Comparative analysis of autoregressive models for linear prediction of ultrasonic speech

Farzaneh Ahmadi[1], *Ian V. McLoughlin*[2], *Hamid. R. Sharifzadeh*[3]

[1,2,3] School of Computer Engineering, Nanyang University, Singapore
[1] ahmadi@ntu.edu.sg

Abstract

Ultrasonic speech is a novel research area with significant applications: as a speech-aid prosthesis for patients with voice box difficulties, silent speech interfaces, secure mode of communication in mobile phones and as a communication medium in high noise industrial environments.
Feature extraction is a critical part of the ultrasonic speech system. Linear prediction analysis (LPA) has been recently proven to be viable for extracting features from the three dimensional ultrasonic propagation in the vocal tract (VT). A one-dimensional autoregressive model based on averaging the LP coefficients, analysed in different recording positions has been investigated by the authors to fit the LF ultrasonic resonances of the VT.
To reach a state of maturity for the LPA of ultrasonic speech and in continuum of the previous work, this paper compares the application of two major conventional methods of averaging and least squares error - already applied in room acoustics - for deriving the coefficients in autoregressive modelling of ultrasonic speech.

Index Terms: ultrasonic speech, linear prediction analysis, autoregressive modelling, averaging, least squares

1. Introduction

Ultrasonic speech is defined as "a system which augments a low frequency (LF) ultrasound excitation to the human voice production mechanism as a substitute or supplement to the glottal excitation and extracts feature sets from the resulting ultrasonic output to be used in several tasks including conversion to the audible speech, recognition, enhancement and communication" [1]. The applications of the technology include speech generation for voice loss patients, secure mode of communication when one needs his talk over a mobile phone while not being heard by a third party and communication in high noise environments such as factories.
Although the term ultrasonic speech has been coined 20 years ago [2], the field has been a neglected research area, being reintroduced recently by the research activities of the authors [3]. The present team has already established the mathematical and physical basics of ultrasonic speech analysis, and has proven that a slightly modified version of linear predictive analysis which is already applied in room acoustics [4] can be used for extracting features out of ultrasonic output of the vocal tract. To reach a state of maturity, linear prediction of ultrasonic speech urges a comparative analysis of major linear prediction methods of room acoustics when being applied to the analysis of ultrasonic propagation inside the VT. Since to the knowledge of the authors such comparison was not present in the room acoustics literature, this paper analyses the major autoregressive modeling approaches of averaging and least squares in the continuum of the previous work [5]. The result of this analysis will aid the choice of autoregressive model of ultrasonic speech.

2. Autoregressive analysis of ultrasonic speech

Physical analysis of the system shows that ultrasonic propagation inside the VT has several similarities to the audible propagation. The negligible dispersive effects of Carbon dioxide in exhaled air of the VT and small frequency dependant attenuation in the frequency range of this application (less than 100 kHz) permit the linear source filter modeling of the system [1]. The major difference with audible propagation however, is that ultrasound propagation inside the vocal tract can no longer be considered to be majorly along the axis of the VT. Due to the smaller wavelength compared to audible sound, the ultrasonic wave can propagate in three dimensions inside the tract.
Emitting out of the VT, the ultrasonic output is captured in front of the mouth. The ultrasonic speech system then extracts feature sets out of this signal for processing and conversion to audible domain. Linear prediction is a major tool for feature extraction of audible speech, which is based on a one dimensional autoregressive model of the transfer function of the VT. Since the propagation of sound is not majorly axial in the VT, the one-dimensional model cannot be applied directly to the three dimensional VT transfer function. Extension of linear prediction to the three dimensional propagation of ultrasound in the VT was proven to be viable in the previous work of the authors [5].
This paper will implement and compare the two major methods of averaging and least square error, already applicable to room acoustics, in evaluation of linear prediction coefficients of ultrasonic speech. The results of this paper will complete the previous work, which introduced linear predictive analysis of ultrasonic speech using averaging the AR coefficients to derive common resonances of the VT for ultrasonic propagation. A discussion of the stability of both methods is also necessary to reach a final decision.
According to acoustic scale modeling [6], methods of room acoustics for audible domain may be applicable to ultrasonic speech. Linear prediction methods in room acoustics are majorly based on the common acoustical poles modelling of the transfer function of a room [4]. Prior to implementing major common poles autoregressive modelling approaches, the viability of the presence of common acoustic poles in the ultrasonic transfer function of the VT should be investigated.

