JOINT ESTIMATION OF VOCAL TRACT AND NASAL TRACT AREA FUNCTIONS FROM SPEECH WAVEFORMS VIA AUTO-REGRESSION MOVING-AVERAGE MODELING AND A POLE ASSIGNMENT METHOD

Shang-Hsuan Peng, Chao-Wen Li, and Yi-Wen Liu

Dept. Electrical Engineering, National Tsing Hua University, Hsinchu, 30013, Taiwan

ABSTRACT

Nasal resonance is utilized in certain languages to differentiate word meanings. The joint filtering effect by the vocal tract and the nasal tract can be modeled by the auto-regression moving-average (ARMA) approach. However, unlike all-pole (i.e., AR) modeling, it has been difficult to derive the equivalent vocal-tract area function directly from an ARMA model due to the nonlinear nature in the relation between model coefficients and vocal-tract geometry. In this paper, we propose a method to decompose an ARMA model approximately into $\alpha/C(z) + \beta/D(z)$; in our context, $1/C(z)$ and $1/D(z)$ represent the filtering effects of the oral and the nasal tract, respectively. Once the decomposition is performed, equivalent oral-tract and nasal-tract area functions can be obtained by converting $C(z)$ and $D(z)$ to their respective lattice representation. The proposed method was applied to non-nasalized and nasalized vowels produced by three speakers, and it was found that the ratio $r = \beta/\alpha$ tends to be higher in nasalized vowels than in their non-nasalized counterparts. The vocal-tract area function estimated by the present approach was also fairly stable for sustained vowels.

Index Terms— Speech, ARMA modeling, nasalization, vocal-tract area function

1. INTRODUCTION

Speech coding has been a well-studied field in signal processing. Nowadays, speech signals are routinely encoded via linear prediction-based algorithms [1] for mobile-phone communication. Modern speech encoders mostly exploit all-pole modeling, and for any stable all-pole model (i.e., all the poles are located inside the unit circle) there is an equivalent acoustic tube model [2]. The correspondence can be established via the lattice representations [3]; the key technique is to find the *reflection coefficients* of the lattice sequentially via a step-down procedure such as the Levinson-Durbin method [4]. The reflections in the lattice can be regarded as if they are caused by impedance mismatch between cascaded acoustic tubes. Thus the vocal tract cross sectional area, as it varies along the airway, can be derived.

The procedure mentioned above does not easily generalize if nasal resonance is also considered. Nasal resonance changes the transfer function into the form of a pole-zero model $H(z) = B(z)/A(z)$. Though it is possible to derive optimal $A(z)$ and $B(z)$ from speech signals via ARMA approaches [5, 6], converting an ARMA model into an equivalent tube model is not straightforward. Lim and Lee [7] attempted to convert any given pole-zero model (i.e., ARMA model) into an acoustic tube model that consisted of three branches representing the pharynx, the oral tract, and the nasal tract, respectively. However, the lossless assumption imposed a constraint that $B(z)$ had to be symmetric. The constraint was later removed so the tube model was allowed to be lossy [8]. Lim and Lee verified that the reflection coefficients in the revised tube model could be effectively estimated from synthetic speech that the model produced.

However, Lim and Lee's approach [8] relied on a simplifying assumption that the mouth is shut so the sounds radiate from the nose. Because of this assumption, the reflection coefficients corresponding to the oral tract did not reside in the expression of $A(z)$, the denominator of the transfer function. More recently, Huang et al [9] attempted to relax the assumption so sound could radiate from both the lips and the nostrils in the tube model. However, the reflection coefficients consequently coupled to both $A(z)$ and $B(z)$ in a nonlinear fashion, and approximate solutions were difficult to find. In the present research, we avoided solving the nonlinear coupled equations. Instead, the poles are assigned to the oral tract (OT) and the nasal tract (NT) first, and then the reflection coefficients and the area functions of OT and NT can be derived via a standard step-down procedure. The proposed method is described in Sec. 2, experimental results are presented in Sec. 3, and discussion and conclusions follow.

