THE PERCEPTUAL EFFECTS OF ALIASING DISTORTION IN GLOTTAL FLOW MODELLING

Zihan Wang, Christer Gobl

Phonetics and Speech Laboratory, School of Linguistic, Speech and Communication Sciences,
Trinity College Dublin, Ireland
zwang7@tcd.ie, cegobl@tcd.ie

ABSTRACT

When modelling the glottal flow signal in a discrete-time system, the aliasing distortion that is produced is typically ignored. The assumption is that the perceptual effects are negligible if the sampling frequency is sufficiently high. In this paper, we implement a recently developed version of the LF model, which eliminates aliasing distortion. By comparing it to the standard application of the LF model, which introduces aliasing distortion, we can explore if, and to what extent, aliasing distortion is perceptible in the modelled voice source signal. The results of a listening test demonstrate that the aliasing distortion is almost always perceptible, even when the sampling frequency is relatively high. The perceptual artefacts are very noticeable when the fundamental frequency is high, particularly in combination with a tense voice quality. In conclusion, therefore, if high-quality modelling of the glottal flow signal is required, aliasing-free source modelling is recommended.

Keywords: glottal flow, voice source generator, aliasing distortion, LF model, voice quality.

1. INTRODUCTION

The glottal flow signal, i.e. the voice source, plays a fundamental role is speech communication. It contributes to the linguistic prosody, such as variations in prominence, accentuation and phrasing e.g., [1, 2, 3]. Furthermore, variation in the voice source carries paralinguistic information, which signals interpersonal information concerning the speaker's state, attitude to the interlocutor and to the discourse [4, 5, 6]. It also coveys extralinguistic information, such as specific characteristics of a speaker's voice, e.g., [7, 8, 9].

Accurate and detailed modelling of the source is therefore important. For the analysis of the glottal flow, a parametric model is often used, where the source parameter data, which define the model waveform, serve as useful descriptors of the source characteristics. Source models can also be used in synthesis, e.g., for the purpose of voice transformation or for the generation of acoustic stimuli to explore the perceptual role of the voice source [10-14].

One aspect of glottal flow modelling generally overlooked is the effect of the aliasing distortion which is produced when the model waveform is sampled. Most source models are defined in the time domain as a set of piecewise elementary functions e.g., [15-21]. Given the discontinuities that will occur in the derivative of these functions, the spectrum is not bandlimited, and thus aliasing will always be produced. Components above the Nyquist frequency (half the sampling frequency) will appear as components at frequencies below the Nyquist frequency, thereby distorting the spectrum.

Since the amplitude level of the source spectrum decreases with increasing frequency, the amount of aliasing is less when the sampling frequency is high. It is therefore often assumed that if the sampling frequency is sufficiently high, the effect will be relatively small and can be disregarded.

However, the perceptual consequences of this distortion in the modelling have not been systematically studied. Hence, in this paper a perception test was carried out to explore the effect of the aliasing distortion and the extent to which it is perceptible.

Two sets of voice source stimuli were used, involving pairs of stimuli which are identical except for the aliasing distortion being present in one and not in the other. Among the stimuli further variables were manipulated to explore other factors that might impact on the perceptibility of aliasing. These tested the following hypotheses: aliasing is consistently audible, even at high sampling frequencies; aliasing is more strongly perceived when the sampling frequency is low and when the f_0 level is high; it is more strongly perceived for a tense voice quality with strong higher harmonics than for a lax voice quality with weak higher harmonics; voice source dynamics might influence the perceived aliasing, the initial expectation being that greater dynamics would results in the aliasing being less strongly perceived, since it is conceivable that there would be greater masking of the aliasing components in these conditions.

2. METHODOLGY

Recently, an alternative discrete-time implementation of the widely used Liljencrants-Fant (LF) model [17]

was developed, which eliminates the aliasing distortion present in the standard application [22]. By comparing the two versions of the model, it is possible to study the perceptual effects of the aliasing distortion in the glottal flow signal.

