Estimating Glottal Parameters in Nasalized Speech: An Analysis by Synthesis

James Mahshie† and Christer Gobl‡

† Gallaudet University, Department of Audiology and Speech-Language Pathology Washington, DC, USA

‡ University of Dublin, Trinity College, Centre for Language and Communication Studies, Dublin, Ireland

E-mail: james.mahshie@gallaudet.edu, cegobl@tcd.ie

ABSTRACT

The present study examines the extent to which increased nasal coupling affects estimates of glottal parameters, when derived from inverse filtering based on an all-pole assumption of the vocal tract. The analysis involved synthesis of five steady-state vowels using the HLSYN quasi-articulatory synthesizer. For each vowel, the parameter controlling the nasal aperture, ***An***, was systematically varied from 0 to 100 mm^2. The acoustic signal for each utterance was subsequently inverse filtered and estimates were made for a range of glottal source parameters. Preliminary results suggest that for most vowels, many of the source parameter estimates remain relatively unaffected by increased nasal coupling. However, for the high vowels (the high front vowel in particular) substantial changes in the source estimates were found in some of the source parameters with increasing degree of nasal coupling.

1 INTRODUCTION

Inverse filtering is a technique for obtaining a glottal source signal by eliminating the acoustic effects of vocal tract resonances. It has been used to examine the voice parameters of both normal speech, e.g. [4], and disordered speech, e.g. [7]. Because of the all-pole assumptions typically made about the vocal tract in the filtering schema, however, the method is in theory limited to utterances that are non-nasalized.

There have been attempts to inverse filter nasalized utterances using a more complex filter schema, e.g. [5]. However, this approach is difficult, particularly when a formant and a nasal zero fall in close proximity to each other. Others, e.g., [1], have suggested that the all-pole model provides fairly robust results for utterances that are moderately nasalized. No research to date has been reported that quantifies the extent and nature of changes in glottal source parameter estimates from inverse filtered utterances containing varying degrees of nasalization. Accordingly, the present study examines the extent to which systematically varying the degree of nasal coupling affects estimates of glottal parameters when derived from inverse filtering based on an all-pole approximation of the vocal tract transfer function.

2 GENERAL PROCEDURES

2.1 Synthesis of utterances

To enable careful control of nasal coupling, the HLSYN synthesizer [9] was employed to generate the speech utterances. HLSYN is a quasi-articulatory speech synthesizer in which a small set of parameters control the Klatt formant synthesizer KLSYN88 [6]. HLSYN permits control of 13 high level parameters including F1, F2, F3, F4, and nasal aperture. Each high level parameter of HLSYN in turn controls multiple KLSYN88 parameters. An appealing aspect of HLSYN is that it controls KLSYN88 parameters in ways that are physiologically feasible, and thus provides a realistic means of synthesizing utterances.

The degree of nasal coupling is set by a parameter called ***An***, which varies the size of the nasal aperture from 0 mm^2 to 100 mm^2. This parameter causes changes primarily in three KLSYN88 parameters: FNP, BNP and FNZ, i.e. the frequency and bandwidth of the nasal pole, and the frequency of the nasal zero.

To examine the effect of different degrees of nasal coupling on different vowels, the following vowels were synthesized: /a/, /i/, /u/, /ɑ/ and /ə/. The formant frequencies, given in Table 1, are based on Swedish vowels.

Vowel	F_1 (Hz)	F_2 (Hz)	F_3 (Hz)	F_4 (Hz)
/a/	750	1250	2500	3350
/i/	250	2200	3150	3750
/u/	300	600	2350	3250
/ɑ/	600	950	2550	3330
/ə/	500	1500	2500	3500

Table 1: Formant frequency input values for HLSYN.

For each vowel quality, 15 versions were synthesized with the nasal coupling parameter ranging from 0 to 100 mm^2. The specific ***An*** values employed were 0, 5, 10, 15, 20, 25, 30, 35, 40, 50, 60, 70, 80, 90 and 100 mm^2. The output sampling rate was 10 kHz and the number of formants was 5 (NF = 5).