2. METHODS

In this section, we shall first give a brief introduction to ARMA modeling and review an iterative approach proposed by Schnell and Lacroix [6]. After a step-down procedure, a revised ARMA model is obtained, and its poles are assigned to two components $C(z)$ and $D(z)$ that represent the OT and the NT, respectively. Then, the cross-section areas of the OT and the NT can be derived.

2.1. ARMA modeling

Given a discrete-time signal $x(n)$, ARMA modeling looks for linear coefficients $\{a_1, ..., a_N\}$ and $\{b_1, ..., b_M\}$ so as to minimize the variance of the modeling error $e(n)$ defined as follows,

$$e(n) = x(n) - \sum_{k=1}^{N} a_k x(n-k) - \sum_{k=1}^{M} b_k e(n-k). \qquad (1)$$

Taking the z-transform of Eq. (1), a transfer function $H(z)$ for the ARMA model is obtained,

$$H(z) = \frac{B(z)}{A(z)} \triangleq \frac{\sum_{k=0}^{M} b_k z^{-k}}{1 - \sum_{k=1}^{N} a_k z^{-k}} \equiv \frac{X(z)}{E(z)}, \qquad (2)$$

Thanks to the Ministry of Science and Technology of Taiwan for funding under Grant No. 102-2220-E-007-020.

where $b_0 = 1$, and we follow the convention that the uppercase $X(z)$ and $E(z)$ denote the z-transform of the corresponding lowercase signals $x(n)$ and $e(n)$, respectively. Applying Parseval's theorem, the energy ε of $e(n)$ can be expressed as

$$\varepsilon = \sum_{n=-\infty}^{\infty} |e(n)|^2 = \frac{1}{2\pi}\int_{-\pi}^{\pi} \left| \frac{1}{H(e^{j\omega})} \cdot X(e^{j\omega}) \right|^2 d\omega. \quad (3)$$

In the present research, Schnell and Lacroix's iterative approach [6] is adopted to find $\{a_1, ..., a_N\}$ and $\{b_1, ..., b_M\}$ such that ε is minimized. The approach is summarized in Figure 1; the key idea is to apply Burg's method [4, 10] alternately and refine the estimate of $A(z)$ and $B(z)$ in every iteration. The process continues until the the present modeling error $\varepsilon[i]$ cannot be reduced further, i being the iteration step number. Note that the difference between path 1 and path 2 in Fig. 1 is whether to find $A(z)$ given $B(z)$ first, or in the opposite order. In every step, it is hoped that at least one of the paths lead to a reduced modeling error. If both paths successfully reduce the modeling error (i.e., when $\varepsilon', \varepsilon'' < \varepsilon[i]$), Schnell and Lacroix suggested to update the ARMA model by setting $A(z)$'s *reflection coefficients* k_j to be $k_j = (k'_j + k''_j)/2$, where $\{k'_1, ..., k'_N\}$ and $\{k''_1, ..., k''_N\}$ are the reflection coefficients of $A'(z)$ and $A''(z)$ in their respective lattice-structure representations [2, 4]. Similarly, a separate set of reflection coefficients are determined for $B(z)$. Therefore, the condition $|k_j| < 1$ is satisfied ($j = 1, ..., N$), which guarantees that the updated model would remain stable at every iteration [11].