2.1. The LF model of glottal flow

The LF model is defined by the piecewise function shown in (1).

$$(1)\quad U_g'(t) = \begin{cases} E_0\, e^{\alpha t} \sin \omega_g t & 0 \le t < t_e \\ -\dfrac{E_e}{\varepsilon\, T_a}\left(e^{-\varepsilon(t-t_e)} - e^{-\varepsilon\, T_b} \right) & t_e \le t < t_c \end{cases}$$

It models the derivative of the glottal flow pulse and involves two sub-functions. The first is a segment of an exponentially growing sinusoid, which models the flow derivative during the open phase of the glottal cycle. The duration of the open phase is T_e, and at the end of the open phase the amplitude of the waveform is $-E_e$, which is the negative amplitude of the main glottal excitation (see Fig. 1).

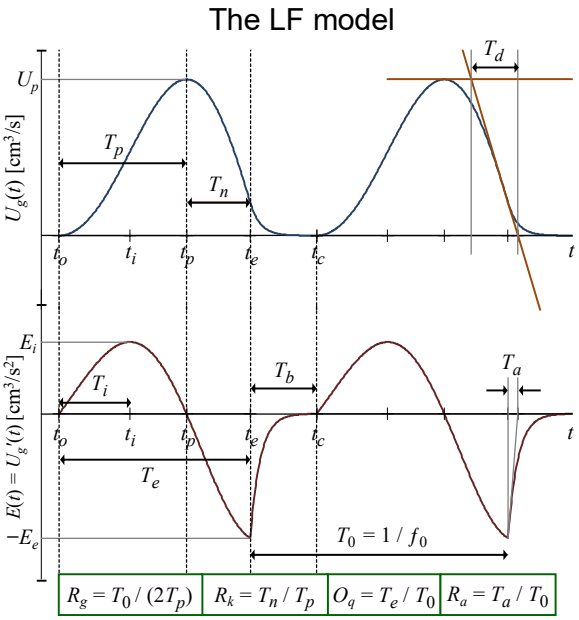

Figure 1: Two LF pulses and parameter definitions. Flow derivative (bottom) and corresponding glottal flow (top).

The second sub-function is a segment of an exponential function, which models the flow derivative during the return phase, i.e. the part of the cycle for which the vocal folds return to full or maximum closure. The duration of this segment is $T_b = t_c - t_e$ (Fig. 1).

The LF model pulse shape is often described by the parameters R_g, R_k and R_a (Fig. 1). Together with E_e, these R-parameters are sometimes called 'the LF-parameters'. However, as has been discussed in [23], the actual parameters of the model, which are required to generate the LF pulse, are E_e, T_e, ω_g, α, ε and T_b.

2.2. The LF model without aliasing

The LF glottal waveform is continuous in time, and when digitised, this is typically done by sampling the time domain functions in (1). The sampling process will introduce aliasing distortion, since the spectrum of the LF model pulse is not bandlimited, and consequently it will always extend beyond the Nyquist frequency regardless of how high the sampling frequency is. Thus, the standard discrete-time version of LF model will have a distorted frequency spectrum.

A method is presented in [22] which avoids aliasing distortion in voice source models and a mathematical framework for generating an aliasing-free version of the LF model is outlined. To avoid aliasing distortion, the model needs to be described in the frequency domain, which can be done by deriving the Laplace transform of the model. This enables the calculation of the true spectrum of the model directly from the model parameters.

The ideal discrete spectrum of the model can then be obtained by sampling the true model spectrum up to the Nyquist frequency. Given this ideal spectrum, which effectively bandlimits the model spectrum to the Nyquist frequency, no aliasing is introduced. The discrete-time LF pulse corresponding to the ideal discrete spectrum is derived by applying the inverse discrete Fourier transform (IDFT). For full details on how the aliasing-free LF model is calculated, see [22].

2.3. The voice source generator

To produce the stimuli to explore the perceptual effects of the aliasing distortion introduced by the sampling process, a system for generating voice source signals was developed. This system, referred to as the voice source generator (VSG) was designed using the MATLAB App Designer [24].

The VSG currently incorporates the two version of the LF model described above: the standard version with aliasing and the aliasing-free version, implemented according to [22].

Furthermore, the VSG offers two different sets of control parameters for the source modelling: one set is the commonly used 'LF parameters', i.e. E_e, R_g, R_k, R_a and f_0. The other set of parameters is R_d, E_e, O_q and f_0. Note that only four parameters need to be specified in this case: R_a is derived from the R_d value using the formula provided in [25].

