) is a lateral profile of the lower part of the vocal tract (i.e., the hypopharynx and lower part of the mesopharynx). This sketch was simplified from overlaid tracings on sagittal 3D MRI obtained from an adult male during production of vowel /a/. In the anatomical definition, the hypopharynx includes the laryngeal cavity (LC), piriform fossa (PF), and vallecula. The LC forms a short conduit above the glottis, which opens into the lower part of the mesopharynx with its open end at the aryepiglottic fold. The PF in the figure is shown for the longer side of the bilateral cavities. Anatomically, the vallecula and PF are partially bounded by the lateral glossoepiglottic folds, but this structure is not shown in the figure. In the traditional view, the LC is part of the main vocal tract, while the PF and vallecula are side branches to the tract. More realistically, however, the LC opens into the expanded lower pharynx near at the center, where the PF also opens, as shown in Figure 1.

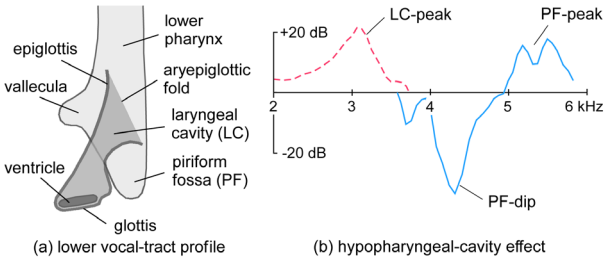

Figure 1. *Anatomy and acoustic response of the lower pharynx. (a) Lateral profile of the lower vocal tract, and (b) Spectral effect of the lower pharynx measured as the difference between conditions with and without cavities derived from an acoustic experiment[17]*

Figure 1(b) is an example of the frequency response of the hypopharynx based on an acoustic experiment using an MRI-based solid vocal-tract replica for vowel /a/ obtained from the same male speaker [17]. The two curves indicate the differences between spectra measured in two conditions with and without the corresponding cavities. When those cavities are removed from the vocal tract in the experiment, then the peaks and dip(s) disappear from the spectra. In the condition with the cavities, the LC generates a peak resonance (or Helmholtz resonance) at around 3 kHz, and the PF causes one or two dips (anti-resonances) at 4 – 5 kHz. Besides those known observations above, the PF is also found to contribute a resonance peak at 5 – 6 kHz as shown in Figure 1. This extra peak has been observed in the results of the previous reports by others [9][11][12], but it was not mentioned presumably because a simple side-branch only causes a spectral dip having no peak from an account of duct acoustics.

3.2. Modeling

A model-based acoustic simulation was attempted in search of the causal mechanism of the second peak of the PF (PF-peak). Figure 2(a) shows a simplified duct model of the lower vocal tract according its geometry in Figure 1(a). This model is an analogy from a simple duct silencer having an expansion chamber and inlet extension inside [18]. In the figure, the side cavity corresponds to the PF, and the inlet extension is part of the LC. In this model, the inlet is a short closed tube differing from a long open duct in a common silencer. The acoustic characteristics of such a duct complex are new to our knowledge, since such a silencer duct with an expansion chamber is reported to have varied patterns of transmission losses depending on the geometry of the duct components.

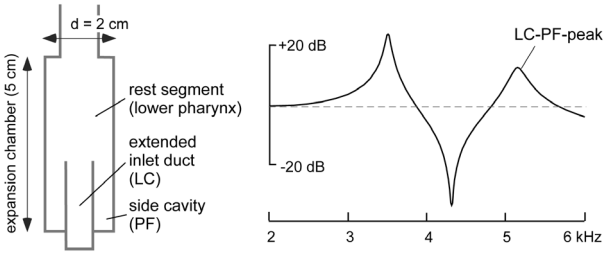

Figure 2. *Model geometry and acoustic simulation. (a) Duct model of the lower vocal tract and (b) Result of duct model simulation showing a peak-dip-peak transfer function.*

Figure 2(b) depicts a result of our preliminary simulation of an acoustic response of the duct complex in Figure 2(a). In this simulation, the peak at 3.5 kHz is due to the resonance of the LC, and the dip at 4.3 kHz is generated by the side cavity of the PF. The second peak (labeled LC-PF-peak) at 5.15 kHz is derived from the LC and PF components of the duct.