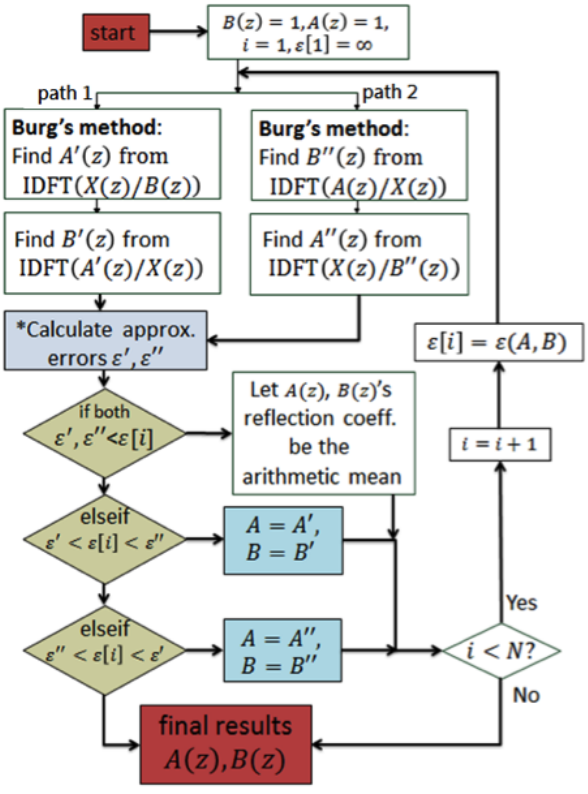

Fig. 1. The iterative ARMA algorithm adopted from Schnell and Lacroix [6]

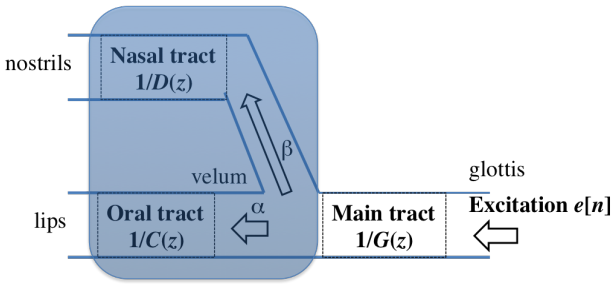

Fig. 2. Modeling the filtering effect of the upper respiratory airway according to the form of Eq. (5). The transfer function for the shaded area is $B(z)/\tilde{A}(z)$.

2.2. Three-branch modeling and a step-down procedure

Figure 2 illustrates the geometry three-branch modeling for the upper respiratory airway. When an excitation signal $e(n)$[1] is produced at the glottis, Fig. 2 suggests that $e(n)$ is first filtered by the main tract's transfer function $1/G(z)$, then the resulting signal splits into the OT and the NT, which are characterized by transfer functions $1/C(z)$ and $1/D(z)$, respectively, with unknown scaling factors α and β. The orders of $G(z)$, $C(z)$, and $D(z)$, denoted as N_G, N_C, and N_D respectively, should be proportional to the lengths of the tracts. For instance, $N_G = 2f_s l/c$, where f_s denotes the sampling rate, l denotes the length of the main tract, and c is the speed of sound [3]. In the present research, $f_s = 16$ kHz, and the orders are set as $N_G = 7$, $N_C = 8$, and $N_D = 10$ based on the typical length of each tract for humans.

Assume that the transfer function of the entire system is $H(z) = B(z)/A(z)$ given by ARMA modeling. We apply a step-down procedure to $A(z)$ [12],

$$A_{n-1}(z) = \frac{A_n(z) - k_n z^{-n} A_n(1/z)}{1 - k_n^2}; \quad (4)$$

the procedure is initialized at $n = N$ where $A_N(z) \triangleq A(z)$, and it repeated for N_G times (i.e., from $n = N$ to $N - 6$). The resulting polynomial $A_{N-7}(z)$, denoted as $\tilde{A}(z)$, is regarded as the AR part for the joint filtering effect of the OT and the NT.

2.3. Pole assignment and transfer function decomposition

Next, we attempt to perform the following decomposition,

$$\frac{B(z)}{\tilde{A}(z)} = \frac{\alpha}{C(z)} + \frac{\beta}{D(z)} = \frac{\alpha D(z) + \beta C(z)}{C(z) \cdot D(z)}. \quad (5)$$