KLSYN88 offers a choice of three different glottal source models. Here we used the default source model, KLGLOTT88, for the generation of the glottal source signal. This model has three control parameters: AV, amplitude of voicing, which sets the overall amplitude of

ISBN 1-876346-48-5

the pulses; OQ, the open quotient; and TL, which controls the amount of spectral tilt of the source spectrum. The TL value is the additional attenuation in dB at 3 kHz. In HLSYN, the source parameters were set as follows: AV = 60 dB, TL = 5 dB, OQ = 50%. The skew of pulses generated by KLGLOTT88 is constant [9]. For each utterance, approximately 500 ms was synthesized and then analyzed as indicated below.

2.2 Glottal waveform analysis

To examine the glottal waveform characteristics of each synthesized utterance, they were initially interactively inverse filtered to obtain the source waveform [2]. Then the LF model [3] was matched to the glottal pulses derived from the inverse filtering. This process involves manipulating the shape of the LF model by adjusting six cursors in the time domain, specifying five time-points and one amplitude point. From these points the LF waveform is generated and superimposed on the estimated glottal waveform [2]. The matching software not only shows the match in the time domain, but also allows for comparisons of the corresponding spectra. The frequency domain matching is essential for achieving good estimates of certain source parameters, and is particularly useful when trying to optimize the matching when a perfect temporal match is impossible.

The inverse filtering strategy employed uses five anti-formants, one for each of the formants used in the synthesis. Furthermore, formant values were constrained to remain realistic for a given vowel quality. For instance, the anti-formant used for F5 could not be 'reassigned' to cancel the nasal pole.

Given these constraints, the optimization criterion in the time domain was to achieve maximum cancellation of oscillations in the glottal closed phase. In the frequency domain, the aim was to maximally cancel the spectral peaks while retaining realistic spectral continuity.

Since the inverse filtering parameters were constant for each utterance, only one pulse from each utterance would need to be analyzed. However, to minimize possible measurement errors, three pulses were analyzed and results were averaged.

3 RESULTS

To examine the effect of altering ***An*** on various glottal source features, a number of different source parameters were estimated. The results for a subset of these – *EE*, *RA*, *RG*, *RK*, and *OQ* – are presented here.

The settings used for AV and TL were altered by HLSYN in some of the utterances. AV was in certain cases lowered by one dB. TL values ranged from 2 dB to 8 dB. These variation would obviously affect our source estimates, and were thus compensated for, as indicated below.

3.1 Excitation Strength, *EE*

EE is a measure of the strength of the glottal excitation, as defined by the negative amplitude at the time of maximum discontinuity of the differentiated glottal pulse. This amplitude is typically equivalent to the maximum negative amplitude of the differentiated glottal pulse.

Figure 1 presents the results for the *EE* parameter, after normalization. This was simply a matter of adding 1 dB to the original *EE* values from the analysis for the utterances where HLSYN imposed a lowering of AV by one dB.

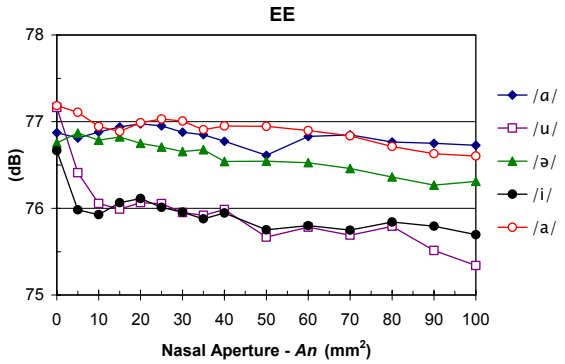

Figure 1: Normalized *EE* values for all utterances.

The variation in *EE* was on the whole small as a function of ***An***. The general trend was for a small reduction in the *EE* estimates with increasing nasal coupling. This drop was larger for the high vowels (1.9 dB for /u/), but most of the change occurred between 0 and 10 mm^2 nasal aperture. For /a/, /ə/, and /ɑ/ the change was 0.6 dB or less. Note that for utterances with ***An*** = 0, there were differences among the five vowels of approximately 0.5 dB.