This simulation was conducted to obtain a transfer function with a second peak from the duct segment in Fig. 2(a). The simulation technique basically agrees with the usual transmission-line matrix model, where losses due to friction and heat-exchange with the vocal tract wall were taken into account (but with no wall-yielding effect). However, two extra conditions were assumed. The first is no sound reflection at the open end to dismiss formants originating from the main vocal tract, and the other is a treatment of air in the lower pharynx expansion as an air cushion (having stiffness only). The second is analogous to the treatment of the bilateral PFs as “a coupled

two-oscillator system" [19]. The authors explain that anti-resonance dips generated by two PFs of varied dimensions can be assimilated due to their acoustic coupling mediated by the air surrounding the two open ends of the PFs. By adopting the same treatment, our simulation demonstrates that the LC and PF resonate together in the same phase at the peak frequency of 5.15 kHz in Fig. 2(b), thus accounting for the second peak that was questioned in this study.

Acoustic characteristics of the lower vocal tract are discussed above based on our re-consideration on the regional anatomy and acoustic modeling. The LC and PF contribute a peak and zero(s) as seen in previous studies. In addition, the two cavities together generate an extra peak at the higher frequency. This peak is mediated by the air in the lower pharynx in such a way that is unexplained by the plane-wave account on vocal-tract resonance. The frequency of the peak appears to reflect a full-wavelength resonance in the expansion chamber shown in Fig. 2(a). To summarize, the whole lower vocal-tract segment generates a peak-dip-peak pattern in the transfer function, and complex acoustic pattern can be observed in the three frequency regions of the LC peak, PF dip, and LC-PF peak.

4. Database analysis for F-ratio

Above considerations on the resonance pattern of the lower vocal tract suggest a hypothesis: Static speaker characteristics may be observed in the three frequency regions that demonstrate the peak-dip-peak pattern in the higher spectra. In what follows, this hypothesis is tested through a statistical analysis for obtaining discriminative frequency divisions. To do so, mean frequency curves of the Fisher's F-ratio for male and female speakers are computed from a gender-balanced subset of the TIMIT database.

4.1. Speech Database and Spectral Feature

The TIMIT database [20] was used to compute the Fisher's F-ratio. The database was collected from eight dialectal regions of the United States, including ten short sentences for each speaker, and recorded under clean conditions. The original database is unbalanced for the gender with the larger number of male speakers. To compensate for the gender unbalance in the database, a subset of the database is composed for this analysis by reducing male data while keeping the age and dialect variations similar between male and female datasets. The total number of speakers is 354 (177 for each gender).

Figure 3 shows the procedure to obtain the F-ratio. Our method is identical to that in [10]. The whole speech signals (sampled at 16 kHz) were processed with windowing (25 ms), shifting (10 ms), and FFT (512-point). The squared FFT spectra were passed through 60 overlapping triangular bandpass filters in the linear frequency scale to obtain an array of log-amplitude subband energy. Then, the mean subband energies were used to compute the F-ratio.

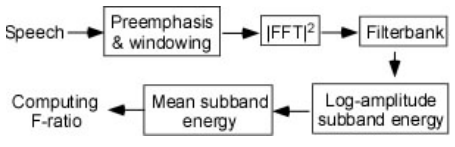

Figure 3. *Procedure to obtain the mean subband energy.*

4.2. Fisher's F-ratio

The Fisher's F-ratio is widely used to measure acoustic features' discriminative ability for a classification task. In previous reports [10][21], outputs of nonlinear filter-banks were tested with the F-ratio analyses with the results showing improved speaker discriminative performance. In our study, F-ratio values for 60 filter outputs from the database are used to obtain the discriminability curves. The responsible frequency bands are estimated for male and female data separately.

