PERCEPTUAL CORRELATES OF SOURCE PARAMETERS IN BREATHY VOICE

Christer Gobl and Ailbhe Ní Chasaide
Centre for Language and Communication Studies, Trinity College, Dublin, Ireland

ABSTRACT

The perceptual relevance of the major source correlates of breathy voice quality were tested with stimuli synthesised using KLSYN88 with the modified LF voice source option. Starting from an auditorily realistic breathy-voiced [a] reference vowel, a series of stimuli was constructed, where the following parameters were varied toward more typical modal values: excitation strength (EE), aspiration noise (AH), open quotient (OQ), speed quotient (SQ), spectral tilt (TL), F1/F2 bandwidths (B1/B2), and fundamental frequency (f_0). These were varied individually and in combinations and tested on 12 subjects. Results suggest that spectral tilt is the main determinant of perceived breathiness. However, a high TL value does not necessarily yield a breathy voiced-percept, as the effect may be offset when the parameters OQ, SQ and B1/B2 are at more modal-like settings. Surprisingly, aspiration noise appeared to contribute relatively little. Differences between these results and those of an earlier study are discussed.

1. INTRODUCTION

The LF model of differentiated glottal flow [1] has been used as an analytic tool in the description of voice quality variation, e.g., [2, 3]. These and other studies report a number of production correlates of breathy voice quality. The aim of the present study was to test the perceptual relevance of these correlates, using the modified LF model of KLSYN88 [4], one of three voice source models offered by this synthesiser. The pulse (shape and amplitude) of the modified LF model is controlled in KLSYN88 by the parameters OQ, SQ, TL and AV.

Klatt and Klatt [4] have tested the relevance of most of these parameters to the perception of breathy voice. In that study the default voice source of KLSYN88 was used (KLGLOTT88). One difference between the two source models is that the latter has one parameter less, SQ. In the modified LF model SQ is used for controlling the skew of the glottal pulse. In KLGLOTT88, the pulse skewing is fixed at 200% (for definition of SQ, see below).

In the Klatt and Klatt experiment, the stimuli were modelled on female speech. The stimuli were constructed on the basis of a modal reference stimulus: the other stimuli involved deviations from the modal values. Listeners were asked to judge how much the altering of a particular parameter in the direction of 'breathy' induced a perception of breathy voice. In some cases more than one parameter was manipulated.

On the basis of informal listening tests, it was decided for this study to use an auditorily good breathy voice stimulus as the reference stimulus. It was felt that modifying (typically) a single parameter in a modal stimulus would not yield a quality sufficiently close to breathy voice, and this would make it harder to judge on the contribution of the individual parameters to the perception of breathiness.

2. STIMULI AND PERCEPTION TEST

17 stimuli were generated for this experiment, and these were modelled on male speech. The reference stimulus was a sustained [a] of 500 ms duration, deemed by the authors to have an auditorily good breathy voice quality. The chosen values were guided by earlier descriptions, such as [2, 3]. The settings for the parameters crucial for the perception test are shown in Table 1, with those of the reference stimulus at the top of the table. The parameters AV, AH and f_0 were time-varying and their values for the reference stimulus are illustrated in Figure 1. Note that the onset and offset of the vowel are gradual: in the first 100 ms, the AV value rises by 6 dB, and in the final 100 ms, the AV value falls by 14 dB. The level of AH also falls in the final 100 ms, but only by 9 dB. The remaining parameters used in the synthesis were common to all the stimuli and were set as follows: F1 = 740 Hz, F2 = 1450 Hz, F3 = 2400 Hz, F4 = 3500 Hz, F5 = 4500 Hz, B3 = 150 Hz, B4 = 200 Hz, B5 = 200 Hz. The parameters DI (diplophonia) and FL (flutter) were not employed. The sampling rate was 10 kHz.

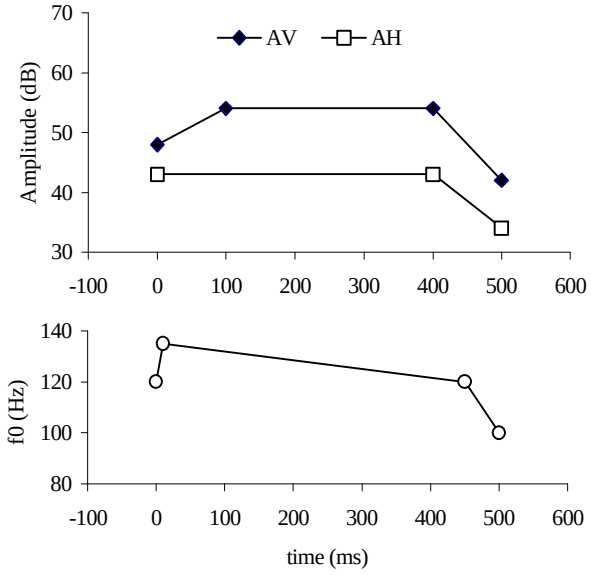

Figure 1. Time-varying parameters AV, AH and f_0.

