RETRIEVING VOCAL-TRACT RESONANCE AND ANTI-RESONANCE FROM HIGH-PITCHED VOWELS USING A RAHMONIC SUBTRACTION TECHNIQUE

Zhao Zhang, Kiyoshi Honda, Jianguo Wei[*]

College of Intelligence and Computing, Tianjin University

williezz@163.com, khonda@sannet.ne.jp, jianguo_fr@163.com

ABSTRACT

Vocal tract resonances give rise to core spectral information of speech signals. Linear prediction and cepstral methods are widely used for this purpose. However, both approaches are prone to fail as the fundamental frequency (F0) rises. In this study, a new cepstral method is developed combined with a refined rahmonic subtraction technique (RS-CEPS) to extract spectral envelopes excited by glottal noise sources. A vowel synthesis system based on 3D-printed solid vocal tract models is used to obtain reference transfer functions for accuracy verification. A series of stable vowels /a/ was synthesized for a wide F0 range. By analyzing the synthetic vowels, the results showed that the RS-CEPS yields accurate estimates of resonance-peak and anti-resonance frequencies in comparison to those from the conventional methods. The RS-CEPS is simple and stable, offering a potential for expanding speech analysis applications.

Index Terms— cepstral analysis, high-pitched vowels, rahmonic subtraction, formant estimation, spectral envelope

1. INTRODUCTION

Understanding the nature of speech sounds via spectral analysis has been a long expectation by researchers because acoustic processes of speaking could be deciphered by certain techniques. However, this target appears still in distant, even moving away from us as we discover more on the complexities of vocal tract geometries. The real vocal tract is a complicated conduit containing cavities and branches, and the resonance of the structures modifies vowel spectra in certain frequency regions [1]. In male voices, an extra resonance peak at about 3 kHz is caused by the supraglottic laryngeal cavity. This peak is followed by a trough-and-peak pattern above 4 kHz due to the bilateral piriform fossa. This regional resonance has been known as one of the causal factors of speaker characteristics [2], playing an important role in speaker recognition by humans and machines.

Another longstanding issue is the interference by voice fundamental frequency (F0). The two common approaches described below are known for inaccuracy at high F0.

The LPC offers a parametric envelope estimation based on an all-pole digital filter principle [3]. However, due to the large energy concentrations at the harmonics, the optimized all-pole LPC model results in inaccuracies for high-pitched vowels [4]. Many efforts were paid to the conventional LPC models [5][6] to reduce the influence of F0. Shadle et al. [7] compared several formant estimation methods and drew the conclusion that the LPC-based WLP-AME algorithm [8] reached the best performance among others. However, the LPC-based methods are incapable of describing true vocal-tract resonance with zeros generated by anti-resonances.

The cepstral analysis traces contours of power spectra with poles and zeros [9], thus being capable of extracting both peaks and troughs. The "true envelope" method such as Imai's [10] favors low-F0 vowels, however it fails at high F0 due to sparsely populated harmonics. A hint for improvements is found in FFT power spectra of vowels produced at high F0, as seen in Fig. 1. The spectrum shows a baseline contour that infers a finer pattern of vocal-tract resonance. This is because turbulent airflow noise at the glottis undertakes vocal tract resonance and displays baseline spectral information. In voice production, both periodic signal and random noise are generated at the glottis as the airflow passes by [11]. Instead of estimating the harmonic based spectral outline, tracing the envelope driven by glottal noise in vowels is a way to avoid the interference by F0. Fig. 1 is a spectrum of female vowel /a/ with F0 at 530 Hz. The dashed line is the true envelope obtained by the Imai's method, and the solid line is the noise baseline envelope computed by the "inverted" Imai's method. By comparison, the latter explores the finer spectral detail.

