ARMA Modelling for Whispered Speech

Xue-li LI(栗学丽), Wei-dong ZHOU(周卫东)

(*School of Information Science and Engineering*, *Shandong University*, *Jinan* 250100, *China*)

Abstract – The Autoregressive Moving Average (ARMA) model for whispered speech is proposed. Compared with normal speech, whispered speech has no fundamental frequency because of the glottis being semi-opened and turbulent flow being created, and formant shifting exists in the lower frequency region due to the narrowing of the tract in the false vocal fold regions and weak acoustic coupling with the subglottal system. Analysis shows that the effect of the subglottal system is to introduce additional pole-zero pairs into the vocal tract transfer function. Theoretically, the method based on an ARMA process is superior to that based on an AR process in the spectral analysis of the whispered speech. Two methods, the least squared modified Yule-Walker likelihood estimate (LSMY) algorithm and the Frequency-Domain Steiglitz-Mcbride (FDSM) algorithm, are applied to the ARMA model for the whispered speech. The performance evaluation shows that the ARMA model is much more appropriate for representing the whispered speech than the AR model, and the FDSM algorithm provides a more accurate estimation of the whispered speech spectral envelope than the LSMY algorithm with higher computational complexity.

Key words – ***ARMA model***; ***AR model***; ***whispered speech***; ***LSMY algorithm***; ***FDSM algorithm***

Manuscript Number: 1674-8042(2010)03-0300-04

dio: 10.3969/j.issn.1674-8042.2010.03.22

1 Introduction

Whispering is a natural but unusual mode of speech. The research on whispered speech rooted in the need for private communication and speaking-aid system for laryngectomees. When people talk in the whispery mode, the glottis is semi-opened, and turbulent flow created by exhaled air passing through this glottal constriction provides a noise source of sound. With no vocal cord vibration, the whispered vowel is not quasi-periodic and has no pitch, but the formants still exist and the first formant (F1) shifts to higher frequency[1]. Since the mechanism of whisper production is different from that of the normal speech, significantly different spectral structure exists between whispered and normal speech.

Information extraction from the short-time speech spectrum with the all-pole AR model has been studied in the past and has achieved good performance in digital transmission, synthesis and recognition of speech. For a more accurate descrip-tion of speech, especially nasal and fricative sounds, the pole-zero ARMA model is proposed[2]. One of the consequences of semi-opening glottis during whispering is an acoustic coupling to the subglottal airways. The subglottal system produces a series of resonances, and these subglottal resonances could introduce additional pole-zero pairs into the vocal tract transfer function from the glottal source to the mouth output[3]. Theoretically, the method based on an ARMA process is superior to that based on an AR process in the spectral analysis of the whispered speech. Thus the representation of whispered speech with the quasi-stationary Autoregressive Moving Average (ARMA) model is proposed in this work.

At present, most of the whispered speech analysis is based on the AR model. An autoregressive model with exogenous input (ARX) was applied for the measurement of the formant frequencies of the whispered vowel[4]. Studies show that there is weak acoustic coupling with the subglottal system during whispering vowels. The ARX model parameters are estimated using the Kalman filtering algorithm. However the stability of estimated IIR filter could not be theoretically guaranteed, especially for the consonantal segment with small amplitudes[5]. No research on ARMA-based whispered speech analysis has been reported so far. In this paper, two methods, the least squared modified Yule-Walker likelihood estimate (LSMY) algorithm and the frequency-domain Steiglitz-Mcbride (FDSM) algorithm, are applied to the ARMA modelling for the whispered speech.

The remainder of the paper is arranged as follows. In section 2 the two ARMA algorithms are introduced. The performance evaluations are presented in section 3. Experimental results are given in section 4. Finally, the conclusions are drawn.

2 ARMA Modelling Methods

The ARMA model has been extensively studied in various fields, such as speech processing, biomedical signal processing, control engineering, economics and others. Numerous methods have been developed for estimating the parameters of an ARMA model.