$R_d = 0.11^{-1} f_0 U_p / E_e = 0.11^{-1} T_d / T_0$ (see Fig. 1) is the global waveshape parameter proposed by Fant [25, 26], which is the central parameter in the alternative parameter control system for the LF model. The R_d

value has been shown to be a good indicator of the tense-lax dimension of voice quality, being low for tense voice and high for lax voice. O_q is the glottal open quotient, here defined as T_e/T_0 (Fig. 1).

To ensure that a possible LF pulse is always produced from the input parameters, parameter constraints as defined in [27] are imposed on the input values to the VSG. Furthermore, amplitude modulated aspiration noise can be added to the voice source pulses, where the modulation is determined by the glottal pulse shape according to [28].

The VSG can also generate LF pulses from the analysis output of the GlórCáil system, a system that can automatically extract LF parameter data from a speech signal [13, 29] or from the output the ISF system, a manual interactive inverse filtering and source modelling system described in [30, 31, 32].

In the future, it is envisaged that additional source models will be included in the VSG, as well as an extended choice of control parameters.

2.4. Acoustic stimuli

The voice source generator described above was used to create 38 acoustic stimuli. These stimuli made up 19 pairs, where the only difference between the stimuli in each pair was the version of the LF model: one producing aliasing and the other aliasing-free.

Out of these 19 pairs, two sets of 8 pairs were generated by specifying input values using the second option of control parameters, i.e. R_d, E_e, O_q and f_0 (aspiration noise was not used). The two sets differed only in terms of the sampling frequency, 10 kHz vs. 20 kHz, to test the hypothesis that aliasing is less perceptible at a higher sampling frequency. These are referred to as stimulus pairs 1 to 8 (see Table 1).

To test if a source signal with higher f_0 and a tenser voice quality produces aliasing that is more perceptible, stimuli 1 to 6 (both sets) differed in terms of f_0 level and voice quality: there were three f_0 levels with either lax or tense settings. Stimuli 1 and 4 had low f_0 (mean f_0 109 Hz), stimuli 2 and 5 medium f_0 (mean f_0 219 Hz), stimuli 3 and 6 high f_0 (mean f_0 437 Hz), stimuli 1 to 3 had high R_d (lax voice, mean R_d was 2.1), stimuli 4 to 6 had low R_d (tense voice, mean R_d was 0.58). These 6 stimuli (both sets) all had relatively little variation in the parameter values (small dynamics). To test the effect of the dynamic variation in the source, stimuli 7 and 8 (both sets) on the other hand involved relatively large dynamics. The only difference between stimuli 7 and 8 was the f_0 level: relatively low for stimulus 7 (mean f_0 166 Hz) and relatively high for stimulus 8 (mean f_0 322 Hz). The R_d mean was 1.2 for both stimuli 7 and 8.

So far, all stimuli are based on stylised synthetically generated source signals. The remaining three pairs of stimuli were generated from source data obtained in a previous study [33] derived from natural speech using the manual interactive ISF system. The three utterances analysed were of the sentence “We were away a year ago” spoken by a male speaker in neutral, angry and sad voice. These stimuli were added to ascertain that the findings on the other stimuli should hold for naturally spoken utterances.

2.5. Listening test

To evaluate the perceptual effect of the aliasing in glottal flow modelling, a listening test was carried out with the 19 pairs of stimuli described above. The test was administered online, and participants were instructed to take the test in a quiet room using high-quality headphones. They were also informed that they could listen to each audio file as many times as they wished. Before the main test, a file containing all the test stimuli was presented. This allowed the participants to familiarise themselves with the range of differences involved in the different pairs of stimuli and to adjust the volume to a comfortable level.

Participants then proceeded to the main test to assess the perceived degree of difference between the pairs of stimuli. To do this, they used a scale from 0 to 5. If no difference between the stimuli was perceived, 0 would be selected. If a difference was perceived, the strength of the perceived difference was rated between 1 and 5, where 1 meant a very small difference and 5 a very large difference (approximately the maximum differences found among the pairs of stimuli). The stimuli were presented in a randomised order. The order of the two stimuli (with or without aliasing) in each pair was also random.