However, given $B(z)$ and $\tilde{A}(z)$, Eq. (5) might not have an exact solution in general. Here, we propose to find an optimal decomposition by exhaustively looking at all possible ways of writing $\tilde{A}(z)$ as the product of two polynomials $C(z) \triangleq \sum_{k=0}^{N_C} c_k z^{-k}$ and $D(z) \triangleq \sum_{k=0}^{N_D} d_k z^{-k}$. Denote the roots of $\tilde{A}(z)$ as $\{z_1, ..., z_{N-N_G}\}$, and without loss of generality assume that the first N_C of them are assigned to $C(z)$. Then, $C(z)$ can be calculated as follows,

$$C(z) = \prod_{k=1}^{N_C} \left(1 - z_k z^{-1}\right), \quad (6)$$

[1] here we intentionally re-use the notation $e(n)$ for excitation because the modeling error has been treated as the excitation signal in the literature of linear prediction-based speech synthesis.

and similarly $D(z) = \prod_{k=N_G+1}^{N-N_G}(1 - z_k z^{-1})$. Given any arbitrary combination of $C(z)$ and $D(z)$, we define the parameters $\mathbf{p} = [\hat{\alpha}, \hat{\beta}]^T$ to be the ones that minimizes a target function $J(\cdot)$,

$$\left[\hat{\alpha}, \hat{\beta}\right]^T = \arg \min_{\alpha,\beta} J\big(\alpha, \beta\,; C(z), D(z)\big), \tag{7}$$

where J is defined as follows:

$$J = \frac{1}{2\pi}\int_{-\pi}^{\pi} \left|B(e^{j\omega}) - \alpha D(e^{j\omega}) - \beta C(e^{j\omega})\right|^2 d\omega. \tag{8}$$

In practice, J is more straightforward to calculate in the time domain:

$$J \equiv \sum_{n=0}^{M} \left|b_n - \alpha d_n - \beta c_n\right|^2, \tag{9}$$

where we implicitly assume that $B(z)$ and $\alpha D(z) + \beta C(z)$ have the same order (as polynomials of z^{-1}). Thus, Eq. 7 is a least-square problem and $\mathbf{p}$ can be found by standard pseudo-inverse approaches; let us define $\mathbf{b} = [b_0, ..., b_M]^T$, $\mathbf{c} = [c_0, ..., c_M]^T$, and $\mathbf{d} = [d_0, ..., d_M]^T$.[2] Then, the solution to Eq. 7 is given as follows,

$$\mathbf{p} = (Q^T Q)^{-1} Q^T \mathbf{b},$$

where $Q = [\mathbf{d}|\mathbf{c}]$ is a matrix of size $(M+1) \times 2$. When performing the factorization $\tilde{A}(z) = C(z)D(z)$, we require that the coefficients c_k and d_k must all be real. Equivalently, this means that each pair of conjugate roots of $\tilde{A}(z)$ should stay together after root assignment.

Finally, the optimal decomposition is achieved by choosing the combination that globally minimizes J,

$$\{C^*(z), D^*(z)\} = \arg \min_{C(z),D(z)} J\left(\hat{\alpha}, \hat{\beta}; C(z), D(z)\right) \tag{10}$$

subject to the constraint that $C(z) \cdot D(z) = \tilde{A}(z)$. The optimal scaling parameters α^* and β^* are simultaneously determined by substituting $C^*(z)$ and $D^*(z)$ back to Eq. (7).

2.4. Deriving the effective cross-sectional areas

The polynomials $C^*(z)$ and $D^*(z)$ each corresponds to a lattice structure characterized by reflection coefficients. Ways of converting from filter coefficients to reflection coefficients can be found in [11]. Denote the coefficients as $\{\mu_1^d, ..., \mu_{10}^d\}$ for $D^*(z)$, $\{\mu_1^c, ..., \mu_8^c\}$ for $C^*(z)$, and $\{\mu_1^g, ..., \mu_7^g\}$ for $G(z)$. Then, the reflection coefficients determine the cross-section area (CSA) ratios S_{m+1}/S_m between adjacent segments of tubes in the following manner [2],

$$\frac{S_{m+1}}{S_m} = \frac{1-\mu_m}{1+\mu_m}, m = 1, ..., N_G \text{ (or } N_C, N_D). \tag{11}$$

Here, increasing m by 1 means moving one step into the throat toward the glottis. To visualize the results, we set the following boundary conditions: $S_1^d = 1.1$ cm^2 at the nostrils, $S_8^g = 2.5$ cm^2 at the glottis, and $S_9^c = S_1^g$ at the velum.