In the frequency domain, harmonic components are mixed up seriously with glottal noise components, which makes it difficult to separate these two components. Contrarily, in the cepstrum domain, the harmonics are transformed into localized prominent regions called "rahmonic". Thus, the disturbing harmonics could be eliminated by a rahmonic subtraction technique. Childers et al. [12] and Randall et al. [13] proposed the use of a notch or comb lifter to remove the rahmon-

This work was supported by the National Key R&D Program of China (No.2018YFC0806802), National Natural Science Foundation of China (No.61573254, No.61876131, No.U1936102). E-mail of corresponding author Jianguo Wei: jianguo_fr@163.com

ics. However, such lifters also remove components of vocal tract resonances elicited by glottal airflow noise.

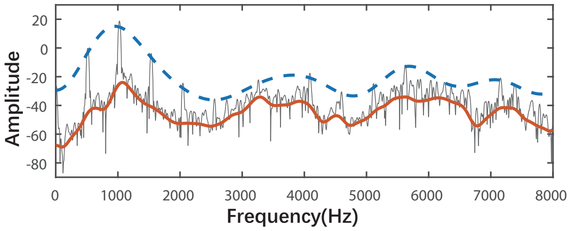

Fig. 1. A spectrum of natural vowel /a/ produced by a female speaker at F0 = 530 Hz with a with a 60-ms window length. The dashed line is the "true envelope" of the harmonic peaks, and the solid line is the envelope of noise baseline.

In this study, we propose an improved cepstral method using finer rahmonic subtraction (RS-CEPS) to separate the harmonics and glottal noise components and then trace the glottal-noise driven envelope. To validate the accuracy, a vowel synthesis system based on 3D-printed solid vocal tract models is employed to obtain natural-sounding vowels. Evaluation of the proposed analysis method is based on the comparison between obtained vowel spectra with varied F0s and the transfer function of the solid model.

2. DEVELOPING THE RS-CEPS METHOD

The RS-CEPS aims at obtaining noise-excited spectra according to the following steps. Firstly, the spectrum, cepstrum and fundamental period are obtained as basic data. Secondly, the rahmonic peaks and boundaries are determined to isolate the rahmonic structure in the quefrency domain. Thirdly, the cepstrum of harmonic components is extracted from the original spectrum, and then it is used to remove harmonics from the original vowel spectrum. With this procedure of rahmonic subtraction, also shown in Fig. 2, the spectral envelope is finally acquired by cepstral liftering for desired spectral resolutions.

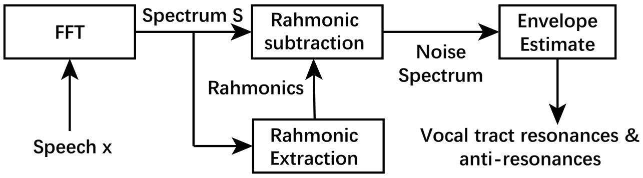

Fig. 2. Illustration of the proposed method.

2.1. Basic data acquisition

As an initial step, the standard log spectrum S and cepstrum C are obtained from vowels. Then the quefrency value of the first rahmonic peak, defined as P_1, is obtained by searching the maximum within the quefrency range $(fs/F0_{max}, fs/F0_{min})$, where fs is the sample rate, and $F0_{max}$ and $F0_{min}$ are the upper and lower limits of $F0$ search, respectively.

2.2. Rahmonic extraction

A smoothed spectral envelope E is obtained by cepstral liftering using $fs/F0_{max}$ as the lifter order, which approximatively segmented harmonic and noise components. The spectrum S_H above E is extracted from S by

$$S_H = max(S - E, 0). \tag{1}$$

The purpose of this step is to extract the upper spectrum that contains harmonic components alone so that the cepstrum C_H with a clear rahmonic structure could be obtained through S_H. This step is illustrated in Fig. 3.

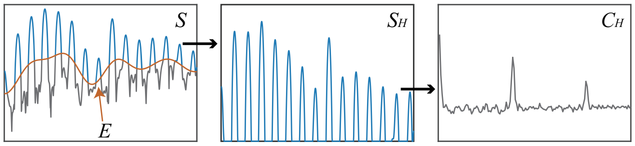

Fig. 3. Illustration of the upper spectrum extraction.