The F-ratio is the ratio of the variance between samples to the variance within samples. In this study, the samples correspond to the mean subband energy values of speakers. The F-ratio is defined as:

$$F-ratio = \frac{\frac{1}{M}\sum_{i=1}^{M}(u_i-u)^2}{\frac{1}{M \bullet N}\sum_{i=1}^{M}\sum_{j=1}^{N}(x_i^j-u_i)^2}\ , \qquad (1)$$

where x_i^j is the j_{th} sample feature vector of speaker i with $i = 1, 2, ..., M$, and $j = 1, 2, ..., N$. u_i and u are the average vectors for speaker i and for all the speakers, respectively. They are defined as

$$u_i = \frac{1}{N}\sum_{j=1}^{N} x_i^j, \quad u = \frac{1}{M}\sum_{i=1}^{M} u_i \ . \qquad (2)$$

Equation (1) obtains an F-ratio vector of the inter-speaker variance to intra-speaker variance in each filter band. In this analysis, the larger F-ratio value suggests rich speaker discriminative information.

4.3. Results

Analysis of the F-ratio for vocal-tract resonance features served two roles: (1) comparing the discrimination ability in male and female spectra, and (2) determining the F-ratio-based frequency divisions. Figure 4 shows male and female F-ratio curves. Comparing those two curves in the figure, the F-ratio is generally greater for male than for female, which are also seen in [21] [22]. Both F-ratio curves show a trend of rise toward the higher frequencies, agreeing with [15][23]. It may be that this trend reflects the resonance pattern of the lower vocal tract. Notably, the male F-ratio is the highest at around 5.5 kHz, while the female's is around 7.2 kHz. Considering the male-female difference in the dimension of the piriform fossa (PF) [13], the highest peaks in the F-ratio curves (at 5.5 kHz for male and 7.2 kHz in female) appear to correspond to the peaks of male and female PF resonances. This result suggests that the LC-PF peak region could be the major acoustic component elucidating salient speaker characteristics in spectral analyses.

Based on the visual comparison of the male and female curves in Figure 4, the whole frequency region was divided into five subdivisions adjusted for male and female spectra separately. As shown by the vertical lines in the figure, boundaries for the divisions are set at dips of the F-ratio curves. Three most discriminative frequency subdivisions are labeled in then figure: the LC-peak (resonance of laryngeal cavity), the PF-dip (anti-resonance of the piriform fossa), and LC-PF-peak (resonance of the lower pharynx). Possible explanations for those parts are given next.

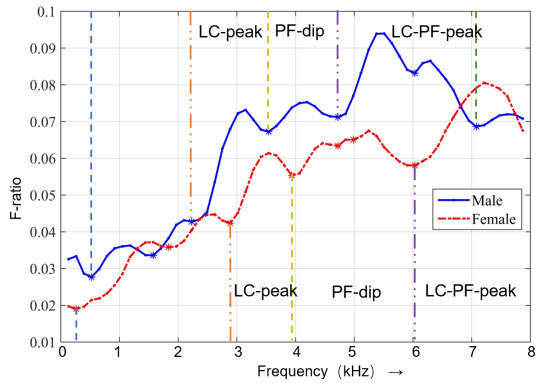

Figure 4 *F-ratio curves for male and female data in the gender-balanced subset of the TIMIT database.*

In the first two frequency subdivisions (unlabeled, below LC-peak), both male and female F-ratio curves suggest less speaker discriminative information, as previously suggested in [24]. The first subdivision is relevant to voice fundamental frequency (F0), showing relatively poor discrimination. The second subdivision includes the lower vowel formants, and this frequency band contributes to phonetic discrimination but not to speaker discrimination. The third frequency subdivision (LC-peak) corresponds to the frequency region of the LC resonance proposed in [14]. As seen in our previous data, the resonance is sharper in male at about 3 kHz, while it is less obvious in female, both characterizing spectral variations in the higher frequency. This observation agrees with the male-female difference of the F-ratio in this frequency region. The fourth subdivision (PF-dip) corresponds to the region including the spectral dip(s) caused by the anti-resonance of the PF. The fifth subdivision (LC-PF-peak) is the region for the in-phase resonance by the LC and PF, which was discussed earlier in Section 2 as a type of the couple two-oscillator mechanisms.