2.1. Common acoustical poles in ultrasonic transfer function of the VT

Having infinite small ultrasonic waves, propagating inside the vocal tract, and limiting the ultrasonic bandwidth to values below 100 kHz (to discard the dissipative and dispersive

behaviors of ultrasound in the air bounded by VT volume), ultrasonic propagation inside the vocal tract can be described in the linear lossless acoustic domain [1]. This is explained using the three dimensional Helmholtz equation, in the frequency domain:

$$(\nabla^2 + \mathrm{k}^2)p(r,\omega) = 0 \quad (1)$$

Where, ω is the angular frequency of the sound wave, $\boldsymbol{r}$ is the spatial coordinates vector, $k = \frac{\omega}{c}$, c is the speed of sound and $p(r,\omega)$ is the Fourier transform of pressure signal ($p(r,t)$). In any enclosure (e.g. the vocal tract) bounded by boundaries with Dirichlet and Neumann conditions, equation (1) has no solutions except for discrete values of $k = k_n$ (eigen-value) and their corresponding solutions of $p_n(r)$ (eigen-functions). The transfer function of the enclosure $H(\omega, r_s, r_o)$ between a point source having frequency ω located at position r_s and a receiver located at r_0 is then derived as [6]:

$$H(\omega, r_s, r_o) = C\frac{\sum_{i=1}^{\infty} P_i(r_s)P_i(r_o)j\omega}{(\omega^2 - \omega_i^2 - 2\delta_i\omega_i)} \quad (2)$$

where C , is a constant and $P_i(r)$ is any of the eigen-functions of the enclosure, ω_i are the eigen-frequencies and δ_i are their corresponding damping consta[1]. The independence of the denominator of the transfer function form the location of source and receiver (r and r_0) and its dependence on the eigen-frequencies (δ_i and ω_i), is observed in (2) and is also reflected in the Z domain representation of (3) [4].

$$H(z, r_s, r_o) = \frac{p(z,r_o)}{p(z,r_i)} = \frac{B(z,\mathbf{r_S},\mathbf{r_O})}{A(z)} \quad (3)$$

Where A(z) denotes the polynomial formed by the set of common poles and $B(z, r_s, r_o)$ denotes the dependence of the zeros to source and receiver locations.

The vocal tract geometry can be simply modeled as being bounded by boundary definitions of Dirichlet and Neumann conditions (zero pressure in front of the mouth; hard walls elsewhere). As a consequence of (2), the three dimensional ultrasonic transfer function of VT between a point source and a receiver will have a set of common poles which do not depend on the locations of the source and receiver and thus are common to all transfer functions for various positions of source and the receiver points.

The transducer source being used in ultrasonic speech, is however distributed over a range in the space. As a conclusion of superposition, this source can be approximated, as an integration of several point sources $S_i(\mathbf{r_{S_i}}, z)$, $1 \le i \le N$ each located at position r_{S_i} distributed on the surface of the transducer. Each source signal then propagates out of the VT to one of the output points and on its path, it will face a transfer function with a set of common poles but different zeros to produce the output.

