AN IMPROVED CORRECTION FORMULA FOR THE ESTIMATION OF HARMONIC MAGNITUDES AND ITS APPLICATION TO OPEN QUOTIENT ESTIMATION

Markus Iseli and Abeer Alwan

University of California Los Angeles
Dept. of Electrical Engineering
405 Hilgard Ave., Los Angeles, CA 90095
iseli@icsl.ucla.edu, alwan@icsl.ucla.edu

ABSTRACT

Many voice quality parameters, such as the open quotient (OQ), depend on an accurate estimate of the source spectrum. It is known that OQ, for example, is correlated with the magnitude difference of the first two harmonics ($H_1 - H_2$) of the speech source spectrum. In order to compare OQ estimates across different vocal tract configurations, magnitude correction, achieved by removing the influence of vocal tract resonances, has to be used. The improved correction described in this paper is inspired by a correction formula in [1]. The new correction formula accounts for the bandwidths of all vocal tract resonances, and most importantly, is not limited to the analysis of non-high vowels as is the case in [1]. $H_1 - H_2$ estimates, using the proposed technique with synthesized vowels generated with the LF and the KLGLOTT88 models, are very accurate.

1. INTRODUCTION

Glottal source characteristics, such as fundamental frequency (F_0), spectral tilt, aspiration noise, and open quotient (OQ) are often referred to as voice quality parameters. Applications which can benefit from better estimates of voice quality parameters include speech synthesis, voice morphing, speaker recognition, identification, and adaptation, as well as speech recognition and several medical applications. OQ is defined as the ratio of the time in which the vocal folds are open in one cycle of the fundamental period. Since OQ indicates the duty ratio of the glottal airflow, the spectrum of a glottal excitation is highly dependent on it. OQ is also highly correlated to physiological constraints which are reflected in different phonation types, such as creaky, breathy, and whispery [2].

In this paper, we focus on obtaining an accurate estimate of the magnitudes of the first two speech source spectral harmonics H_1 and H_2, since the difference $H_1 - H_2$ is correlated with OQ [3]. In [1], Hanson presents a correction formula, which removes the effect of the first formant on H_1 and H_2, making it possible to compare OQ measurements across different vowels. In [4] and [5], Hanson's seminal work provides a thorough comparison of voice quality parameters for male and female speakers, using her correction formula [1]. All results show that on average female speakers have higher OQ than male speakers and that the range and standard deviation are slightly larger for female speakers. Hanson's results are applicable to non-high vowels (/æ, ϵ, Λ/) and assume that the fundamental frequency is at least 100 Hz away from the first formant frequency (F_1). It does not take into account the bandwidth at F_1 nor the effects of higher formants.

In this study, we are interested in estimating $H_1 - H_2$ for the three corner vowels /a, i, u/. To accomplish this, a more complete correction formula, which takes into account formant bandwidths is proposed. The technique is evaluated with synthetic vowels produced with the Liljencrants-Fant (LF) and the KLGLOTT88 source models.

Work supported in part by the NSF.

2. HANSON'S CORRECTION FORMULA

The magnitudes of the source harmonics are influenced by the vocal tract filter. Assuming an all-pole model, the vocal tract transfer function $T(s)$ with n formants can be written as:

$$T(s) = K \prod_{i=1}^{n} \frac{\sigma_i^2 + \Omega_i^2}{(s - (\sigma_i + j\Omega_i))(s - (\sigma_i - j\Omega_i))}, \quad (1)$$

with $s = \sigma + j\Omega$, gain factor $|K| < 1$, $\sigma_i = \pi B_i$, and $\Omega_i = 2\pi F_i$. $T(s)$ is normalized such that $T(s = 0) = K$.