* **Received**: 2010-06-25

Project supported: This work was supported by the Independent Innovation Foundation of Shandong University (No. 2009JC004) and the Natural Science Foundation of Shandong Province (No. Y2007G31)

Corresponding author: Xue-li LI(lixueli@sdu.edu.cn)

The ARMA (or pole-zero) model for the vocal tract filter of the whispered speech is investigated. It can be described by the rational transfer function

$$h(z) = \frac{B(z)}{A(z)} = \frac{\sum_{n=0}^{q} b_n z^{-k}}{\sum_{n=0}^{p} a_n z^{-k}}, \quad (1)$$

where the a_n are the pole coefficients, b_n are the zero coefficients, p and q are the pole and zero orders, respectively.

The parameter estimation may be obtained by using either non-iterative or iterative methods. There is a trade-off between the accuracy of spectral representation and the computational time. In this paper, the least squared modified Yule-Walker likelihood estimate (LSMY) algorithm is used for the non-iterative method[6-7], and the frequency-domain Steiglitz-McBride (FDSM) algorithm is used for the iterative method[8]. Their performances are compared with that of AR model which uses the classical autocorrelation algorithm.

3 Performance evaluation

The major purpose of this work is to model the vocal tract more accurately by including the effect of the subglottal system. Hence, performances of the methods mentioned above and the AR model are compared on the basis of the estimated speech spectrum.

Two error measures are chosen to evaluate the effectiveness of the estimated ARMA model and AR model. They are the Spectral Distortion (SD) and the Spectral Flatness Measure (SFM)[9]. The SD is used in the case of modelling a speech spectrum with known special envelopes, and is defined by

$$SD = \sqrt{\frac{1}{N}\sum_{n=0}^{N-1}[10\log_{10}\frac{P(n)}{\hat{P}(n)}]^2}, \quad (2)$$

where $P(n)$ is the actual power spectrum of the vocal tract filter obtained by FFT, $\hat{P}(n)$ is the estimated power spectrum using the model method, and N is the number of frequency points (i.e. the order of the required FFT). It has the property that the minimum SD occurs if and only if $\hat{P}(n)$ is identical to $P(n)$.

The Spectral Flatness Measure (SFM) measures the whiteness of the residual error spectrum. The residual signal is obtained by inverse filtering the original speech with the vocal tract filter which has been determined. The SFM is defined by

$$SFM = \frac{\exp[\frac{1}{N}\sum_{n=0}^{N-1}\ln S(n)]}{\frac{1}{N}\sum_{n=0}^{N-1}S(n)}, \quad (3)$$

where $S(n)$ is the power spectrum of the residual signal. If the ARMA model is effective, it should remove the correlated components of the signal, leaving a white residual. That is, the ideal residual error spectrum should be flat, and the SFM should be unity when the residual is completely white.

4 Experiments and results

To illustrate the performance and the effec-tiveness of the ARMA model, some simulations are conducted on the practical whispered speech segments.

4.1 Materials

To cover most classes of Chinese speech, 352 Chinese whispered syllables were spoken separately by 2 males and 2 females with a headset microphone in the laboratory. 1408 (352 * 4) syllables were recorded with the background noises from computers and air-condition. The average Signal Noise Ratio (SNR) is 5 dB. The detailed data are listed in Ref.[10].

These whispered syllables are constructed with 23 initial consonants and 34 final vowels with the Chinese syllable rule. 23 initial consonants include fricative (C1), stop-fricative combination (C2), stop (C3), nasal (C4), liquid (C5), retroflexion (C6) and semivowel (C7). 34 final vowels consist of simple vowels (V1), complex vowels (V2) and compound nasal vowels (V3). The letter and its associate IPA for all the phonemes used in the present study are listed as follows:

1) Fricative (C1): f[f], s[s], sh[ʂ], x[ɕ], h[x];

2) Stop-fricative combination (C2): j[tɕ], z[ts], zh [tʂ], q[tɕʰ], c [tsʰ], ch [tʂʰ];

3) Stop (C3): b[p], d[t], g[k], p[pʰ], t[tʰ], k [kʰ];

4) Nasal (C4): m[m], n[n];

5) Liquid (C5): l[l];

6) Retroflexion (C6): r[ʐ];

7) Semivowel (C7): w[w], y[j];

8) Simple vowels (V1): i[i], ü[y], a[a], o[o], e [ɣ], u[u];

9) Complex vowels (V2): ai[aI], ei[eI], ao[aʊ],ou [oʊ], ia[ia], ie[iɛ], iao[iaʊ], iu[iəʊ], ua[ua], ui[ueI], uo[uo], uai[uaI], üe[yɛ];

10) Compound nasal vowels (V3): an[an], en[ən], ang[aŋ], eng[əŋ], ong[oŋ], ian[iɛn], in[in], iang[iaŋ], ing[iŋ], iong[yŋ], uan[uan], un[uən], uang[uaŋ], üan [yɛn], ün[yn].