3. TESTING THE METHOD ON SPEECH SIGNALS

In this section, we first present results of CSA estimation for sustained vowels. Then we compare the results of transfer function decomposition for nasalized vs. non-nasalized vowels.

[2]Because the NT is longer than the OT for humans, we have $N_C < N_D = M$, and the highest $M - N_C$ coefficients in $\mathbf{c}$ are set to zero.

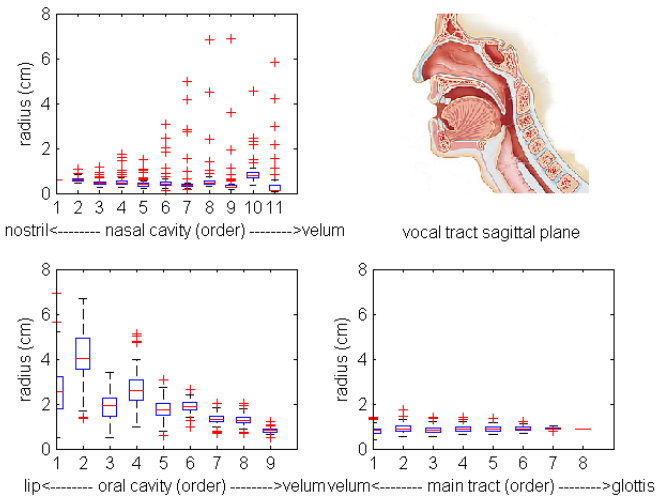

Fig. 3. Tukey's box plot of CSAs derived from a recording of the vowel /a/. The top-left, bottom-left, and the bottom-right panels show results for the nasal tract, the oral tract, and the main tract, respectively. The "radius" is defined as $\sqrt{\text{CSA}/\pi}$. Statistics were obtained from 79 frames of length 50 ms.

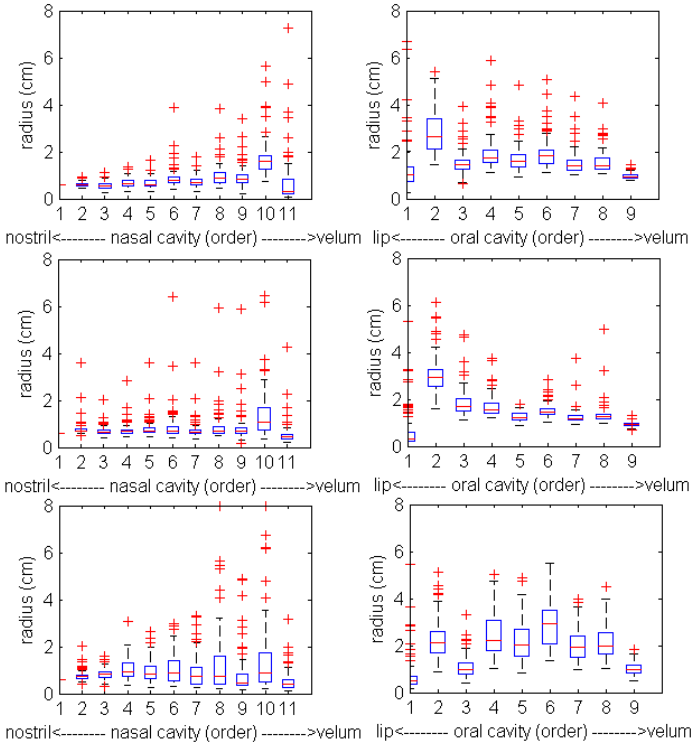

Fig. 4. CSAs of the nasal tract and the oral tract derived from recordings of the vowels /o/ (top row), /u/, and /i/ (bottom row), respectively. Statistics were obtained from 79 frames of length 50 ms.