Based on the previous formulations [1], the model can be simplified further with a source having uniform spatial distribution ($S(r,z) = S(z)$) on the transducer surface and the ultrasonic speech process can be modeled by (4):

$$H(z, \mathbf{r}_j) = \frac{\sum_{i=0}^{Q} b_i(\mathbf{r}_j)z^{-i}}{1 + \sum_{i=1}^{P} a_i z^{-i}} \quad (4)$$

where a_i denotes the polynomial coefficients incorporating the common poles while $b_i(\mathbf{r}_j)$ $1 \le j \le M$ determine the different zeros between source and any of the M receiver points located at $\mathbf{r}_j = \mathbf{r_{O_j}}$. Figure 1 is a and the simplified diagram of the autoregressive model.

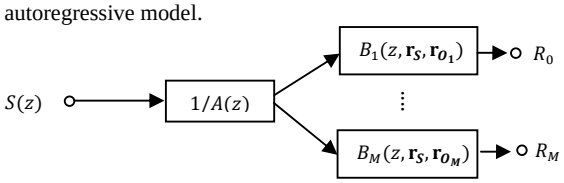

Figure 1: *Approximate model of ultrasonic propagation inside the VT with uniform spatial distribution on the transducer*

2.2. Autoregressive modelling of ultrasonic speech using common acoustic poles

Averaging the AR coefficients for multiple output points and least squares error [4], are the major approaches used in room acoustics to evaluate the set of common resonances which will be applied in the linear prediction of ultrasonic speech here.

2.2.1. Least squares error

Presented by Haneda [4], this method evaluates the coefficients a_i and b_i in a least squares approach to minimize the cost function J:

$$J = \sum_{j=1}^{M}\sum_{n=0}^{\infty} e^2(\mathbf{r}_j, n) \quad (5)$$

Where n is the discrete time and M is the number of output points used for estimation (each output point j is located at $\mathbf{r}_j$) and e can be described is the "equation error" between the modeled and the actual transfer function:

$$e(\mathbf{r}_j, n) = h(\mathbf{r}_j, n) - \sum_{i=1}^{P} a_i h(\mathbf{r}_j, \boldsymbol{n} - \boldsymbol{i}) \quad (6)$$

Because autoregressive modelling is meant in this work, Q (the order of zeros) in (4) is considered to be equal to zero.

2.2.2. Averaging

An alternative to the above method is minimizing the cost function (7) at each receiver point (to fit an autoregressive model to the transfer function at each point) and taking the average of the coefficients $a_i(\mathbf{r}_j)$ resulted for each location. In this case the cost function $J(\mathbf{r}_j)$:

$$J(\mathbf{r}_j) = \sum_{n=0}^{\infty} e^2(\mathbf{r}_j, n) \quad (7)$$

is minimized for each of the M points ($1 \le j \le M$), resulting the auto regressive coefficients $a_i(\mathbf{r}_j)$ (Q is again considered to be zero in (4)). An average of the coefficients $a_i(\mathbf{r}_j)$ for the M points results the AR coefficients of the model:

$$a_i = \sum_{j=1}^{M} a_i(\mathbf{r}_j) \quad (8)$$

Both methods result the autoregressive modeling of the transfer function (4) as:

$$\hat{H}(z, \mathbf{r}_j) = \hat{H}(z) \approx \frac{1}{1 + \sum_{i=1}^{P} a_i z^{-i}} \quad (9)$$

3. Comparative analysis of the autoregressive models

To compare the effectiveness of the two methods a finite element analysis of ultrasonic propagation inside the VT is provided to have more flexibility on the choice of measurements points.

[1] $k_n = \frac{\omega_n}{c} + \frac{\delta_n}{c}$ where ω_n is the Eigen-frequency and δ_i is a damping constant. c is the speed of the sound.

3.1. Finite element analysis

A unit-delay Kelly Lochbaum model consisting 22 same-length tubes [5] is used to model vocal tract geometry of English vowels. The area functions of the vowels were provided from Baer [7]. The impedance value of VT walls, was considered to be equal to the mean acoustic impedance of soft tissues in human body ($1.7*10^6$ Rayl). Two measuring points with a distance more than one wavelength were selected and the ultrasonic frequency responses were evaluated for the range of 20-25 kHz. These were converted to time domain to evaluate the autoregressive coefficients using averaging and the least squares methods. The results of this simulation follow.