Boundaries of rahmonic structures are acquired by following steps:

i. Calculate the squared cepstrum $C_E = {C_H}^2$.

ii. Locate the rahmonic peaks. The locations of the m_{th} rahmonic peaks P_m are given by

$$P_m = max(C_E(n)), P_{m-1} < n < P_{m-1} + (3/2) * P_1 \tag{2}$$

Since P_1 is obtained in 2.1, $P_2, ...P_m$ could be located in the squared cepstrum C_E one by one.

iii. For each rahmonic peak P_m, the left/right boundaries L_m/R_m are extracted as the nearest left/right dips of P_m.

Fig. 4 shows detected rahmonic boundaries. Boundaries of the first three rahmonics are segmented by black points.

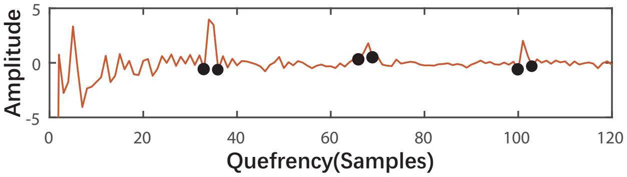

Fig. 4. Cepstrum from a vowel frame with segmented rahmonics.

2.3. Rahmonic subtraction

The harmonic components are subtracted from the original data by the following steps in the quefrency domain.

i. The upper harmonic components S_H from the original spectrum S are extracted using Eq. (1).

ii. The spectra S_H and S are transformed into the corresponding cepstra C_H and C, and a new cepstrum is obtained by subtracting the corresponding rahmonics of C_H from C:

$$C_r(k) = \begin{cases} C(k), & k \notin (L_1, R_1) \cup ... \cup (L_m, R_m) \\ C(k) - C_H(k), & k \in (L_1, R_1) \cup ... \cup (L_m, R_m) \end{cases} \tag{3}$$

where (L_m, R_m) is the quefrency range of the m_{th} rahmonic extracted in 2.2.

iii. Spectrum S_r after rahmonic subtraction is obtained from the subtracted cepstrum C_r, and the step i and step ii are repeated on S_r as a new iteration to further eliminate the residual harmonic components.

After 3 to 5 iterations, the subtraction results C_r and S_r are close to convergence. Fig. 5 shows (a) the cepstra before and after rahmonic subtraction and (b) the corresponding spectra. It is clear that harmonics are removed, and the new spectrum reveals a detailed pattern of vocal tract resonance.

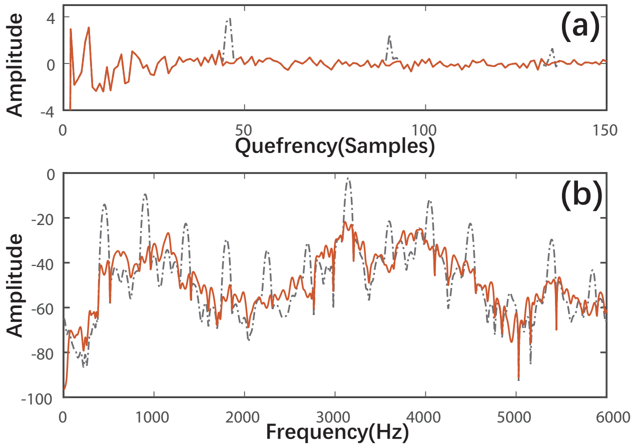

Fig. 5. Effect of rahmonic subtraction, showing (a) the cepstrum in dash-dotted line and cepstrum after subtracting the first three rahmonics in solid line, and (b) the FFT power spectrum in dash-dotted line and rahmonic-subtracted spectrum in solid line.

2.4. Envelope estimation

Since the noise spectrum S_r reveals the true vocal tract information, the spectra envelope will be obtained by certain envelope extracting techniques. In this study, a simple cepstral liftering method is adopted on the glottal noise cepstrum C_r. The liftering window is given by

$$g(n) = \begin{cases} 1, & n \leq L \\ 0.5(1 + cos[\pi(n - L/\Delta L)], & L \leq n < L + \Delta L \\ 0, & n > L + \Delta L \end{cases} \tag{4}$$

where $L = L + \Delta L$ is the length of the lifter window. Since harmonic components were already removed, a simple liftering with a pair of constant L and ΔL suffices at any F0. The frequency values of the formants and dips are extracted manually based on the envelope.