3.1. Sustained vowels

A male speaker was recruited to record his production of the vowels /a/, /o/, /u/, and /i/. The speaker was instructed to sustain the vowel for about three seconds, and the middle 2-second portion of the signal was saved for analysis. Figure 3 shows results of the estimated CSAs from one recording of /a/, displayed using Tukey's boxplot [13]. In particular, each box shows the inter-quartile range with a median line, and every + mark represents an outlier. For this particular recording, the estimated main-tract CSAs were most consistent across frames. The estimated CSAs for the NT had more variation, as depicted by an increased number of outliers in the plot.

Figure 4 shows the results of estimated nasal-tract and oral-tract CSAs for vowels /o/ (top row), /u/ (mid), and /i/ (bottom) produced by the same speaker. The CSAs for the main tract were quite similar across the four vowels so we choose not to show them here.

Comparing the CSAs for the OT across the four vowels, /a/ had the largest radius near the lips at segment #2, followed by /o/ and /u/, and the radius for /i/ was the smallest. Comparing /o/ and /u/, the over-all profile of their radius functions were similar, and the main difference was at segment #1 right at the lips where /u/'s radius is smaller than that of /o/. The profile for /i/, being a front vowel, had surprisingly large CSAs near segments 4, 5, and 6 when compared to /u/, the back vowel that has approximately the same height. This might partially be due to the fact that /i/ in Mandarin is tense so the speaker pulled his cheeks toward both sides when producing it. Consequently, the CSAs near segments 4 to 6 was enlarged even though the tongue position might be high at the moment.

The CSAs for the NT do not have a clear contrasting profile between different vowels. This agrees with the common sense that the NT does not change its shape when different vowels are produced.

3.2. Nasalization

Three male speakers were recruited to produce non-nasalized vowels /a/, /i/, and /e/ and their nasalized counterparts, denoted as /a$^{\rm n}$/, /i$^{\rm n}$/, and /e$^{\rm n}$/, respectively. Two of the speakers are native speakers of *Min Nan* (a Chinese language spoken in Fujian, Taiwan, and Singapore). The other speaker does not speak Min Nan but could understand it. These vowels were chosen because their nasalization can differentiate word meanings; for instance, /wa/ means "me" while /wa$^{\rm n}$/ means "a bowl". So it was expected that a native speaker of Min Nan should be able to produce the contrast with ease.

The speakers were instructed to produce each non-nasalized vowel first and, after taking a brief breath, switch to the nasalized counterpart while holding the shape of their oral cavity as much as they could. Generally this should involve lowering the velum to enhance the nasal resonance. Each speaker produced each vowel for five times, and the recording was conducted inside a sound booth where the noise floor was about 30 dB SPL (sound pressure level).

Signals were partitioned into overlapping 50-ms frames, and the algorithm described in Sec. 2.3 was applied to obtain $C^*(z)$ and $D^*(z)$ in Eq. (10). Then, we counted the frames for which $\beta^* > \alpha^*$. Results are summarized in Table 1. Note that $\beta = 0$ would reduce the transfer function in Eq. (5) to an all-pole model, meaning that the sound does not enter the nasal cavity at all. In general, we can expect that the coefficient β in Eq. (5) should be higher for nasalized vowels than for non-nasalized vowels.

Results in Table 1 suggest that, while in average nasalized vowels had a higher chance to have a larger β^* than α^*, the condition did not always hold — across all vowels recorded in the present research, some nasalized frames ended up having $\alpha^* > \beta^*$, and vice versa. Nevertheless, across all vowel positions and all three speakers, the general tendency shown in Table 1 is consistent with the expectation that the β value should be higher for nasalized vowels than for non-nasalized vowels.