3. RESTULTS AND DISCUSSION

In total, 30 participants completed the listening test. The means and standard deviations of the perceived difference scores for each pair of stimuli are presented in Table 1. One-way analysis of variance was used to test the significance of differences between groups of stimuli. The results show that the aliasing is perceptible in both the stylised stimuli and the stimuli generated using natural spoken data – even at a relatively high sampling frequency. The difference between the difference score and 0 (imperceptible) is statistically significant ($p < 0.001$) for all pairs of stimuli, including the sad stimuli, for which the difference score was the smallest.

Given that the amount of aliasing is less at a high sampling frequency, we would expect the perception scores to reflect this. The overall mean score was indeed lower at the higher sampling rate (Table 1), but the difference was smaller than expected, and was not statistically significant ($p = 0.072$) for these data.

Table 1: Mean and standard deviation values of the difference scores for the pairs of stimuli.

f_s	10 kHz		20 kHz	
Stimuli	*Mean*	*Std*	*Mean*	*Std*
1	1.50	1.18	1.37	1.28
2	2.57	1.43	2.13	1.45
3	2.83	1.46	2.97	1.64
4	1.43	1.28	1.20	1.22
5	2.87	1.36	2.13	1.63
6	3.33	1.58	3.03	1.66
7	1.97	1.45	1.70	1.39
8	2.60	1.52	2.47	1.67
Overall	**2.39**	1.41	**2.13**	1.49
Angry	1.83	1.21		
Neutral	2.63	1.40		
Sad	0.77	0.96		
Overall	**1.74**	1.19		

One would also expect that a high f_0 level would make the aliasing more audible, since when f_0 is high the harmonics close to the Nyquist frequency would be relatively stronger compared with the lower harmonics. Furthermore, the increased frequency gap between the harmonics could potentially lead to less masking of the aliasing components.

This expectation was indeed borne out by the results. Fig. 2 shows that the degree of perceived difference is greater as f_0 increases. The statistical analysis suggests that this effect is significant ($p < 0.01$).

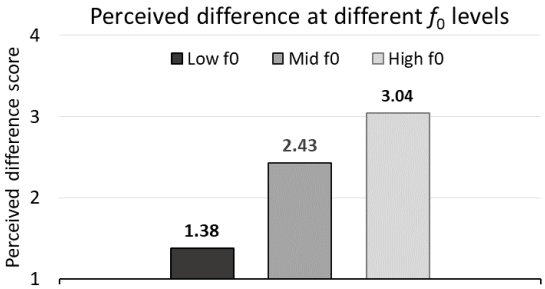

Figure 2: Mean difference scores for stimuli with low, medium and high f_0.

It was also hypothesised that voice quality could affect the perceptibility of aliasing, as a tense voice, with stronger higher harmonics, would be expected to cause more aliasing than a lax voice. Comparing the tense and lax stimuli (stimuli 1-3 vs. 4-6), the overall difference in mean ratings was small (tense 2.33, lax 2.23) and not statistically significant ($p = 0.54$).

However, if one compares only the 10 kHz stimuli for the tense vs. lax voice at low, medium and high f_0 levels (Fig. 3), a more nuanced interpretation is suggested. At low f_0 levels, there is essentially no difference in the scores. At mid f_0 level, a difference does emerge, with tense voice yielding a greater degree of perceived aliasing. At high f_0 level, the difference is greater again. We would tentatively conclude therefore that voice quality plays some role if the sampling frequency is low, with aliasing effects becoming more perceptible with tense voice as f_0 increases.

It was also suggested in the Introduction that a dynamically changing source signal could help mask aliasing components, thus making the aliasing less audible. However, our results do not support this hypothesis. Although the mean rating was slightly lower (2.18 vs. 2.32) for the stimuli with high dynamics, the difference was not significant ($p = 0.48$).