3.2 Dynamic Leakage, *RA*

The *RA* parameter is a measure corresponding to the residual flow during the return phase, which occurs from the time of the main excitation to the time of complete or maximum closure.

Figure 2 shows the *RA* data after normalization for variation in TL. The values have been normalized to correspond to a TL value of 5 dB.

The effect on the source spectrum of the LF model return phase is approximately that of a first order low-pass filter with a cutoff frequency $FA = f_0/(2\pi(RA/100))$ [Hz]. Given its definition, TL can be expressed as a function of *FA* as follows:

$$\mathrm{TL}(FA) = 10 \cdot \log_{10}\left(1 + \left(\frac{3000}{FA}\right)^2\right) \quad [\mathrm{dB}] \qquad (1)$$

From this equation we can derive an expression for the normalized *RA*, given a positive TL change, ΔTL [dB]:

$$RA_{\mathrm{norm}}(\Delta\mathrm{TL}) = RA \cdot 10^{\frac{\Delta\mathrm{TL}}{20}} \cdot \frac{\sqrt{3000^2 + FA^2\left(1 - 10^{\frac{-\Delta\mathrm{TL}}{10}}\right)}}{3000} \qquad (2)$$

The data in Figure 2 show some striking effects on the *RA* estimates. Estimates of *RA* for the high vowels were most affected. For /i/, values range between 1.4% (***An*** = 0) to 6.5% (***An*** = 90 mm^2). For /u/, the trend is similar, but the *RA* increase plateaus at ***An*** = 30 mm^2. Although the variability across values of ***An*** were smaller for /a/ and /ə/, the relative amount of the changes were nevertheless substantial. The smallest variation in *RA* was for /ɑ/. The variation in *RA* estimates for utterances with no nasal coupling was between 1.0% and 1.4%.

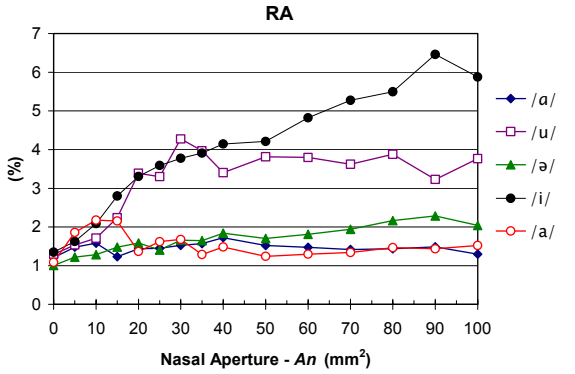

Figure 2: Normalized RA values for all utterances.

3.3 Relative glottal frequency, *RG*

RG is the relative glottal frequency, FG/f_0, where the glottal frequency, FG, has a period time defined as double the duration of the opening phase of the glottal pulse [2]. *RG* estimates were quite similar throughout the range of ***An*** values for the vowels /a/, /ə/, and /ɑ/. For /u/, *RG* values decreased for ***An*** between 15 and 30 mm^2. Beyond 30 mm^2 the parameter values were fairly constant, although smaller than for the other vowels. For /i/, *RG* varied in a rather complex way. For increasing ***An*** to 25 mm^2, values tended to drop. At ***An*** = 30 mm^2, the *RG* value then increased by about 10% and stayed relatively high up to ***An*** = 40 mm^2. For further increases in ***An***, *RG* values again dropped. The *RG* variation for ***An*** = 0 was very small: 143-144%.

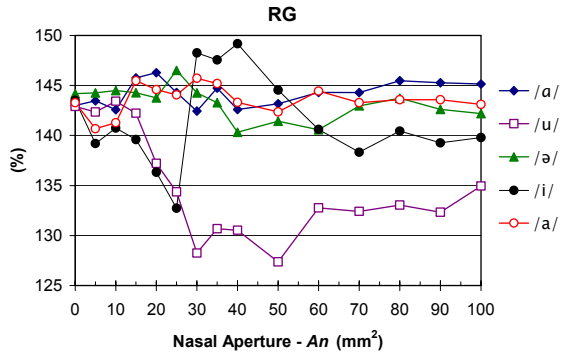

Figure 3: *RG* values for all utterances.