In [1] it is assumed that only F_1 is present and that $\sigma_1^2 \ll (\Omega_1 - \Omega)^2$, that is, F_1 bandwidth is negligible. Hanson's formula [1] for the correction of a harmonic magnitude H in the log domain (dB) is given by:

$$H^* \approx H - 20 \log_{10} \frac{F_1^2}{F_1^2 - f^2}, \quad (2)$$

where f is the first or second harmonic frequency ($f = F_0$, or $f = 2F_0$). Equation 2 further assumes that $f < F_1$ by at least 100 Hz, which is often not true, especially for children and female speakers with high F_0 and for high vowels such as /i/ and /u/, where often $f = 2F_0 > F_1$. Assuming that the harmonics' amplification is symmetric for f close to F_1, we can rewrite Eq. (2) to include such cases:

$$H^* \approx H - 20 \log_{10} \frac{F_1^2}{|F_1^2 - f^2|} = H - 10 \log_{10} \frac{F_1^4}{(F_1^2 - f^2)^2}. \quad (3)$$

Since Eq. (3) produced smaller correction errors than Eq. (2), it will be used for the remainder of this paper when referring to Hanson's correction formula.

3. PROPOSED ALGORITHM

Our improved algorithm takes into account the formant bandwidths B_i. It solves for the actual log magnitude contribution of the i-th formant at frequency Ω_i in the s-domain and not only for F_1:

$$H^* = H - \sum_{i=1}^{n} 10\log_{10}\frac{(\sigma_i^2+\Omega_i^2)^2}{(\sigma_i^2+(\Omega_i-\Omega)^2)(\sigma_i^2+(\Omega_i+\Omega)^2)}, \quad (4)$$

with the number of formants n, $\Omega = 2\pi f$, and $\Omega_i = 2\pi F_i$. Note that Eq. (4) reduces to Eq. (3) by setting $\sigma_i = 0$ and $n = 1$.

For sampled data, the corresponding z-domain formula is used. All frequencies are normalized to the sampling frequency F_s. The correction formula in the z-domain is:

$$H^* = H - \sum_{i=1}^{n} 10\log_{10}\frac{(r_i^2+1-2r_i\cos(\omega_i))^2}{(r_i^2+1-2r_i\cos(\omega_i+\omega))(r_i^2+1-2r_i\cos(\omega_i-\omega))}, \quad (5)$$

with $r_i = e^{-\pi B_i/F_s}$, $\omega_i = 2\pi F_i/F_s$, and $\omega = 2*\pi*f/F_s$. Results based on the improved harmonics' correction use Eq. 5.

4. GLOTTAL FLOW MODELS

The glottal flow models used in this paper are described briefly in this section.

4.1. Modified or simplified LF model

This model for the glottal flow derivative is described in [6]. Its basic equations for the open phase ($E_1(t)$) and the return phase ($E_2(t)$) in continuous time are:

$$E(t) = \begin{cases} E_1(t) & = E_0 e^{\alpha t}\sin(\omega_g t) & (t \le t_e) \\ E_2(t) & = (\frac{-E_e}{\epsilon T_a})[e^{-\epsilon(t-t_e)} - e^{-\epsilon(t_c-t_e)}] & (t_e < t \le t_c). \end{cases} \quad (6)$$

Parameters are the growth factor α, the amplitude scaling factor E_0, the exponential time constant of the return phase ϵ, the duration of the return phase T_a, the instant of glottal closure t_c, the instant of minimal glottal flow derivative t_e, and E_e, which is the magnitude of the signal at time t_e. Defining the instant of maximal glottal airflow as t_p, we have $\omega_g = \frac{\pi}{t_p}$. The asymmetry coefficient is $\alpha_m = \frac{t_p}{t_e}$. In our experiments it was kept at a constant value of $\alpha_m^{(LF)} = 0.8$, which is within the range of [0.5, 0.9] used by most implementations [7].