3.2. Simulation results

Figure 2, shows the results of the averaging and the least square methods being applied on a finite element model for two points within VT configuration for vowel /a/, for ultrasonic frequency range of 20-25 kHz. The transfer function response of the two points is projected in figure 2.a. An eigen-frequency analysis confirms the location of common poles in this figure. Figure 2.b demonstrates the frequency response of the LPC filter whose coefficients are derived by averaging and least square methods. The arrow demonstrates the behaviour of the two methods in extracting a common pole which is being masked by a spatial zero located at one of the points.

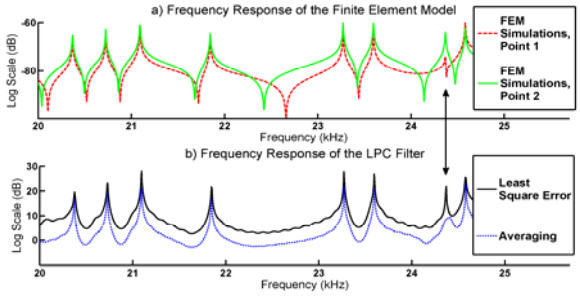

Figure 2: *Linear Prediction using averaging and Least squares methods for extracting the common resonances of vowel /a/ a) Finite element analysis of two receiving points b) LPA using the values of the FEM in the two points*

Figure 2 clearly shows that in given sufficient order, least squares better extracts the uncommon poles. Another simulation was performed to complete the above mentioned result. The aim was to check the robustness of the two methods to the decrease in the order of LPC. Figures 3 summarizes the finite element analysis of the 22 tube model for the vowel /æ/ in two receiver points. The figure demonstrates two less significant poles below 21 kHz.
LPC analysis is performed using the values of this figure by the two methods of least squares errors and averaging with two different orders of 30 and 100. The results are reflected in figure 4.(a,b). As observed in this figure, the averaging method is more robust to the decrease of LPC order and can successfully extract the less significant poles with the decrease of order down to 30.

4. Discussion

The above mentioned results needs to be completed with a mathematical investigation of the problem to observe the efficiency of both methods in terms of extracting the common and uncommon poles and the stability of the resulting LPC filter.

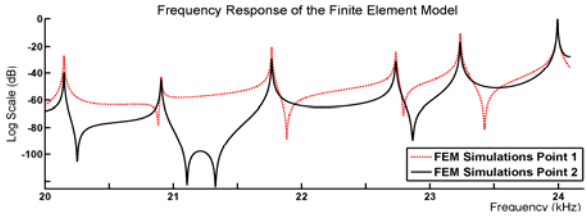

Figure 3: *Finite element analysis of ultrasonic propagation in the VT in the configuration of vowel /@/*

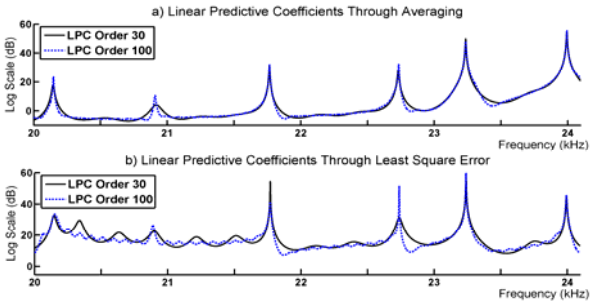

Figure 4: *Linear predictive analysis of ultrasonic output of the VT for vowel /@/ for two orders of 30 and 100 via a) Averaging method, b) Least squares*

4.1. Uncommon poles

Having explained the common poles being present in the analysis of ultrasonic speech in section 2.1, it is necessary to explain that some of the common poles may be eliminated in the output signal by the presence of space dependant zeros. The following small mathematical proof compares the application of the two methods in case of these uncommon poles.