3. MATERIALS AND EXPERIMENT

3.1. Acoustic materials

Synthetic vowel sounds obtained from a solid vocal tract are used to evaluate the accuracy of the proposed method.

3.1.1. Solid vocal tract vowel synthesis system

The solid vocal tract models are constructed by a 3D printer based on our MRI database [14]][15]. A horn driver unit in a hermetic enclosure was used as a source sound generator.

The vowel synthesis system is illustrated in Fig. 6. The transfer function of the solid vocal tract model was measured employing the swept-sine method [16], and resonance and anti-resonance frequencies were measured for accuracy evaluation on the proposed method.

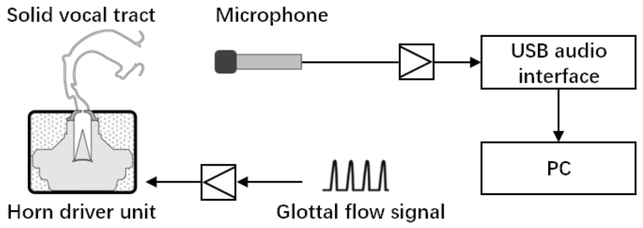

Fig. 6. Vowel synthesis system using a solid vocal tract model.

3.1.2. Vowel synthesis

Two female subjects' solid vocal tract models of vowel /a/ were employed for testing the proposed method. The glottal volume velocity waves with varying F0s from 150 Hz to 750 Hz were generated according to the Rosenberg model [17]. To synthesize the natural-sounding vowels, the jitter of a natural range was added to the signals. White noise was also added to simulate the glottal noise component to have a normal range of the harmonic-to-noise ratio (HNR) [18][19]. The synthetic vowels were recorded in a soundproof room at a sample rate of 44100 Hz.

3.2. Transfer function of the model

Performance evaluation of the proposed method is conducted on the synthesized vowels with varied F0s, using two cepstral analysis methods, the standard cepstral (CEPS) and RS-CEPS methods. The vowel sounds were down-sampled to 24000 Hz, and the Blackman-windowed analysis frame was set to 25 ms. Both resonance and anti-resonance frequencies are selected as validation data. The accuracy of formant frequencies is also compared with a formant tracking method in Praat employing the Burg's method.

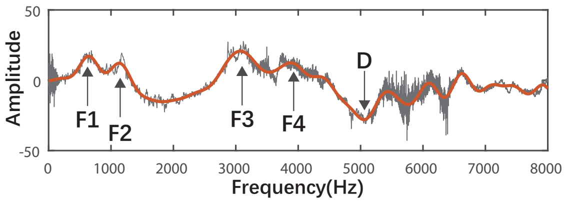

Fig. 7. The transfer function of a solid vocal tract model of vowel /a/. The solid line is the smoothed transfer function, and the gray line is the raw transfer function. F1 to F4 are the first four formants. D is the anti-resonance due to the piriform fossa.

Fig. 7 shows the transfer function of a vocal tract model for vowel /a/ obtained by the swept-sine method, and selected formants and an anti-resonance dip were used as the reference frequency values for the comparison. The formant/dip estimation errors are quantified by

$$d_{err} = 100\% \times \frac{|V_{est,i} - V_{tru,i}|}{V_{tru,i}}, \tag{5}$$

where $V_{est,i}$ is the estimated i_{th} formant/dip frequency, and $V_{tru,i}$ is the corresponding formant/dip frequency on the transfer function.

The error measure that summarizes for all four formants was examined by computing the Euclidean distance (in Hz) between the true and estimated formants:

$$d_{euc} = \sqrt{\sum_{i=1}^{4}(V_{est,i} - V_{tru,i})^2}, \tag{6}$$

where $V_{est,i}$ and $V_{tru,i}$ are defined in Eq. (5).