4.2. KLGLOTT88 model

As described in [8] and [9] the basic equations of the KLGLOTT88 model in discrete time are

$$g(n) = \begin{cases} 2an/F_s - 3b(n/F_s)^2 & (0 \le n \le T_0 OQ F_s) \\ 0 & (T_0 OQ F_s < n \le T_0 F_s) \end{cases} \quad (7)$$

with

$$a = \frac{27AV}{4OQ^2T_0} \quad (8)$$

and

$$b = \frac{27AV}{4OQ^3T_0^2}. \quad (9)$$

Parameters are OQ, amplitude of voicing AV, and fundamental period duration T_0. For this model the asymmetry coefficient is a constant: $\alpha_m^{(KL)} = 2/3$.

5. RESULTS

The following results were produced using the KLGLOTT88 model, since the LF model with asymmetry coefficient $\alpha_m = 0.8$ yielded almost exactly the same harmonic magnitudes H_1 and H_2. To get a fair comparison between our correction algorithm and that introduced by Hanson [1], in all experiments only the influence of the first formant was removed.

5.1. $H_1 - H_2$ error analysis

In a first step, the correction formulae were applied to synthetic signals with F_0 varying between 100 and 300 Hz, F_1 between 200 and 800 Hz with constant bandwidth of $B_1 = 50$ Hz, and OQ between 30% and 70%. Since the signals are synthetic, the actual values for H_1 and H_2 are known beforehand and the error of correction between the actual and the measured harmonics' magnitude difference can be calculated. For any F_0 and F_1, the average error of correction over all values of OQ showed a standard deviation which is close to zero (around 0.3 dB). This implies that the error is statistically independent of OQ. In fact, the correction terms in Eqs. (3) and (5) do not depend on H_1 nor on H_2, and since OQ is strongly related to their difference, this result seems reasonable. For the remainder of this paper we will refer to this average error of correction simply as $H_1 - H_2$ error.

Figure 1 shows the $H_1 - H_2$ error using Hanson's formula from Eq. (3). A cut through Fig. 1 is depicted in Fig. 2. The $H_1 - H_2$ error for our formula from Eq. (5) is the slash-dotted line at 0 dB. This result was expected, since our algorithm performs an exact inverse filtering operation. The $H_1 - H_2$ error for Hanson's formula is shown as a solid line, where the positive peak corresponds to the error at $F_1 \approx F_0$ (first harmonic) and the negative peak to $F_1 \approx 2F_0$ (second harmonic, see Eq. (3)). The peaks can be as high as 28 dB, and due to overcompensation, even higher than the maximum error without correction.

As was stated by Hanson in [5], her correction formula is only suited for non-high vowels and for F_1 more than 100 Hz away from the harmonic frequencies. More precisely, it can be seen from Fig. 2 that the formula should be applied to non-high vowels only if F_1 is more than its bandwidth (B_1) away from the harmonic frequencies:

$$|F_1 - f| > B_1. \quad (10)$$

5.2. Synthesis of corner vowels using LF and KLGLOTT88 models

The vowels /a/, /i/, and /u/, were synthesized using formant frequencies from [10]. Since no bandwidths were provided in that paper, formant bandwidths were calculated according to the formula [11]

$$B_i = (80 + 120F_i/5000)\cdot v, \quad (11)$$

where v is equal to 1 for voiced sounds and equal to 2 for unvoiced sounds.

These values are depicted in Table 1.

The KLGLOTT88 voice source signal was filtered with an all-pole model of the vocal tract. For each vowel, F_0 was varied in 54 steps between 100 and 300 Hz, and OQ in 9 steps between 30%