Table 1. Percentage of frames with $\beta^* > \alpha^*$. The total frame number $N = 395$ for each nasalized or non-nasalized vowel.

speaker 1	/a/	/i/	/e/
non-nasal.	31.6%	31.3%	18.4%
nasalized	69.5%	45.2%	43.6%
speaker 2	/a/	/i/	/e/
non-nasal.	27.9%	19.8%	27.9%
nasalized	70.6%	55.0%	45.2%
Speaker 3	/a/	/i/	/e/
non-nasal.	41.3%	37.7%	38.3%
nasalized	71.3%	67.2%	45.7%

4. DISCUSSION AND CONCLUSIONS

In principle, nasalized vowel should be produced by lowering the velum, thus allowing the sound waves to propagate into the nasal cavity. We could have presented the statistics of the estimated CSA near the velum to see if present results support this hypothesis. However, we fell short of doing so because wave scattering at the junction of the three tracts in Fig. 2 could be modeled more accurately. In general, the scattering coefficients should relate to the area ratios between the three tracts and, due to our favor of simplicity, this has not been considered yet. Future research along this direction is warranted.

Outliers in Figs. 3 and 4 indicate that the present method for CSA estimation might be sensitive to small fluctuations in the outcome of ARMA modeling. Perhaps the requirement $C(z)D(z) = \tilde{A}(z)$ can be relaxed a little so roots of $\tilde{A}(z)$ can be jittered for the purpose of minimizing the target function J. This is also a possible future research direction.

The main contribution of the present research would likely be the idea of transfer function decomposition in Eq. (5). Through exhaustive search among all combinations of $C(z)$ and $D(z)$, the global optimum decomposition can always be found, and effective CSAs of the OT and the NT can be derived via lattice representation. The computation time for searching through all possible ways to perform root assignment $\tilde{A}(z) = C(z)D(z)$ turns out to be not so demanding for $N_C = 8$ and $N_D = 10$. If a higher sampling rate (e.g., 44.1 kHz) is used, the order of the all-pole models would become higher, and the algorithm for optimal decomposition may consequently need to be improved.

To summarize, in this research a novel way is proposed to transform any ARMA-based speech production model into a three-branch waveguide model. Globally optimal parameters could be obtained through exhaustive search, and the method has been tested on nasalized and non-nasalized vowels, in particular produced by speakers of the Min Nan language. In the future, the proposed method could be developed into an OT and NT visualization tool so as to help people master the skill of nasalization when learning a new language.

5. REFERENCES

[1] M. R. Schroeder and B. S. Atal, "Code-excited linear prediction (CELP): High-quality speech at very low bit rates," in *Proc. Int. Conf. Acoust., Speech, and Signal Process.*, 1985, vol. 10, pp. 937–940.

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

[3] J. D. Markel and A. H. Gray, "Linear prediction of speech," in *Techniques in Speech Acoustics*. Springer Verlag, New York, 1976.

[4] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, no. 4, pp. 561–580, Apr. 1975.

[5] S. M. Kay, *Modern Spectral Estimation*, Prentice-Hall, Englewood Cliffs, NJ, 1988.

[6] K. Schnell and A. Lacroix, "Pole zero estimation from speech signals by an iterative procedure," in *Proc. Int. Conf. Acoust., Speech, and Signal Processing, Salt Lake City, USA*, 2001, vol. 1, pp. 109–112.

[7] I.-T. Lim and B. G. Lee, "Lossless pole-zero modeling of speech signals," *IEEE Trans. Speech Audio Process.*, vol. 1, no. 3, pp. 269–276, Jul. 1993.

[8] I.-T. Lim and B. G. Lee, "Lossy pole-zero modeling for speech signals," *IEEE Trans. Speech Audio Process.*, vol. 4, no. 2, pp. 81–88, Mar. 1996.

[9] H.-K. Huang, Y.-W. Liu, and R. P.-Y. Chiang, "Detection of obstructive sleep apnea by estimation of oral and nasal cavity cross-section areas from acoustic recordings of snore," *Proc. Meetings on Acoustics*, vol. 19, no. 060172, pp. 1–7, 2013.

[10] J. P. Burg, *Maximum entropy spectral analysis*, Ph.D. thesis, Stanford University, Stanford, CA, USA, 1975.

[11] A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Procssing*, Pearson-Prentice Hall, Boston, 3rd edition, 2010.

[12] Julius O. Smith, *Introduction to Digital Filters with Audio Applications*, W3K Publishing, 2007.

[13] J. W. Tukey, *Exploratory Data Analysis*, Addison-Wesley, 1977.