4.1.1 Averaging

Considering two measurement points 1,2, this method uses two autoregressive models to fit the transfer function at each point:

$$H_1(\boldsymbol{r},z) \approx \frac{1}{p(z)},\quad H_2(\boldsymbol{r},z) \approx \frac{1}{q(z)} \tag{10}$$

The denominator polynomials ($p(z)$ and $q(z)$) are not a function of location vector $\boldsymbol{r}$ meaning that the poles of $H_1(\boldsymbol{r},z)$ and $H_2(\boldsymbol{r},z)$ do not depend on the location.

The averaging method builds a new function H so that:

$$\widehat{H}(z) = \frac{1}{a(z)} \tag{11}$$

Where: $a(z) = 0.5 * (p(z) + q(z))$. In case of a common root at $z = z_0$ in $p(z)$ and $q(z)$:

$$p(z_0) = 0\ ,\ q(z_0) = 0 \Rightarrow a(z_0) = 0 \tag{12}$$

So the averaging reflects the common poles of $H_1(\boldsymbol{r},z)$ and $H_2(\boldsymbol{r},z)$. In case of an uncommon pole, the wavelength of the LF ultrasound and the position of the receiver points in front of the mouth permit no more than one real zero or one pair of complex conjugate zeros. For one uncommon real zero at $z = z_0$, all except one of the poles of $p(z)$ and $q(z)$ are similar.

$$q(z) = (z - z_0)p(z) \tag{13}$$

Averaging polynomials $p(z)$ and $q(z)$ in this sense will result:

$$a(z) = 0.5 * (p(z) + q(z)) = 0.5 * p(z) * (z - z_0 + 1) \quad (14)$$

So in this sense, the averaging method still finds a value for the uncommon real zero z_0 but as a shifted value at $z_0 - 1$. For a pair of conjugate poles at z_0, z_0^*, The new pole location will be $\mathrm{Re}(z_0) \pm \sqrt{Im^2(z_0) - 1}$, So the averaging method shifts the location of the uncommon pole. This is in compliance with the simulation results of figure 2 which demonstrated that the averaging method has shifted the value of the uncommon pole.

4.1.2 Least squares error

Least square method discussed here can be considered a generalization of the conventional single point linear prediction [8] for multiple receiver points. As a result it uses (9) to model the transfer function in all the receiver points. If the measured values of the transfer function at each point *j* are denoted by $h(\mathbf{r_j}, n)$, taking the Z transform of (6) will result:

$$E(z, \mathbf{r_j}) = H(z, \mathbf{r_j}) / \hat{H}(z) \quad (15)$$

Consequently

$$H(z, \mathbf{r_j}) = E(z, \mathbf{r_j}) . \hat{H}(z) \quad (16)$$

Minimizing cost function *J* is equivalent to:

$$e(n, \mathbf{r_j}) = \begin{cases} A & n = 0 \\ 0 & n \neq 0 \end{cases} \quad (17)$$

In this case $H(z, \mathbf{r_j})$ and $\hat{H}(z)$ have all the poles in common in all the receiving points, and *J* is minimized. If however we have one set of complex conjugate poles at $z = z_0$ which are absent in one of the receiving points (e.g. point *jj*):

$$\hat{H}(z) = \frac{1}{(z - z_0)(z - z_0^*) p(z)} \quad (18)$$

$$H(z, \mathbf{r_{jj}}) = \frac{1}{p(z)} \quad (19)$$

In this case if $\hat{H}(z)$ models the rest of the receiving points correctly, (17) is valid for all the receiving points except *jj*. For point *jj* using this model:

$$E(z, \mathbf{r_{jj}}) = H(z, \mathbf{r_j}) / \hat{H}(z) = (z - z_0)(z - z_0^*) \quad (20)$$

$$e(n, \mathbf{r_{jj}}) = \begin{cases} |z_0|^2 & n = 0 \\ 2Re(z_0) & n = -1 \\ 1 & n = -2 \\ 0 & \text{elserwher} \end{cases} \quad (21)$$