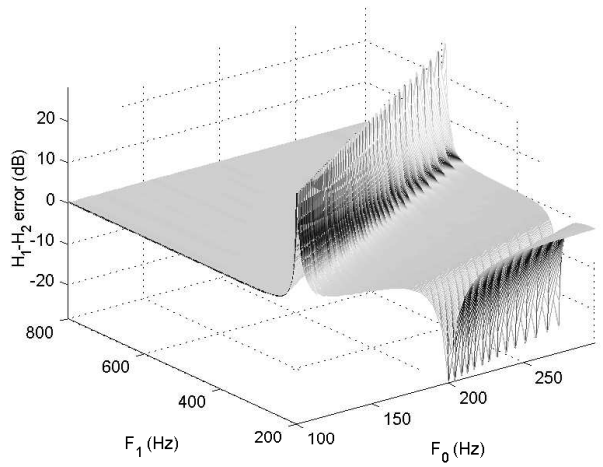

Fig. 1. $H_1 - H_2$ error in dB (mean over all OQs) using the formula in [1]. F_0 is between 100 and 300 Hz, F_1 between 200 and 800 Hz, and $B_1 = 50$ Hz.

Table 1. *Formant frequencies and bandwidths in Hz for the 3 corner vowels for female, male, and child speakers.*

Vowel	F_1	F_2	F_3	B_1	B_2	B_3
Female speakers						
/a/	850	1220	2810	100	109	147
/i/	310	2790	3310	87	147	159
/u/	370	950	2670	89	103	144
Male speakers						
/a/	730	1090	2440	98	106	139
/i/	270	2290	3010	86	135	152
/u/	300	870	2240	87	101	134
Children						
/a/	1030	1370	3170	105	113	156
/i/	370	3200	3730	89	157	170
/u/	430	1170	3260	90	108	158

and 70%. F_s was at 8kHz. From the resulting $486\ (9 \cdot 54)$ $H_1 - H_2$ error values, for each gender, vowel, and correction method, the minimal, average, and maximal $H_1 - H_2$ errors were calculated. Results are listed in Table 2. It uses synthesized vowels with F_1 only.

Table 2. *Min/Mean/Max $H_1 - H_2$ error in dB without correction and with correction in [1]. Single-formant vowels for a synthetic female voice are used.*

Vowel	No correction	Using [1]
/a/	0.38/1.31/4.65	0.00/0.02/0.03
/i/	0.20/6.68/19.23	0.01/3.58/25.19
/u/	0.56/5.32/11.49	0.02/2.32/26.24

In contrast to results for the synthesized female voice with one formant frequency in Table 2, results in Table 3 are for vowels synthesized with three formants. From Eq. (5) we can see that each additional formant, in the log domain adds an offset, which should be subtracted in the correction, or inverse filtering process. Therefore the values in Table 3 are generally higher than those in Table 2. One would expect the error to increase most for vowels

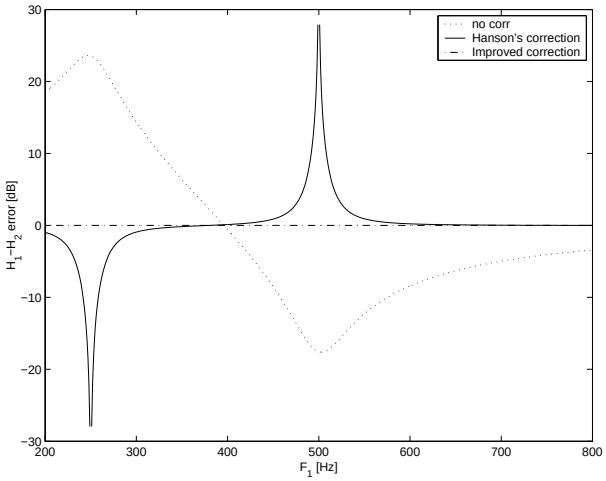

Fig. 2. Error in dB (mean over all OQs), with $F_0 = 250$ Hz and $B_1 = 50$ Hz . Without correction (dotted line), with correction in [1] (solid line), and with our formula (slash-dotted line). The solid line is a vertical cut at $F_0 = 250$ Hz through Fig. 1. Maximum error with no correction is about 24 dB, with the correction in [1] can be as high as 28 dB, and is 0 dB for our case.

having a low F_2, i.e. for /u/ ($F_2 = 950$ Hz) and less for /a/ ($F_2 = 1220$ Hz). For /i/ there should almost be no change, since $F_2 = 2790$ Hz has very little impact on lower frequencies. Interestingly though, the error increase when adding more formants for /u/ is much lower than it is for /a/. One explanation is that since /u/ has a low F_1 at 370 Hz, the first harmonic frequency component is amplified, whereas the second harmonic is attenuated most of the time. Adding a second formant at $F_2 = 950$ Hz, results in little change for the very low frequency first harmonic, but definitely amplifies the second harmonic component, compensating the error introduced by F_1.