4. RESULTS AND DISCUSSIONS

The results comparing the conventional cepstrum (CEPS), Praat (Burg) and proposed (RS-CEPS) methods are shown as the error functions in Fig. 8 (a)-(d) for the peaks and in Fig. 8 (e) for the dip in the high frequency. The estimated errors for the four peaks, i.e., formants, demonstrate that the accuracy of RS-CEPS (circles) for the four formants excels those of the standard cepstral analysis (asterisks) and Praat (Burg) method (triangles). Especially, the RS-CEPS maintains stable performance on F1 and F2 estimates throughout the whole F0 range, while the CEPS fails at certain higher F0 ranges.

From the estimate errors for the anti-resonance dip, even though the RS-CEPS method (circles) does not show significant advantages in vowels below 500 Hz compared with the CEPS method (asterisks), it reaches better performance in higher frequencies, while the stable accuracy seen in the lower frequencies is maintained.

Fig. 8 (f) shows the Euclidean formant estimation error of the first four formants for the CEPS (asterisks), Praat (triangles) and RS-CEPS (circles) methods. According to the overall comparison, the RS-CEPS method excels at representing stable and precise vocal tract resonances in both low and high frequency ranges.

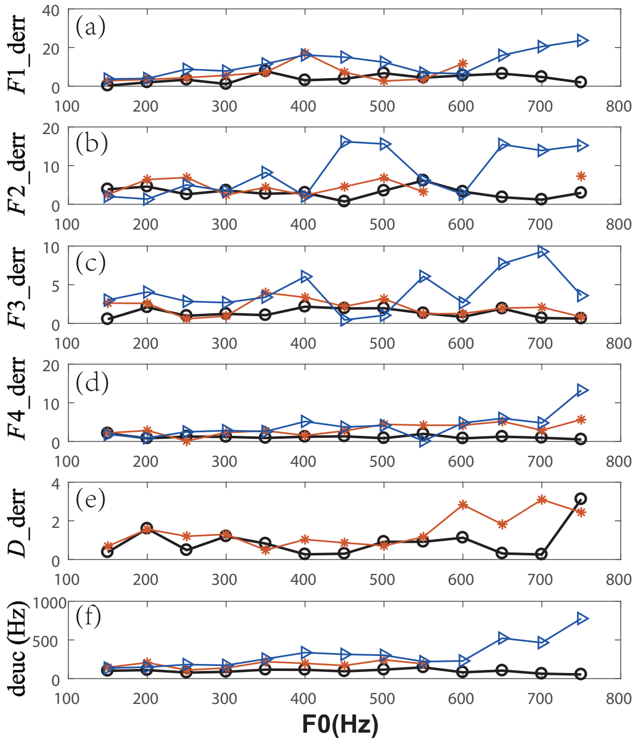

Fig. 8. Estimation error results as a function of F0, showing (a)-(d) the estimation errors (%) for the first four formants, (e) the estimate errors (%) for the anti-resonance dip, and (f) the Euclidean formant estimation errors (Hz) for the first four formants.

5. CONCLUSION

In this study, a new cepstral approach using a rahmonic subtraction technique (RS-CEPS) is proposed to retrieve both vocal tract resonance and anti-resonance from high-pitched vowels. This technique focuses on the spectral components generated by glottal noise to obtain their resonance pattern. The subtraction procedure preserves cepstral information near the rahmonic peak regions after subtraction with minimal distortions on spectral information over a wide range of F0. Comparisons were made among the proposed method, a tool in Praat and standard cepstral analysis technique. The results demonstrate that our new method successfully restores accurate spectral representations with minimal effects of harmonics. In the future, further optimizations of the RS-CEPS will be sought to develope automatic formant tracking technique and improve the computational efficiency toward advancement in speech technology.

