INTERSPEECH 2009 BRIGHTON

A Novel Codebook Search Technique for Estimating the Open Quotient

Yen-Liang Shue[1], *Jody Kreiman*[2], *Abeer Alwan*[1]

[1] Department of Electrical Engineering, University of California, Los Angeles
[2] Division of Head and Neck Surgery, UCLA School of Medicine

yshue@ee.ucla.edu, jkreiman@ucla.edu, alwan@ee.ucla.edu

Abstract

The open quotient (OQ), loosely defined as the proportion of time the glottis is open during phonation, is an important parameter in many source models. Accurate estimation of OQ from acoustic signals is a non-trivial process as it involves the separation of the source signal from the vocal-tract transfer function. Often this process is hampered by the lack of direct physiological data with which to calibrate algorithms. In this paper, an analysis-by-synthesis method using a codebook of harmonically-based Liljencrants-Fant (LF) source models in conjunction with a constrained optimizer was used to obtain estimates of OQ from four subjects. The estimates were compared with physiological measurements from high-speed imaging. Results showed relatively high correlations between the estimated and measured values for only two of the speakers, suggesting that existing source models may be unable to accurately represent some source signals.

Index Terms: open quotient, voice source, speech analysis

1. Introduction

According to the linear acoustic theory of speech production [1], speech signals are generated by a source or excitation signal filtered by the vocal tract transfer function (VTTF). In many applications, we are interested in the underlying acoustic features of the source signal because it can carry information regarding stress (or emphasis), emotional status, prosodic events, or even an underlying disease of the vocal cords. An important parameter in many source models is the open quotient (OQ), which is loosely defined as the proportion of time the glottis is open during a cycle of phonation. OQ has been associated with some aspects of voice quality, but is difficult to estimate directly. For this reason, it is often alluded to indirectly using correlates, such as H_1^*-H_2^* [2], the difference between the magnitudes of the first two spectral harmonics corrected for the effects of the vocal tract. Unfortunately, the relationship between OQ and H_1^*-H_2^* is more complex than previously assumed ([3], [4]), making direct estimation more desirable, but this requires separating the source signal from the VTTF. This is a non-trivial process and is often hampered by the lack of direct physiological data, that is, the ground truth, with which to calibrate algorithms.

Typical source estimation methods involve the initial estimation of the vocal tract filter, followed by inverse-filtering of the speech signal to obtain the source signal. While this method is suitable for some applications, it can be inaccurate because (1) it relies heavily on accurate estimation of the vocal tract filter, and (2) it enforces the linear model onto the vocal tract leaving the residual signal to carry the non-linear source-tract interaction information. In [5], the accuracy of source estimation through inverse-filtering was improved by using variable window lengths to better capture the VTTF in the glottal closure regions. There have also been many proposed joint source-tract estimation algorithms ([6], [7], [8]). In [6], Liljencrants-Fant (LF) source model parameters were estimated iteratively using multi-dimensional optimization techniques that were initialized based on the results of an exhaustive parameter search. During the exhaustive search, a source parameter set was tested by removing its spectrum from the speech spectrum, estimating the VTTF and comparing the output of inverse-filtering with the source spectrum from the parameter set. Multi-dimension optimizers were then employed to further refine the parameters. In [7] and [8], a global optimization scheme was used to estimate the parameters of the source and vocal-tract filter simultaneously. While joint estimation usually does well in minimizing the re-synthesized output error (i.e. analysis-by-synthesis), it is not clear how accurately the source can be estimated with this method. Often, calibrations are done with electroglottographic (EGG) signals ([5], [6]), which is an indirect observation of the glottis that can be affected by noise sources such as DC offsets, mucus bridges and calculation thresholding.

In this paper, an analysis-by-synthesis method using a codebook of source models in conjunction with a constrained optimizer was applied to estimate the OQ of a source signal from acoustic data. Calibration of the results was performed by comparing the estimated OQ values with physiological measurements from high-speed imaging of the larynx.

2. Data and Method

2.1. Data

The data used in this study are the same as those used in [4]. A summary of the data collection procedure is presented here.