Stimuli	Parameters EE	TL	OQ	SQ	AH	B1, B2	f_0	On/off
Breathy reference stimulus	ref: 0 dB	25 dB	85%	200%	43 dB	B1/B2 = 300/180 Hz		see text
EE+2dB	+2 dB							
ONOFF								more abrupt
EE+ONOFF	+2 dB							more abrupt
OQ60			60%					
OQ40			40%					
SQ350				350%				
SQ500				500%				
BW						B1/B2 = 60/90 Hz		
F0+7Hz							+7 Hz	
AH-12dB					-12 dB			
AH0					0			
TL15		15 dB						
TL5		5 dB						
LOW			60%	350%		B1/B2 = 60/90 Hz		
HIGH		5 dB			-12 dB			
MODAL	+2 dB	5 dB	60%	350%	-12 dB	B1/B2 = 60/90 Hz	+7 Hz	more abrupt

Table 1. Parameter values which were changed from the reference breathy voice stimulus to generate the other stimuli. On/off, EE, AH and f_0, are time varying parameters (see text).

The other 16 stimuli were generated by changing one or more parameters toward more typical modal values, and these are detailed in Table 1. In a few stimuli (OQ40, SQ500, AH0) values for a particular parameter were more extreme than a typical modal setting, corresponding more to tense voice.

The parameters varied were EE, OQ, SQ, TL, AH, B1/B2 and f_0, and in Table 1 the individual stimuli are referred to in terms of the parameter that was changed relative to the breathy voice reference stimulus. A few stimuli involved a more complex alteration to the reference, and will be explained below.

EE, the excitation strength, is essentially a volume control, but differs from the AV parameter in that it is not a simple scaling factor, but rather corresponds to the amplitude of the glottal pulse at the point of maximum waveform discontinuity. EE cannot be independently controlled in the synthesiser, as the excitation varies not only as a function of AV, but is also affected by the particular settings of OQ, SQ and TL. Therefore, when all parameter settings other than EE had been effected, it was necessary to reset the AV parameter to yield either the same EE as in the reference stimulus, or an increase of 2 dB when this was the targeted value. Note that the smallest AV change permissible in the synthesiser is 1 dB, which means that the targeted EE values were within ± 0.5 dB.

OQ, the open quotient, determines the open to closed ratio in the glottal flow pulse. It was varied as between 85% in the reference stimulus, 60% in the OQ60, LOW and MODAL stimuli, and 40% in OQ40.

SQ, the speed quotient, determines the skew of the glottal flow pulse as the ratio of the opening branch to the closing branch. The opening branch is the interval from glottal opening to the peak glottal flow, and the closing branch is the interval from peak glottal flow to the main excitation. For these stimuli, SQ was 200% in the reference stimulus, 350% in SQ350, LOW and MODAL, and 500% in SQ500.

TL, spectral tilt, determines the slope of the source spectrum: the numeric value of TL is a measure of the additional attenuation in dB at 3 kHz. In the reference stimulus TL was set to 25 dB, and this was changed to 15 dB in TL15 and 5 dB in TL5, HIGH and MODAL. The default value in the synthesiser is 0 dB, but TL = 5 dB was chosen here as the most different value from the reference stimulus for the range we wanted to explore: a TL value of 0 dB yields a rather extreme tense quality.

B1/B2 are the bandwidths of the first and second formants. In the reference stimulus B1 = 300 Hz and B2 = 80 Hz: these values were reduced to the default values of B1 = 60 Hz and B2 = 90 Hz in the BW, LOW and MODAL stimuli.

AH, aspiration noise, varied over time as is illustrated in Figure 1. Its peak value was 43 dB in the reference stimulus and this was reduced by 12 dB in the AH-12dB, HIGH and MODAL stimuli. AH was switched off in AH0.

f_0, fundamental frequency, was also time-varying, as shown in Figure 1. In the stimulus F0+7Hz and in the MODAL stimulus, values throughout the vowel were increased by 7 Hz.

In the stimulus ONOFF, the duration of the onset and offset ramps were altered relative to the breathy voiced reference to

yield a more abrupt onset and offset to the vowel as would be more typical for modal production. Whereas for the reference stimulus, AV rose by 6 dB over 100 ms to its steady-state value, in the ONOFF stimulus, this rise was 3 dB over 50 ms. In the vowel offset, AV dropped from the steady-state value by 8 dB over 50 ms, compared to a drop of 14 dB over 100 ms in the reference stimulus.

The stimuli referred to as LOW and HIGH in Table 1 involve complex parameter settings that simulate changes in spectral balance that occur as a consequence of certain parameters working together. The LOW stimulus involved manipulation to parameters that affect mainly the lower part of the spectrum. In crude terms breathy voice tends to have a strong fundamental component and weak formants (particularly the first formant). The LOW stimulus reverses this as the OQ, SQ and B1/B2 are set closer to default values so that the levels of the fundamental component and the formants should be close to those found in modal voice. In the HIGH stimulus, parameters were changed so as to affect mainly the higher part of the spectrum. TL was lowered to 5 dB and a concomitant drop in AH of 12 dB should give a spectrum similar to modal in the higher frequency range. The stimulus labelled MODAL involved essentially modal-like settings in all the parameters.