A comparison of vowel differences with data from Table 3 is depicted as a bar diagram in Fig. 3. It shows that Hanson's correction, as stated by her, works well for the non-high vowel /a/ and that the error for high vowels /i/ and /u/ is considerable. Our new

Table 3. *Min/Mean/Max $H_1 - H_2$ error in dB without correction, with correction in [1], and the correction in this paper. Synthesis included all three formants and correction was done only on the first formant.*

Vowel	No correction	Using [1]	This paper
Female Speakers			
/a/	0.62/2.07/7.02	0.25/0.78/2.36	0.24/0.76/2.37
/i/	0.17/6.67/18.39	0.11/3.58/24.96	0.09/0.28/0.84
/u/	0.41/5.70/11.53	0.02/2.29/24.92	0.36/1.19/3.93
Male Speakers			
/a/	0.82/2.85/10.46	0.30/0.93/2.68	0.30/0.95/3.01
/i/	0.07/7.70/18.95	0.04/5.56/30.55	0.11/0.34/1.01
/u/	0.06/6.74/15.00	0.04/4.46/26.46	0.43/1.45/4.98
Child Speakers			
/a/	0.46/1.48/4.72	0.21/0.65/1.98	0.20/0.62/1.90
/i/	0.08/5.40/10.73	0.06/2.23/25.95	0.09/0.26/0.76
/u/	0.35/5.17/12.51	0.03/1.59/23.89	0.25/0.79/2.49

approach reduces this error significantly.

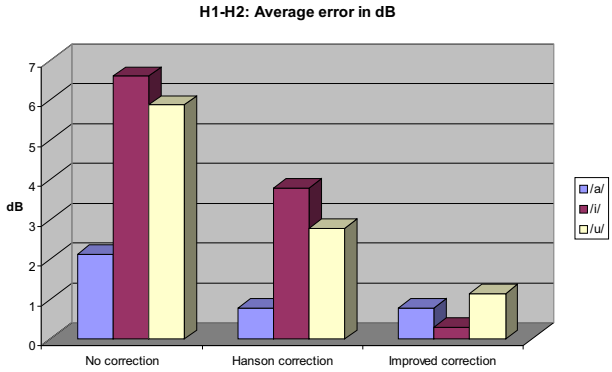

Fig. 3. Comparison bar diagram for average $H_1 - H_2$ error measurements for the three vowels.

A more detailed result can be seen in Fig. 4, where the measure $H_1 - H_2$ in dB is shown as a function of OQ for a female speaking /u/ at a fundamental frequency of 186 Hz. The solid line shows the actual $H_1 - H_2$ (no vocal tract). Again, the average error is independent of OQ, which is manifested in parallel curves. Our formula produced near optimal estimates for $H_1 - H_2$.