Audio and high-speed video data were recorded synchronously from 4 speakers (2 females/2 males). The video was recorded with a laryngoscope, positioned to visualize the larynx, at a speed of 3000 frames/second and resolution of 512×512 pixels. Audio signals were transduced with a Bruel & Kjaer microphone at a sampling rate of 60 kHz; the signals were then downsampled to 16 kHz for analysis.

Table 1. *Mean F_0 values for the four speakers.*

Speaker	low F_0 (Hz)	normal F_0 (Hz)	high F_0 (Hz)
F1	158	215	337
F2	193	211	315
M1	109	166	235
M2	101	136	202

The speakers were asked to produce the vowel /i/ with different voice qualities (pressed, normal and breathy) and different fundamental frequencies (F_0; low, normal and high). The mean F_0 values, as estimated by the Straight algorithm [10], are listed in Table 1 for the two female (F1 and F2) and two male (M1 and M2) speakers. For each recording, speakers sustained the vowel while holding the required voice quality

and F_0 as steady as possible. One-second samples of audio and video were taken from the most stable sections for analysis.

The second author viewed the high-speed video images frame by frame and manually marked the times of the first instants of glottal opening and the points of maximum glottal closure. The OQ was then calculated on a cycle-by-cycle basis as the ratio of the time from the first opening instant to the point of maximum closure to the time from the opening instant to the next cycle's opening instant.

2.2. Definition of OQ

There are many definitions of OQ, but in order to make a direct comparison with the physiological data (high-speed video), OQ will be defined for this study as the proportion of time the glottal flow waveform, normalized to a maximum amplitude of 1, is above a threshold of 0.01. An example is shown in Figure 1.

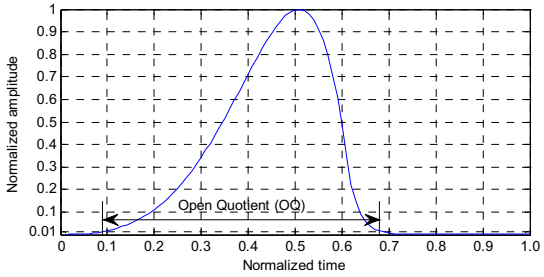

Figure 1. *Definition of OQ used in this study. In this example, OQ = 0.6.*

2.3. OQ estimation process

The linear source-filter model of speech production states that for short-time periods, speech, *s(t)*, can be approximated as a cascade of linear systems involving a source function, *u(t)*, a vocal-tract transfer function, *v(t)*, and a differentiator which is usually incorporated into the source function:

$$s(t) = u(t) * v(t)$$

$$S(\omega) = U(\omega)V(\omega)$$

Taking the magnitudes of each system, V(ω) can be written (in dB) as:

$$|V(\omega)| = |S(\omega)| - |U(\omega)|$$

This equation implies that if a source spectrum was known, the vocal tract spectrum could be calculated exactly. However, estimation of the spectrum, $|S(\omega)| - |U(\omega)|$, is not robust, and can often result in spurious values near the valleys of $S(\omega)$. More robust are the harmonic magnitudes, as used in [6], denoted by $|S(\omega_{Hk})|$, $|U(\omega_{Hk})|$ and $|V(\omega_{Hk})|$, where $\omega_{Hk} = 2\pi k F_0/F_s$, $k \in \mathbb{Z}^+$. Furthermore, the effects of the overall signal power can be neglected if the harmonic magnitudes are normalized to the first harmonic magnitude; e.g. in dB, $S_n = |S(\omega_{H1})| - |S(\omega_{Hn})|$.