As seen above, the male-female difference of the F-ratio curves reveals gender-specific discriminative frequency regions. The discriminative frequency divisions in Fig. 4 are located in the higher and wider frequencies in female. The F-ratio values also differ between male and female in those frequency regions. Above observations on the F-ratio curves conform to our hypothesis: Static speaker characteristics may be observed in the three frequency regions that demonstrate the peak-dip-peak pattern.

5. Discussions

This study was initiated by long-standing three key questions on speaker characteristics that gave rise from our earlier work. The following questions mentioned in Section 1 are partially answered in this study, yet leaving room for further studies.

Question 1 was the peak-dip-peak spectral patterns that were observed as an acoustic effect of the lower vocal tract. This question was to closely understand realistic vocal-tract acoustic phenomena that may contribute to generating speaker characteristics. Previous studies describe that the LC generates a peak, and the PF gives rise to dip(s). However, the second peak in the higher spectra was underexplored. In this study, the second peak was interpreted by a peculiar regional resonance in the lower pharynx involving in-phase resonance of the LC and PF.

Question 2 was how to interpret male-female differences in higher spectra that are derived from male-female anatomical differences of the lower pharynx. Anatomical records on the question are still scanty to suggest gender-related spectral differences. An analysis in this study based on a speech database showed evidence that the F-ratio curves rise toward higher frequencies in both male and female, where the two curves show a clear disparity of peaks between male and female datasets. The observation suggests us finding why automatic speaker recognition performs poorer for female speech. But, this question must be handled by future studies exploring gender-related spectral characteristics for the effect of the quality factor (Q) of the resonance in the lower vocal tract.

Question 3 was regarding vowel-to-vowel variations in observed acoustic response of the lower vocal tract. By comparing available spectral contrasts with and without the LC and PF, the peak-dip-peak pattern in higher spectra is more evident in vowel /a/ than in vowel /i/. The question remains for future studies assuming a possibility of involving certain unknown mechanisms in the vocal tract.

6. Conclusion

This study is our first step to clarify spectral features determining speaker characteristics. The current results suggest the followings for the future work. For Question 1, the regional resonance in the vocal tract has been underestimated in the past, but our preliminary simulation result indicates that the resonance gives a greater influence than ever thought to the lower frequencies. When considering the speech to vocal-tract inversion, this acoustic effect may play a critical role. For Question 2, the male and female F-ratio curves suggest that higher frequency ranges contribute to enhancing speaker discrimination ability when the male and female are treated separately. The importance of higher frequencies for speaker recognition may be accounted for by the peak-dip-peak resonance pattern discussed in this study. Many questions still remain as to how to recover speakers' vocal tracts by inversion for the purpose of speaker recognition by machines.

7. Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 61876131, U1936102, No.61573254), and National Key Research and Development Plan (No.2018YFC0806802).

8. References

[1] Furui, S. & Akagi, M. (1985) Perception of voice individuality and physical correlates. *Tech. Rep. Hear. Acoust. Soc. Jpn.* H85-18, 1-8.

[2] Kitamura, T. & Akagi, M. (1995) Speaker individualities in speech spectral envelopes. *J. Acoust. Soc. Jpn. (E)*, 16: 283-289.

[3] Kitamura, T., Honda, K., & Takemoto, H. (2005) Individual variation of the hypopharyngeal cavities and its acoustic effects. *Acoustical Science and Technology,* 26: 16-26.