These phonemes are obtained by segmenting all the whispered data through the Entropy method[11].

4.2 Methods

The classical autocorrelation algorithm is used to calculate the AR model with 10 poles. The LSMY algorithm and the FDSM algorithm are applied to get the ARMA model with 10 poles and 8 zeros.

4.3 Results

In Tab.1, the average results for all the phonemes

are given. In Fig. 1, the SD results from various methods for all kinds of phonemes are shown in detail. It is seen that the ARMA model is better than the AR model for all the whispered phonemes. For ARMA models, the FDSM algorithm is better than the LSMY algorithm for C1, C2, C3, C5, C7, V3. This is expected since the iterative methods produce optimum poles-zero estimates. The LSMY algorithm is good for the C4, C6, V1, V2. As a whole, the FDSM algorithm is slightly better than the LSMY algorithm on average with higher computational load. The SFM results are also listed in Fig. 2. It gives the conclusion similar to the SD's.

Tab. 1 Average results for all the phonemes

Model		SD (dB)	SFM
AR		5.525 5	0.630 2
ARMA	LSMY	5.109 4	0.660 6
	FDSM	4.995 3	0.668 4

In Fig. 3 the speech spectrum of a Chinese whispered simple vowel /i/ with FFT and the estimated models are shown. The original spectrum of this sound was obtained by computing the Fast Fourier Transform (FFT) of a 20 ms speech signal, with a Hamming window and pre-emphasizer. The fine line represents the original speech spectrum obtained with FFT. AR model and two ARMA models were estimated using the proposed methods. It can be found that the spectral envelopes obtained with the ARMA model fit better for the original speech spectrum than that with the AR model, especially at the formant frequency F1 (350 Hz), F2 (2 800 Hz), F3 (3 370 Hz). The FDSM algorithm is slightly better than the LSMY algorithm.

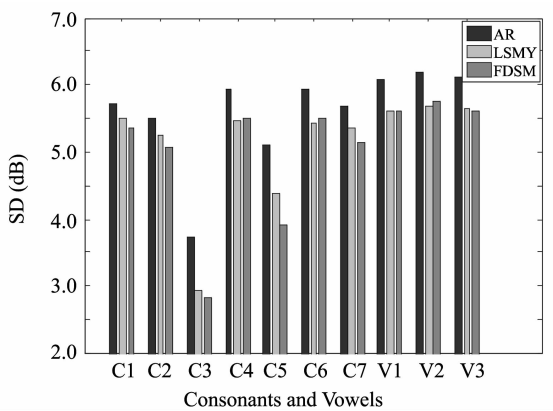

Fig. 1 Histogram of the SD for all kinds of consonants and vowels

Experimental results indicate that the ARMA methods are able to produce better spectral estimation for whispered speech segments in general. This is evident from both the SD and SFM results shown in Fig. 1 and Fig. 2. However, the compu-tational complexity and the number of parameters also increase with ARMA model. The FDSM algorithm can give better representation of the whispered speech than the LSMY algorithm when using ARMA models, but at the cost of considerable increasing in computational load. Since LSMY algorithm involves the least amount of computation and yields quite good estimate, it is promising for incorporating LSMY into low-rate whispered speech coder and real-time voice conversion.