In this paper, a codebook was constructed using the normalized harmonic magnitudes of the four parameter LF model [9]. Note that while codebooks have been used extensively in speech coders, the contents of those codebooks are vastly different from the codebooks of applications that seek to estimate the source signal. Here, grid searches were performed on each of the four parameters (t_e, t_p, t_a and E_e) at the following resolutions: t_e from 0.3 to 0.98 at increments of 0.01, t_p from 0.01 to 0.95 at increments of 0.01, t_a from 0.01 to 0.95 at increments of 0.01 and E_e from 0.1 to 5 at increments of 0.1. Since not every combination of the four parameters constituted a valid glottal flow derivative waveform, the resulting signals were checked to ensure they were physically realizable. Note that while the codebook uses glottal flow derivatives, the OQ calculations use the glottal flow, in line with the OQ measurements from the high-speed video. Normalized harmonic magnitudes and OQ values were then extracted from these waveforms and stored in a codebook. The large number of entries in the codebook was reduced by performing a correlation analysis and discarding those entries which had a correlation coefficient of 0.9 or more. This resulted in a final codebook size of 1726 entries.

The main OQ estimation process for a given codebook is shown in Figure 2. To reduce the number of codebook entries used in each OQ estimation and hence, increase processing speed, a two-iteration approach was used. The first iteration involves a smaller codebook consisting of 13 entries taken from the larger codebook. These entries, based on glottal flow waveforms which have OQ values beginning from 0.35 to 0.95 at increments of 0.05, were selected by averaging the source parameters across the list of entries which have the required OQ value. After the first iteration of the estimation process, the source with the smallest error, denoted by *m*, was used to select the entries from the larger codebook for the second iteration. Assuming that the OQ for entry *m* was OQ_m, then all entries in the larger codebook which had OQ greater than $OQ_m - 0.1$ and less than $OQ_m + 0.1$ were selected for the second iteration. At the end of the second iteration, the entry with the smallest error was returned as the most likely source for the given input signal.

The normalized harmonic magnitudes, S_n, of the input signal were calculated from the spectrum using the pitch information which was extracted by the STRAIGHT algorithm [10]. A Hamming window consisting of 4 pitch periods was used to calculate the spectrum of the input signal. Hence, the window length was different for each speaker. For this study, the number of harmonics used was in the range 0 to 2.6 kHz. For example, for a pitch period of 100 Hz, 26 harmonics would be used. This number is arbitrary and in practice depends only on the number of harmonics that can be reliably estimated from the spectrum.

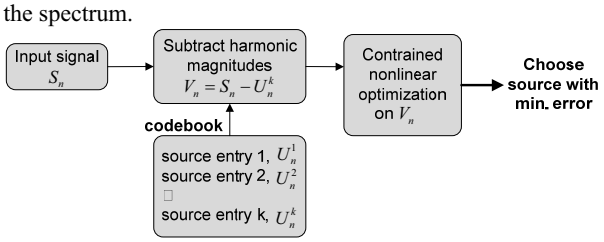

Figure 2. *The main OQ estimation process.*

For each entry in the codebook, denoted by U_n^k for the *k*-th source entry and the *n*-th normalized harmonic magnitude, subtraction with S_n is performed to produce an estimate of the normalized harmonic magnitudes of the vocal tract, i.e. $V_n = S_n - U_n^k$. A constrained nonlinear optimization, using the active-set quadratic programming method, is then performed on V_n to find an estimate of the formant frequencies and their bandwidths as well as an error value. Since this study involved only vowels and used harmonics up to 2.6 kHz, a 3-formant (6-pole) model was used for the vocal tract:

$$|V(\omega)|^2 = \prod_{p=1}^{3} \frac{1}{\left(1 - 2r_p \cos(\omega - \omega_p) + r_p^2\right)\left(1 - 2r_p \cos(\omega + \omega_p) + r_p^2\right)}$$

where $r_p = e^{-\pi B_p / F_s}$, $\omega_p = 2\pi F_p / F_s$, F_s is the sampling frequency and F_p and B_p are the formant frequencies and their respective bandwidths. The constraints on the formants and bandwidths are listed in Table 2. Note that, different to other

source estimation schemes, the optimization here is over the VTTF parameters, allowing the error signal to be modeled by the VTTF. Although the speakers were asked to produce the vowel /i/, the vowel quality ranged from /I/ to /æ/ due to the positioning of the laryngoscope. The optimization constraints for the formant frequencies were based on the mean values obtained from the Snack Sound Toolkit [11], with separate averages for the female and male speakers. For F_1, the constraints were set to ±150 Hz of the mean value for each gender, while for F_2 and F_3, the constraints were set to ±500 Hz of the mean value. Bandwidth constraints were based on the formant-bandwidth mapping formula given in [12]. The optimization criterion is a weighted least squares error function:

$$E_k = \min_{V_n} \sum_{n=2}^{N} \left(S_n - U_n^k - V_n \right)^2 \cdot W_n$$

where N is the number of harmonics up to 2.6 kHz and W_n is a weighting function used to emphasize the lower frequency harmonic magnitudes. For the results presented in this study, W_n was empirically defined as:

$$W_n = \begin{cases} 2^{12-n}, & 2 \le n \le 12 \\ 1, & n > 12 \end{cases}$$

Table 2. *Optimization constraints for formant frequencies and their bandwidths.*

Formants	min. value (Hz) (female/male)	max. value (Hz) (female/male)
F_1	360/320	660/620
F_2	1110/1020	2110/2020
F_3	2195/2050	3195/3050
Bandwidths		
B_1	35/30	150/120
B_2	40/30	180/120
B_3	70/50	270/260

3. Results

For each audio file, OQ estimates were made every 20 ms for a total of 50 measurements per file. Similarly, OQ measurements from the high-speed video were also selected every 20 ms. However, since the OQ measurements from the high-speed video are on a cycle-by-cycle basis, each 20 ms measurement was calculated to be the mean of the OQ values of the four closest cycles from the selected time point. For brevity, in this section, OQ estimates from the audio file will be denoted by OQ_a and those from the high-speed by OQ_v.

Table 3. *The Pearson's correlation coefficient (PCC), F values and null hypothesis probability (sig.) from a linear regression analysis for each speaker.*

Speaker	PCC	*F*	sig.
F1	.971	$F(1, 480) = 7937$	$p < .000$
F2	.778	$F(1, 238) = 366$	$p < .000$
M1	.925	$F(1, 421) = 2506$	$p < .000$
M2	.723	$F(1, 425) = 466$	$p < .000$

Table 3 shows linear regression analysis results for each speaker. The values given are the Pearson's correlation coefficient (PCC), F values, and the null hypothesis probability. Overall, the four speakers showed good correlations although the OQ estimation accuracy was better for speakers F1 and M1.

OQ_a vs. OQ_v plots, grouped in terms of the speakers' F_0 ranges and phonation types, are shown in Figures 3 and 4 for speakers F1 and M1. It can be seen that the correlation is quite high with PCC > 0.9. A few outlier points can be seen for these speakers, however unlike speakers F2 and M2, these outliers do not involve complete groups of points indicating it may be caused by noise in the input signal. The lack of spread in the OQ_a values for speaker F1 could be due to the codebook not having the entries to closely match this speaker's source. Note that OQ_v values which are equal to 1 correspond to incidents where no observable closure of the glottis could be seen. The current LF-based codebook appears to have no source functions which can accurately model these types of phonation.

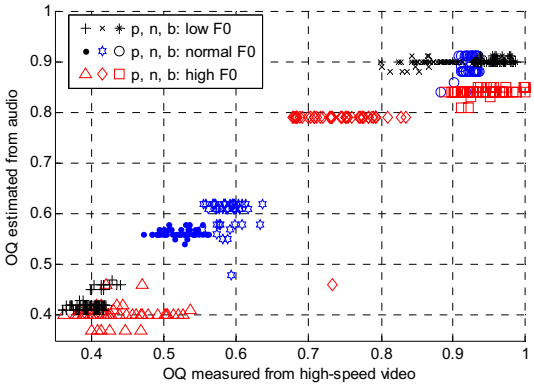

Figure 3. *OQ estimation results for speaker F1 (PCC = 0.971), grouped by low, normal and high F_0, for the phonation types pressed (p), normal (n) and breathy (b).*