3.4 Glottal skew, *RK*

The degree of asymmetry of the glottal pulse (glottal skew) is here measured by the *RK* parameter. *RK* is defined as the duration of the closing portion of the pulse (from the timepoint of peak glottal flow to the timepoint the main excitation) as a proportion of the duration of opening portion of the pulse (from the timepoint of glottal opening to the timepoint of peak glottal flow). Thus *RK* = 100% means a perfectly symmetrical pulse and smaller *RK* values indicate a more skewed glottal pulse.

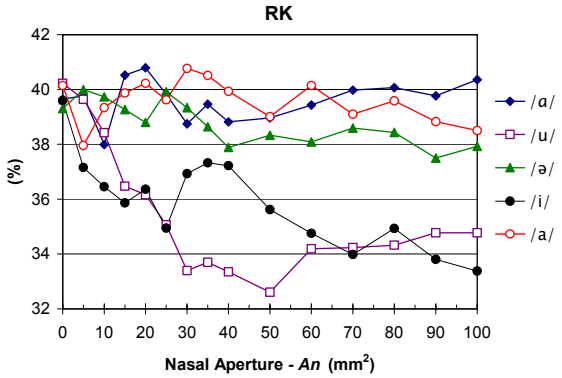

Figure 4: *RK* values for all utterances.

The variation in *RK* as a function of ***An*** was relatively small. For /a/, /ə/, and /ɑ/, the changes are within 3 percent or less. Again, the high vowels tend to display the most variation with increasing ***An***, there being a tendency for *RK* values to decrease (increase in glottal skew) with increasing ***An***. Some of the change in *RK* is of course a direct consequence of the variation in *RG*. In the case of no nasal coupling, *RK* values were very consistent; values varied only between 39.3% and 40.2%.

3.5 Open Quotient, *OQ*

The open quotient represents the portion of the glottal cycle when the glottis is open. The *OQ* values presented here, however, do not include the duration of the return phase. This measure conforms to the definition of *OQ* in KLSYN88. This *OQ* measure is completely determined by *RG* and *RK*.

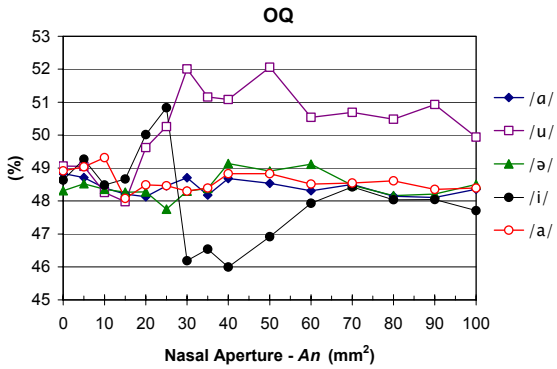

Figure 5: *OQ* values for all utterances.

Thus, it is not surprising that the vowels /a/, /ə/, and /ɑ/

ISBN 1-876346-48-5

showed consistent estimates for *OQ* throughout the range of ***An*** values. For /u/ and /i/, the variation in *OQ* appears mainly a consequence of the variation in *RG* (i.e. variation in the duration of the opening portion of the glottal cycle). Also for *OQ* the results were very consistent for utterances with no nasal coupling; these values were between 48.3% and 49.1%.

4 DISCUSSION AND CONCLUSIONS

The present study examined the effect of nasal coupling on the accuracy of glottal source measures obtained through inverse filtering based on an all-pole assumption of the vocal tract. For the non-high high vowels, there appears to be little effect of increasing ***An*** on the source parameters, regardless of degree of nasal coupling (with exception perhaps of *RA*). It is tentatively concluded that for non-high vowels, reasonably accurate source estimates can be obtained from the analysis procedure outlined here, even with significant degrees of nasal coupling.

Findings thus far further suggest that glottal source estimates of some parameters, such as *RA,* can be adversely affected by nasal coupling. *RA* is a particularly important determinant of the slope of the source spectrum.