6. REFERENCES

[1] Kiyoshi Honda, Tatsuya Kitamura, Hironori Takemoto, et al., "Visualisation of hypopharyngeal cavities and vocal-tract acoustic modelling," *Computer Methods in Biomechanics and Biomedical Engineering*, vol. 13, no. 4, pp. 443–453, 2010.

[2] Tatsuya Kitamura, Kiyoshi Honda, and Hironori Takemoto, "Individual variation of the hypopharyngeal cavities and its acoustic effects," *Acoustical Science and Technology*, vol. 26, no. 1, pp. 16–26, 2005.

[3] John D Markel and AH Jr Gray, *Linear Prediction of Speech*, vol. 12, Springer Science & Business Media, 2013.

[4] Amro El-Jaroudi and John Makhoul, "Discrete all-pole modeling," *IEEE Transactions on Signal Processing*, vol. 39, no. 2, pp. 411–423, 1991.

[5] M Shahidur Rahman and Tetsuya Shimamura, "Formant frequency estimation of high-pitched speech by homomorphic prediction," *Acoustical Science and Technology*, vol. 26, no. 6, pp. 502–510, 2005.

[6] Tianyu T Wang and Thomas F Quatieri, "High-pitch formant estimation by exploiting temporal change of pitch," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 1, pp. 171–186, 2009.

[7] Christine H Shadle, Hosung Nam, and DH Whalen, "Comparing measurement errors for formants in synthetic and natural vowels," *The Journal of the Acoustical Society of America*, vol. 139, no. 2, pp. 713–727, 2016.

[8] Paavo Alku, Jouni Pohjalainen, Martti Vainio, Anne-Maria Laukkanen, and Brad H Story, "Formant frequency estimation of high-pitched vowels using weighted linear prediction," *The Journal of the Acoustical Society of America*, vol. 134, no. 2, pp. 1295–1313, 2013.

[9] Ronald W Schafer and Lawrence R Rabiner, "System for automatic formant analysis of voiced speech," *The Journal of the Acoustical Society of America*, vol. 47, no. 2B, pp. 634–648, 1970.

[10] S. Imai and Y. Abe, "Spectral envelope extraction by improved cepstral method," *Electron. Comm.*, vol. 62, no. 4, pp. 10–17, 1979.

[11] Helen M Hanson, "Glottal characteristics of female speakers: Acoustic correlates," *The Journal of the Acoustical Society of America*, vol. 101, no. 1, pp. 466–481, 1997.

[12] Donald G Childers, David P Skinner, and Robert C Kemerait, "The cepstrum: A guide to processing," *Proceedings of the IEEE*, vol. 65, no. 10, pp. 1428–1443, 1977.

[13] RB Randall, J Antoni, and WA Smith, "A survey of the application of the cepstrum to structural modal analysis," *Mechanical Systems and Signal Processing*, vol. 118, pp. 716–741, 2019.

[14] Congcong Zhang, Kiyoshi Honda, Ju Zhang, and Jianguo Wei, "Contributions of the piriform fossa of female speakers to vowel spectra," in *2016 10th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. IEEE, 2016, pp. 1–5.

[15] Ju Zhang, Kiyoshi Honda, and Jianguo Wei, "Tooth visualization in vowel production mr images for three-dimensional vocal tract modeling," *Speech Communication*, vol. 96, pp. 37–48, 2018.

[16] Angelo Farina, "Simultaneous measurement of impulse response and distortion with a swept-sine technique," in *Audio Engineering Society Convention 108*. Audio Engineering Society, 2000.

[17] Aaron E Rosenberg, "Effect of glottal pulse shape on the quality of natural vowels," *The Journal of the Acoustical Society of America*, vol. 49, no. 2B, pp. 583–590, 1971.

[18] Paul Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," in *Proceedings of the institute of Phonetic Sciences*. Amsterdam, 1993, vol. 17, pp. 97–110.

[19] Eiji Yumoto, Wilbur J Gould, and Thomas Baer, "Harmonics-to-noise ratio as an index of the degree of hoarseness," *The journal of the Acoustical Society of America*, vol. 71, no. 6, pp. 1544–1550, 1982.