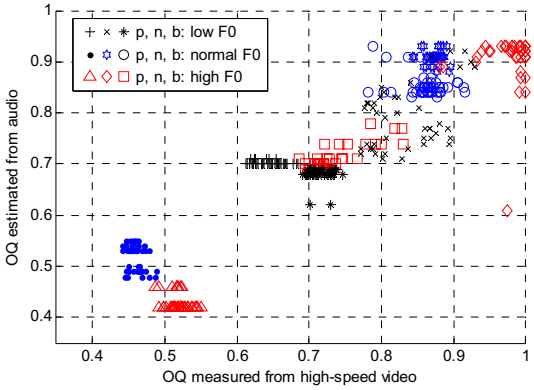

Figure 4. *OQ estimation results for speaker M1 (PCC = 0.925), grouped by low, normal and high F_0, for phonation types pressed (p), normal (n) and breathy (b).*

OQ estimates for speaker F2 and M2 are shown in Figures 5 and 6, respectively. Speaker F2 had fewer data points than the other speakers as there were fewer recordings taken for this speaker. For both speakers, outliers occurred, consisting of complete groups of points. For comparison, outlier groups are defined as those cases where the mean of OQ_a differs from the mean of OQ_v by more than 0.15. A closer analysis of these groups showed that the codebook entry selected by the first iteration of the OQ estimation process was not very accurate. This is most likely due to the codebook lacking the models to capture, with any effectiveness, the source waveforms produced by these groups. A small experiment was performed using the entire codebook (with 1726 entries) in the first iteration of the estimation process; no change in results were seen for either speaker. Another possible cause is the formant/bandwidth constraints used in the optimization. However, when the constraints were changed to be more constrictive (reducing the range of formant/bandwidth values),

the results for speaker M2 did not show any difference while for speaker F2 a slight improvement could be observed, but was not enough to remove the outliers.

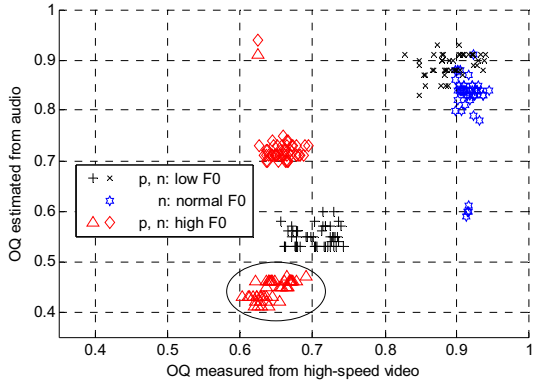

Figure 5. *OQ estimation results for speaker F2 (PCC = 0.778), grouped by low, normal and high F_0, for phonation types pressed (p), normal (n) and breathy (b). The outlier group is circled.*

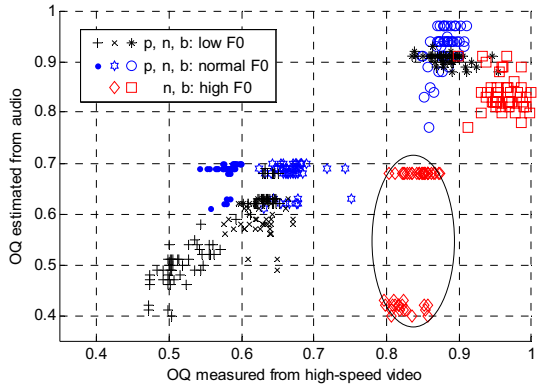

Figure 6. *OQ estimation results for speaker M2 (PCC = 0.723), grouped by low, normal and high F_0, for phonation types pressed (p), normal (n) and breathy (b). The outlier group is circled. Note that no pressed high F_0 phonations were recorded for this speaker.*

The results show that, for these four speakers, the accuracy of the OQ estimation process appears to be fairly independent of F_0.

4. Summary and Conclusions

An analysis-by-synthesis method was used in conjunction with a source codebook to estimate the OQ for four speakers producing the vowel /i/ with varying voice qualities and varying F_0. Significant correlations were found between the estimated OQ and OQ measured from high-speed filming of the larynx. However, for two speakers, outlier groups existed, possibly due to lack of source models in the codebook that could effectively capture the source signal of those particular phonations.