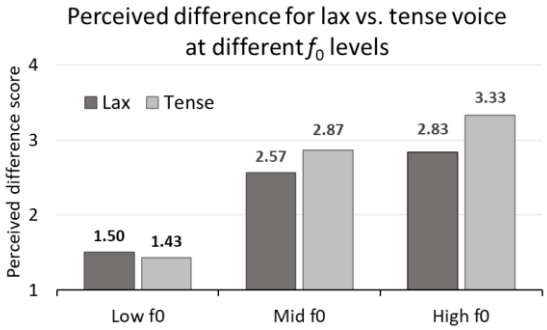

Figure 3: Mean difference scores for stimuli with different voice tension at low, mid and high f_0 ($F_s = 10$ kHz).

Finally, results for the stimuli based on natural spoken utterances showed similar trends to the stylised stimuli. The mean rating may seem somewhat lower, but this is most likely due to the relatively low f_0 values in these stimuli. The mean f_0 across the three stimuli is 106 Hz, comparable to stimuli 1 and 4, which in fact have somewhat lower scores (Table 1). We can therefore be reasonably confident that differences perceived in the stylised voice source stimuli reflect those we would get from real speech data.

4. CONCLUSIONS

The perceptual effect of aliasing distortion is explored by comparing the voice source signals produced by two version of the LF glottal flow model: the standard version which introduces aliasing distortion and a novel implementation based on a frequency domain representation of the model [22]. The perception test shows that the aliasing is almost always perceptible, even when the sampling frequency is high. The distortion is strikingly noticeable when f_0 is high, particularly in combination with tense voice quality. The results therefore suggest that aliasing-free source modelling is beneficial when high-quality modelling of the glottal flow signal is required.

5. ACKNOWLEGEMENTS

This research was carried out in the Róbóglór and ABAIR projects, supported by the Irish Govt. Department of Tourism, Culture, Arts, Gaeltacht, Sport and Media, with funding from the National Lottery, supporting An Stráitéis 20 bliain don Ghaeilge 2010-2030.

6. REFERENCES

[1] Sluijter, A. M. C., Shattuck-Hufnagel, S., Stevens, K. N., van Heuven, V. J. 1995. Supralaryngeal resonance and glottal pulse shape as correlate of stress and accent in English. *Proc. 13th ICPhS*, Stockholm, Sweden, 630–633.

[2] Ní Chasaide, A., Yanushevskaya, I., Gobl, C. 2015. Prosody of voice: declination, sentence mode and interaction with prominence. *Proc. 18th ICPhS*, Glasgow, UK, paper no. 0476, 5 pp.

[3] Yanushevskaya, I., Murphy, A., Gobl, C., Ní Chasaide, A. (2022). Global waveshape parameter R_d in signaling focal prominence: Perceptual salience in the absence of f0 variation. Front. Commun. 7, 1–23.

[4] Sundberg, J., Patel, S., Björkner, E., Scherer, K. R. 2011. Interdependencies among voice source parameters in emotional speech. *IEEE Trans. Affect. Comput*. 2, 162–174.

[5] Patel, S., Scherer, K. R., Björkner, E., Sundberg, J. 2011. Mapping emotions into acoustic space: The role of voice production. *Biol. Psychol.* 87, 93–98.

[6] Yanushevskaya, I., Gobl, C., Ní Chasaide, A. 2018. Cross-language differences in how voice quality and f0 contours map to affect. *J. Acoust. Soc. Am.* 144 (5), 2730-2750.

[7] Hanson, H. M. 1997. Glottal characteristics of female speakers: Acoustic correlates. *J. Acoust. Soc. Am.* 101, 466-481.

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

[9] Iseli, M., Shue, Y.-L., Alwan, A. 2007. Age, sex, and vowel dependencies of acoustic measures related to the voice source. *J. Acoust. Soc. Am.* 121 (4), 2283–2295.

[10] Cabral, J. P., Renals, S., Yamagishi, J., Richmond, K. 2011. HMM-based speech synthesiser using the LF-model of the glottal source. *Proc. ICASSP*, 4704–4707.

[11] Carlson, R., Fant, G., Gobl, C., Granström, B., Karlsson, I., Lin, Q. 1989. Voice source rules for text-to-speech synthesis. *Proc. ICASSP*, Glasgow, UK, vol. 1, 223–226.

[12] del Pozo, A., Young, S. 2008. The linear transformation of LF glottal waveforms for voice conversion. *Proc. INTERSPEECH 2008*, Brisbane, Australia, 1457–1460.