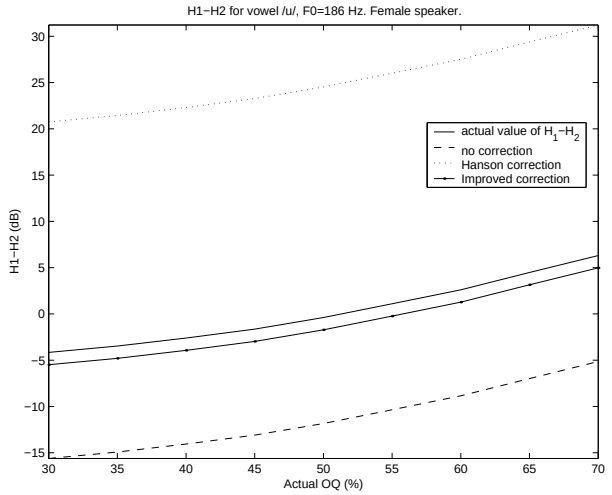

Fig. 4. $H_1 - H_2$ in dB for a synthetic female /u/ with $F_0 = 186$ Hz. Actual value (solid line), no correction (dashed line), with the correction in [1] (dotted line), our correction (solid dotted line). Note that the estimation errors are independent of OQ.

To get an idea on what F_0 ranges are critical for Hanson's correction, Table 4 lists ranges of F_0 where Hanson' correction is more than 3 dB off from the actual value of $H_1 - H_2$. As expected, for the vowel /a/ Hanson's correction formula works very well.

6. SUMMARY AND CONCLUSION

An improved correction formula for the estimation of source harmonics' magnitudes is presented with the goal of obtaining a more accurate estimate of OQ. The new formula accounts for the bandwidths of all vocal tract resonances and is not limited to the analysis of non-high vowels as is the case in [1]. For vowels, synthesized with the LF and KLGLOTT88 models, the new technique can estimate $H_1 - H_2$ almost perfectly. Furthermore, we have shown that the average $H_1 - H_2$ error is independent of OQ. To analyze non-synthetic speech signals, future research will include an automatic and reliable estimation of formant frequencies and their bandwidths.

Table 4. *Ranges of F_0 in Hertz where Hanson's $H_1 - H_2$ error is greater than 3dB.*

Vowel	Male	Female	Child
/a/	–	–	–
/i/	116–154	136–174	167–205
/u/	129–170	167–205	195–235

7. REFERENCES

[1] H. M. Hanson, *Glottal characteristics of female speakers*, Ph. D. Dissertation, Harvard University, Cambridge, MA, 1995.

[2] A. N. Chasaide and C. Gobl, *The handbook of phonetic sciences*, chapter Voice Source Variation, pp. 428–461, Blackwell Publishers Inc., 1997.

[3] 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," *J. Speech Hear. Res.*, vol. 38, pp. 1212–1223, 1995.

[4] H. M. Hanson, "Glottal characteristics of female speakers: Acoustic correlates," *J. Acoust. Soc. Am.*, vol. 101, pp. 466–481, 1997.

[5] H. M. Hanson and E. S. Chuang, "Glottal characteristics of male speakers: Acoustic correlates and comparison with female data," *J. Acoust. Soc. Am.*, vol. 106, pp. 1064–1077, 1999.

[6] Y. Qi and N. Bi, "A simplified approximation of the four-parameter lf model of voice source," *J. Acoust. Soc. Am.*, vol. 96, no. 2, pp. 1182–1185, August 1994.

[7] N. Henrich, C. d'Alessandro, and B. Doval, "Spectral correlates of voice open quotient and glottal flow asymmetry: theory, limits and experimental data," in *Proceedings of EUROSPEECH*, 2001, Scandinavia.

[8] D. H. Klatt and L. C. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *J. Acoust. Soc. Am.*, vol. 87, no. 2, pp. 820–857, February 1990.

[9] H.-L. Lu and J. O. Smith III, "Joint estimation of vocal tract filter and glottal source waveform via convex optimization," in *Proceedings of the IEEE workshop on applications of signal processing to audio and acoustics*, Piscataway, NJ, USA, 1999, pp. 79–82.

[10] G. E. Peterson and H. L. Barney, "Control methods used in a study of the vowels," *J. Acoust. Soc. Am.*, vol. 24, no. 2, pp. 175–184, March 1952.

[11] R. H. Mannell, "Formant diphone parameter extraction utilising a labelled single speaker database," in *Proceedings of the ICSLP*, 1998, Sydney, Australia.