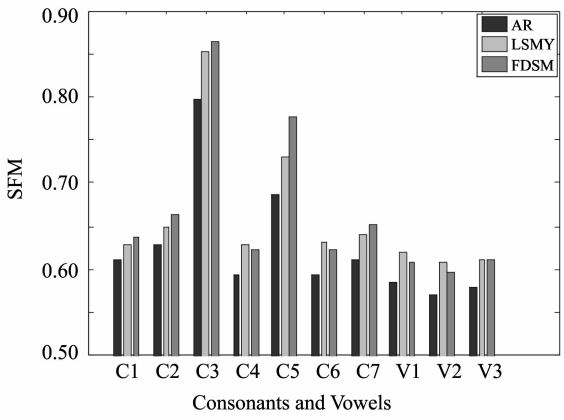

Fig. 2 Histogram of the SFM for all kinds of consonants and vowels

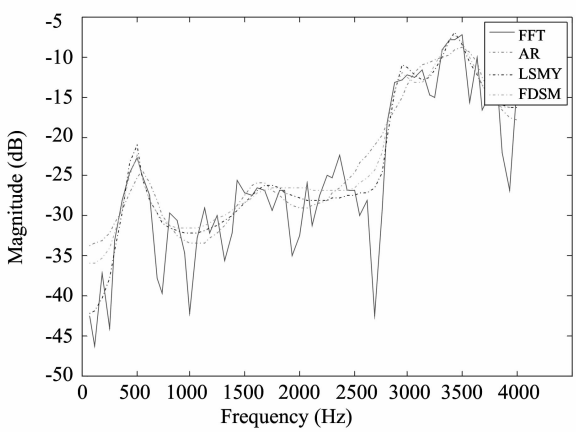

Fig. 3 Estimation of the spectral envelope of a whispered vowel /i/

5 Conclusions

At present, most of the whispered speech analysis is based on the AR model. The ARMA model could be an effective means for precise representation of the whispered speech. In this paper, the experiment results confirmed that the ARMA model is more appropriate for representing the whispered speech than the AR model, and the FDSM algorithm is proved to provide a more accurate estimation of the whispered speech spectral envelope than the LSMY algorithm with higher computational complexity. In the next work, the spectral analysis of whispered speech based on the ARMA model will be studied in detail.

6 Acknowledgments

The authors are sincerely grateful to Dr. Leland B. Jackson, University of Rhode Island, Kingston, USA for his kind and significant help. This work was supported by the Independent Innovation Foundation of Shandong University (No. 2009JC004), and the Natural Science Foundation of Shandong Province (No. Y2007G31).

References

[1] Xue-li Li, Bo-ling Xu, 2005. Formant comparison between whispered and voiced vowels in Mandarin. *Acta Acustica united with Acustica*, 91(6): 1079-1085.

[2] H. Morikawa, H. Fujisaki, 1982. Adaptive analysis of speech based on a pole-zero representation. *IEEE transactions on acoustic speech and signal processing*, ASSP-30(1): 77-88.

[3] K. J. Kallail, F. W. Emanuel, 1984. Formant-frequency differences between isolated whispered and phonated of vowel samples produced by adult female subjects. *Journal of Speech and Hearing Research*, 27(2): 245-251.

[4] M. Matsuda, H. Kasuya, 1999. Acoustic Nature of the Whisper. Proc. EUROSPEECH, p. 137-140.

[5] W. Ding, H. Kasuya, S. Adachi, 1995. Simultaneous estimation of vocal tract and voice source parameters based on an ARX model. *IEICE Trans. Inf. Syst.*, E78-D,(6): 738-743.

[6] S. M. Kay, 1988. Modern spectral estimation: theory and application. Prentice-Hall, Englewood Cliffs, NJ.

[7] D.G. Childers, 2000. Speech processing and synthesis toolboxes. John Wiley & Sons, Inc., New York.

[8] Leland B. Jackson, 2008. Frequency-domain Steiglitz-McBride method for least-squares IIR filter design, ARMA modeling, and periodogram smoothing. *IEEE Signal Processing Letters*, 15: 49-52.

[9] S. Yim. D. Sen, W. H. Holmes, 1994. Comparison of ARMA modelling methods for low bit rate speech coding. Proc. ICASSP, p. 273-276.

[10] Kechu Yi, Bin Tian, Qiang Fu, 2000. Speech Signal Processing. National Defence Industry Press, China, p. 48.

[11] Xue-li Li, Hui Ding, Bo-ling Xu, 2005. Entropy-based initial/final segmentation for Chinese whispered speech. *ACTA ACUS-TICA*, 30(1):69-75.