[4] Honda, K. (2008). Physiological processes of speech production. In J. Benesty, M. Sondhi & Y. Huang (Eds.) *Springer Handbook of Speech Processing* (pp. 7-26). Springer, Berlin, Heidelberg.

[5] Suzuki, Y. & Takeshima, H. (2004) Equal-loudness-level contours for pure tones. *J. Acoust. Soc. Am.* 116: 918-933.

[6] Negus, V. (1949) *Comparative Anatomy and Physiology of the Larynx.* Heinemann

[7] Sundberg, J. (1987) *The Science of the Singing Voice.* Northern Illinois Univ. Press.

[8] Takemoto, H., Adachi, S., Kitamura, T., Mokhtari, P., & Honda, K. (2006) Acoustic role of the laryngeal cavity in vocal tract resonance. *J. Acoust. Soc, Am.,* 120: 2228-2238.

[9] Dang, J. & Honda, K. (1997) Acoustic characteristics of the piriform fossa in models and humans. *J. Acoust. Soc. Am.,* 101: 456-465.

[10] Lu, X. & Dang, J. (2008) An investigation of dependencies between frequency components and speaker characteristics for text-independent speaker identification. *Speech Communication,* 50: 312-322

[11] Mokhtari, P., Takemoto, H., & Kitamura, Y. (2008) Single-matrix formulation of a time domain acoustic model of the vocal tract with side branches. *Speech Communication,* 50: 179-190.

[12] Takemoto, H., Honda, K., Masaki, S., Shimada, Y., & Fujimoto, I. (2006) Measurement of temporal changes in vocal tract area function from 3D cine-MRI data. *J. Acoust. Soc. Am*., 119: 1037-1049.

[13] Zhang, C., Honda, K., Zhang, J., & Wei, J. (2016) Contributions of the piriform fossa of female speakers to vowel spectra. *ISCSLP 2016*, October 17-20, Tianjin, China.

[14] Li, J., Honda, K., Zhang, J., & Wei, J. (2016) Individual difference and acoustic effect of female laryngeal cavities. *ISCSLP 2016*, October 17-20, Tianjin, China.

[15] Kitamura, T., Honda, K., & Takemoto, H. (2005) Individual variation of the hypopharyngeal cavities and its acoustic effects. *Acoustical Science & Technology,* 26: 16-26.

[16] Zhang, J., Honda, K., Wei, J., & Kitamura, T, (2019) Morphological characteristics of male and female hypopharynx: A magnetic resonance imaging-based study." *J. Acoust. Soc. Am.,* 145: 734-748.

[17] Honda, K., Kitamura, T., Takemoto, H., et al. (2010). Visualisation of hypopharyngeal cavities and vocal-tract acoustic modelling. *Computer Methods in Biomechanics and Biomedical Engineering,* 13(4), 443-453.

[18] Alfredson, R.J. & Davies, P.O.A.L. (1971) Performance of exhaust silencer components. *J. Sound Vib.,* 15: 175-196.

[19] Takemoto, H., Adachi, S., Mokhtari, P., & Kitamura, T. (2013) Acoustic interaction between the right and left piriform fossae in generating spectral dips. *J. Acoust. Soc. Am.,* 134: 2955-2964.

[20] Zue, V., Seneff, S., & Glass, J. (1990). Speech database development at MIT: TIMIT and beyond. *Speech Communication*, 9: 351-356..

[21] Orman, Ö. D. & Arslan, L. M. (2001). Frequency analysis of speaker identification. *In 2001: A Speaker Odyssey-The Speaker Recognition Workshop*. Crete, Greece.

[22] Reynolds, D. A. (1995). Speaker identification and verification using Gaussian mixture speaker models. *Speech Communication*, 17(1-2), 91-108.

[23] Kitamura, T. & Akagi, M. (1995) Speaker individualities in speech spectral envelopes. *J. Acoust. Soc. Jpn.,* (E) 16: 283-289

[24] Rabiner, L. & Juang, B.H. (1993) *Fundamentals of Speech Recognition.* Prentice Hall PTR.