In estimating OQ, a shape of the source signal is also estimated from the entries in the codebook. However, more research is needed to determine whether the estimated source signals bear any resemblance to the physiological data. Preliminary analysis of speaker F1 suggests that a speaker's range of glottal configurations extends beyond what can be represented by the LF model. This is consistent with what was reported in [3], which suggested that the modeling capabilities of current glottal models are deficient. Figure 7 shows a comparison of the estimated source signal with the measured source signal, obtained by calculating the area of the glottal opening from the high-speed images. It can be seen that for this particular phonation, the slope of the opening phase is actually steeper than the slope of the closing phase. It can also be seen that, due to a glottal gap, this signal does not return to zero at the end of each cycle. Both of these conditions are not handled by the LF model in the current codebook design, but will be addressed in future work.

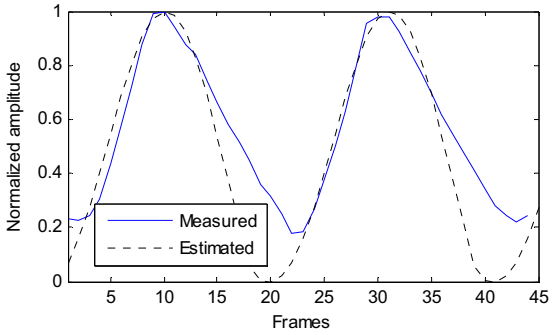

Figure 7. *Comparison of a measured source signal with the estimated source signal for speaker F1.*

5. Acknowledgements

This work was supported in part by the NSF, and by grant DC01797 from the NIH/NIDCD.

6. References

[1] G. Fant, "Acoustic theory of speech production," Mouton, The Hague, Paris, 2nd edition, 1970.

[2] E. B. Holmberg, R. E. Hillman, J. S. Perkell, P. Guiod, and S. L. Goldman, "Comparisons among aerodynamic, electroglottographic, and acoustic spectral measures of female voice", *JSHR*, vol. 38, pp. 1212-1223, 1995.

[3] N. Henrich, C. d'Alessandro and B. Doval, "Spectral correlates of voice open quotient and glottal flow asymmetry: theory, limits and experimental data", *Proc. Eurospeech*, pp. 47-50, 2001.

[4] J. Kreiman, B. R. Gerratt, M. Iseli, J. Neubauer, Y.-L. Shue and A. Alwan, "The relationship between open quotient and H_1^*-H_2^*," *Proc. 6th International Conf. on Voice Physiology and Biomechanics*, Tampere, Finland, August, 2008.

[5] E. Moore, II and M. Clements, "Algorithm for automatic glottal waveform estimation without the reliance on precise glottal closure information," in *Proc. ICASSP*, pp. 101-104, May, 2004.

[6] M. Fröhlich, D. Michaelis, and H. W. Strube, "SIM – simultaneous inverse filtering and matching of a glottal flow model for acoustic speech signals," *JASA*, 110(1): 479-488, July 2001.

[7] A. del Pozo and S. Young, "The linear transformation of LF glottal waveforms for voice conversion", *Proc. Interspeech*, 2008, pp. 1457-1460.

[8] P. Jinachitra and J. O. Smith III, "Joint estimation of glottal source and vocal tract for vocal synthesis using Kalman smoothing and EM algorithm", *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 327-330, October, 2005.

[9] G. Fant, J. Liljencrants and Q. Lin. "A four-parameter model of glottal flow," *STL-QPSR Report*, pp. 1-13, 1985.

[10] H. Kawahara, A. de Cheveigné, and R. D. Patterson, "An instantaneous-frequency-based pitch extraction method for high quality speech transformation: revised TEMPO in STRAIGHT-suite", in *Proc. ICSLP*, 1998.

[11] K. Sjölander, "Snack sound toolkit", KTH Stockholm, Sweden, 2004, http://www.speech.kth.se/snack/.

[12] J. W. Hawks and J. D. Miller, "A formant bandwidth estimation procedure for vowel synthesis", *JASA*, 97:1343-1344, 1995.