[13] Murphy, A., Yanushevskaya, I., Ní Chasaide, A., Gobl C. 2020. Testing the GlórCáil System in a Speaker and Affect Voice Transformation Task," *Proc. 10th Int. Conf. on Speech Prosody,* Tokyo, Japan, 950–954.

[14] Gobl, C., Ní Chasaide, A. 2003. The role of voice quality in communicating emotion, mood and attitude. *Speech Comm.* 40, 189–212.

[15] Rosenberg, A. E. 1971. Effect of glottal pulse shape on the quality of natural vowels. *J. Acoust. Soc. Am.* 49 (2B), 583–590.

[16] Ananthapadmanabha, T. V. 1984. Acoustic analysis of voice source dynamics. *STL-QPSR*, Speech, Music and Hearing, Royal Institute of Technology, Stockholm, 25 (2-3), 1–24.

[17] Fant, G., Liljencrants, J., Lin, Q. 1985. A four-parameter model of glottal flow. *STL-QPSR*, Speech, Music and Hearing, Royal Institute of Technology, Stockholm, 26 (4), 1–13.

[18] Fujisaki, H., Ljungqvist, M. 1986. Proposal and evaluation of models for the glottal source waveform. *Proc. ICASSP*, Tokyo, Japan, 1605–1608.

[19] Milenkovic, P. H. 1993. Voice source model for continuous control of pitch period. *J. Acoust. Soc. Am.* 93 (2), 1087–1096.

[20] R. Veldhuis, 1998. A computationally efficient alternative for the Liljencrants-Fant model and its perceptual evaluation. *J. Acoust. Soc. Am.* 103 (1), 566–571.

[21] Shue, Y.-L., Alwan, A. 2010. A new voice source model based on high-speed imaging and its application to voice source estimation," *Proc. ICASSP*, Dallas, Texas, 5134–5137.

[22] Gobl, C. 2021. The LF model in the frequency domain for glottal airflow modelling without aliasing distortion. *Proc. INTERSPEECH 2021*, Brno, Czechia, 2651–2655.

[23] Gobl, C. 2017. Reshaping the transformed LF model: generating the glottal source from the waveshape parameter R_d. *Proc. INTERSPEECH 2017*, Stockholm, Sweden, 3008–3012.

[24] MATLAB 2020. Version 9.8.0.1451342 (R2020a). Natick, Massachusetts: The MathWorks Inc.

[25] Fant, G. 1995. The LF-model revisited: transformations and frequency domain analysis. *STL-QPSR* 2-3, 119–156.

[26] Fant, G. 1997. The voice source in connected speech. *Speech Comm*. 22, 125–139.

[27] Gobl, C. 2003. *The voice source in speech communication – production and perception experiments involving inverse filtering and synthesis*. Ph.D. thesis, Royal Institute of Technology (KTH), Stockholm.

[28] Gobl, C. 2006. Modelling aspiration noise during phonation using the LF voice source model. *Proc. INTERSPEECH 2006*, Pittsburgh, Pennsylvania, 965–968.

[29] Murphy, A. 2020. *Controlling the Voice Quality Dimension of Prosody in Synthetic Speech using an Acoustic Glottal Model*. Unpublished PhD thesis, Trinity College Dublin.

[30] Gobl, C., Ní Chasaide, A. 2010. Voice source variation and its communicative functions. In: Hardcastle, W. J., Laver, J., Gibbon, F. E. (eds), *The Handbook of Phonetic Sciences* (2nd edition). Oxford: Blackwell, 378–423.

[31] Gobl, C., Ní Chasaide, A. 1999. Techniques for analysing the voice source. In: Hardcastle, W. J., Hewlett, N. (eds), *Coarticulation: Theory, Data and Techniques.* Cambridge: Cambridge University Press, 300–320.

[32] Monahan, P. 1996. *Systems for voice source analysis*. Unpublished PhD thesis, Trinity College Dublin.

[33] Wang, Z., Gobl, C. 2022. Contribution of the glottal flow residual in affect related voice transformation. *Proc. INTERSPEECH 2022*, Incheon, Korea, 5288–5292.