Consequently,

$$J = (M - 1)A + (1 + |z_0|^4 + 4Re(z_0)^2) \quad (22)$$

This means that the error of modeling in this case will depend on the norm of the uncommon pole and its real part. In other words, the closer the pole is to the unit circle (higher $|z_0|$, i.e. more significant pole), the larger the error of estimation is. This matches the fact that missing the more significant poles in the AR modeling will produce larger errors.

4.2. Stability

The stability of the least squares method is already demonstrated for Q = 0 (considering no zeros) [4]. For the stability of the averaging method, in case of all common poles, if the AR representation in each of receiving points is stable, the stability will be clear. For the case of real uncommon pole mentioned in 4.1.1, in case the uncommon pole is located in the left hand side of the unit circle the -1 shifting causes the resulting pole to exit the unit circle and the resulting filter will be unstable, In case of conjugate poles, the stability is guaranteed for one pair of uncommon poles (again with the precondition of having and stable AR model at each point).

5. Conclusion

This paper presented a comparison of the two methods of averaging and least squares error to evaluate the LP coefficients of ultrasonic speech. This was a necessary step to complete the previous effort in feature extraction of ultrasonic speech.
Based on the provided analysis, least squares and averaging can both extract the common poles of ultrasonic transfer function of the VT. However, least squares, fits an AR model to the set of values of multiple receiving points compared to the averaging, which fits several AR models to the values of the points. Consequently, least squares provides more accuracy extracting common poles which was demonstrated to be true by simulation. Both methods face difficulties estimating the uncommon poles. Averaging shifts the values of the uncommon poles. Least squares performs less error if the uncommon pole is not one of the most significant poles of the transfer function. Averaging demonstrates less sensitivity to the decrease of LPC order and has less Computation costs. The stability of least squares method is guaranteed for AR modeling. However averaging can be considered stable only if the AR modeling at each receiver point produces a stable LPC filter.

6. Acknowledgements

The authors wish to thank Mr. Tan Swee Huat who was one of the people we wished to help with ultrasonic speech technology in return to his kind support of our project. It is our deepest regret that we have missed him forever today, when his goal of kindness is fulfilled but ours of responsibility is still incomplete.

7. References

1. Ahmadi, F. and I.V. McLoughlin, *The use of low frequency ultrasonics in speech processing*, in *Recent Advances in Signal Processing*. 2009, Itech Book Publishers: Vienna, Austria.
2. MacLeod, N., *Non-audible speech generation method and apparatus*, United States Patent No. 4821326, Editor. 1987.
3. Dyball, H., *Talking ultrasound.* IET Electronic Letters, 2010. **46**(6): p. 383.
4. Haneda, Y., S. Makino, and Y. Kaneda, *Common acoustical pole and zero modeling of room transfer functions.* IEEE trans. speech and audio proc., 1994. **2**(2).
5. Ahmadi, F., I.V. Mcloughlin, and H. Sharifzadeh, *Linear predictive analysis of ultrasonic speech.* IET Electronic Letters 2009. **46**(6): p. 387–388
6. Kuttruff, H., *Room acoustics* 5th ed. 2009: Taylor & Francis.
7. Baer, T., et al., *Analysis of vocal tract shape and dimensions using magnetic resonance imaging: Vowels.* The Journal of the Acoustical Society of America, 1991. **90**(2).
8. Cheng, E., I.S. Burnett, and C. Ritz, *Time Delay Estimation of Reverberant Meeting Speech: On the Use of Multichannel Linear Prediction*, in *Third International IEEE Conference on Signal-Image Technologies and Internet-Based System, SITIS '07*. 2007. p. 531 – 537.