The results also suggest that the extent to which glottal source estimates are affected is related to the specific vocal tract configuration. In these simulations, the glottal source estimates of the high vowels /i/ and /u/ appear to be particularly affected by the degree of nasal coupling.

The difference between vowels can be related to two factors. The first, and probably the main factor, is the extent to which the frequencies of the nasal pole and zero diverge. When they are close in frequency they have minimal effect on the spectrum; the more they separate, the greater the influence on the spectrum (making all-zero inverse filtering more problematic). In HLSYN the separation of the pole/zero pair varies as function of vowel quality: the maximum separation is for /i/, the smallest is for /a/. The second factor contributing to the differences among vowels concerns the frequency distance between F1 (and sometimes F2) and the nasal pole. When the nasal pole is relatively close to F1 (and/or F2), the effect can be more readily compensated for by adjusting some of the parameters of the all-zero filter. For /i/ in particular, the nasal pole occurs far from both F1 and F2.

Detailed examination of the nasal pole/zero data, the F1 data used in the synthesis, and the data used in the inverse filtering to cancel F1, is currently being carried out to explore the basis for variation in these source estimates.

There was some variation in the source estimates for *EE* and *RA* even when there was no nasal coupling. Some small amount of spread in the measurement is to be expected, but the variations also suggest that there could be differences in the source signal used in the synthesis. As mentioned, AV and TL varied as consequence of the specific settings used for the HLSYN parameters. Although this variation was compensated for (approximately in the case TL), one could envisage other unknown interactions between synthesis parameters, which may lead to variations in the synthesized source signal. To investigate the potential of such effects, an analysis of the actual source signal used in the synthesis is currently under way.

The results reported here are of course only valid insofar as the nasality synthesized in HLSYN represents realistic simulations of nasality in real speech. Further research based on the analysis of natural speech utterances will be necessary to corroborate these results.

REFERENCES

[1] T.V. Ananthapadmanabha, "Acoustic analysis of voice source dynamics", *STL-QPSR*, Speech, Music and Hearing, Royal Institute of Technology, Stockholm, vol. 2-3/1984, pp. 1-24, 1984.

[2] G. Fant, "Glottal source and excitation analysis", *STL-QPSR*, Speech, Music and Hearing, Royal Institute of Technology, Stockholm, vol. 1/1979, pp. 85-107, 1979.

[3] G. Fant, J. Liljencrants and Q. Lin, "A four-parameter model of glottal flow", *STL-QPSR*, Speech, Music and Hearing, Royal Institute of Technology, Stockholm, vol. 4/1985, pp. 1-13, 1985.

[4] C. Gobl and A. Ní Chasaide, "Voice source variation in the vowel as a function of consonantal context", in *Coarticulation: Theory, Data and Techniques*, W.J. Hardcastle and N. Hewlett, Eds., pp. 122-143. Cambridge: Cambridge University Press, 1999.

[5] I. Karlsson, "Voice quality, male/female variations and speaking style", in *Proceedings of the Speech Maps Workshop,* Esprit/Basic Research Action no. 6975, Institut de la Communication Parlée, Grenoble, vol. 2, pp. 9-13, 1995.

[6] D.H. Klatt and L.C. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers", *Journal of the Acoustical Society of America*, vol. 87, pp. 820-857.

[7] J. Mahshie, and A.M. Öster, "Electroglottographic and Glottal Air Flow Measurements for Deaf and Normal-Hearing Speakers", *STL-QPSR*, Speech, Music and Hearing, Royal Institute of Technology, Stockholm, vol. 2-3/1984, pp. 1-24, 1991

[8] A. Ní Chasaide, C. Gobl and P. Monahan, "A technique for analysing voice quality in pathological and normal speech", *Journal of Clinical Speech & Language Studies*, Dublin, vol. 2, pp. 1-16, 1992.

[9] K.N. Stevens and C.A. Bickley, "Constraints among parameters simplify control of Klatt formant synthesizer" *Journal of Phonetics*, vol. 19, pp. 161-174, 1991.