12 subjects listened to pairs of stimuli, where the first member was always the breathy-voiced reference stimulus. For the second vowel they were asked to judge the extent to which the perceived breathiness had changed compared to the first vowel on a scale of +1 to -5, where 0 meant no difference in perceived breathiness relative to the reference stimulus, 5 meant the greatest reduction in breathiness, and +1 catered for the possibility that a particular stimulus could appear to be more breathy than the reference.

3. RESULTS

The results of the perception test are presented in Figure 2, where the average responses to the stimuli are shown in the order of decreasing perceived breathiness. Figure 3 shows some of the same information ordered slightly differently, to facilitate comparison of stimuli where only a single parameter was varied in a stepwise fashion, i.e. the parameters TL, AH, OQ and SQ. Note that in Figure 3, absolute AH steady-state levels are given. Thus, AH43 corresponds to the reference stimulus, and AH31 is the same as the stimulus otherwise referred to as AH-12dB.

As can be seen in both figures, TL turned out to be the single most important determinant of perceived breathiness. When TL = 5 dB (i.e., a relatively flat spectral slope) there was always a large reduction in the perceived breathiness. When TL was high (25 dB), subjects tended to hear stimuli as breathy voiced. However, as is clear in Figure 2, a high TL on its own was not a sufficient condition for a high breathiness rating and could be overridden. Note, for the LOW stimulus, which had a steep spectral slope (TL = 25 dB), the responses are much less breathy than for the reference. This appears to be as a result of more modal-like settings for OQ, SQ and B1/B2.

The AH parameter yielded somewhat surprising results. Lowering AH brought about only a small average decrease in breathy-voiced judgements. The extent of the AH reduction, as reflected in the AH-12dB and AH0 stimuli did not seem to make much difference. However, there was considerable variability in responses to these stimuli: of the 12 subjects, 2 found the reduced AH stimuli more breathy than the reference, 6 showed some reduction in breathiness when AH was reduced by 12 dB, but no further reduction when AH was switched off. There was, however, a group of four subjects for whom breathy-voiced judgements decreased in a way that correlated well with the AH reduction.

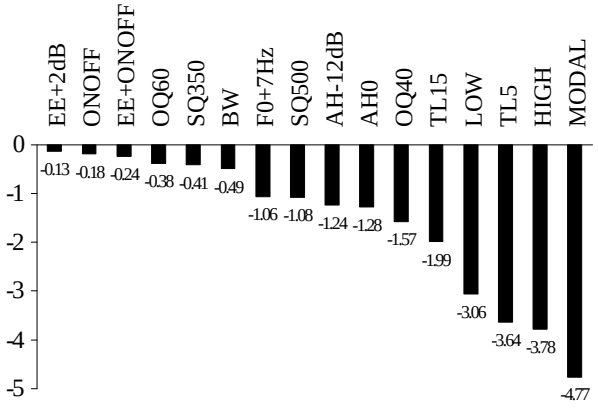

Figure 2. Average ratings for the stimuli in the order of decreasing perceived breathiness. (-5 = least breathy, 0 = as breathy as reference stimulus)

A further indication of the perceptual relevance of AH might be gleaned from the two stimuli TL5 and HIGH, which differ only in the lower AH value of the latter. Average responses differed rather little, suggesting that lower the AH value of HIGH contributes but little to the perceived breathiness. As TL is however low for both of these stimuli, it may be the case that aspiration noise is partially masked in any case.

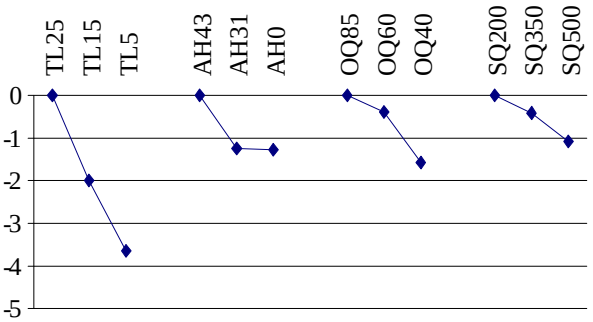

Figure 3. Average ratings for perceived breathiness in stimuli where TL, AH, OQ and SQ were varied in isolation. (-5 = least breathy, 0 = as breathy as reference stimulus)

SQ variation on its own appears to have little effect: only in the extreme case of SQ = 500% is there any appreciable reduction in perceived breathiness. Since the experiment was conducted, however, a parallel study [5] has demonstrated that the SQ range employed here may not have included sufficiently low values to provide a full test of this parameter's contribution.

The reasons for this concerns how the skew parameter in the true LF model maps to the KLSYN88 equivalent, SQ. It would appear that in the Klatt implementation of the LF model, the glottal pulse comes out more skewed than would be expected from the SQ value. For further discussion of this point, see [5].

The OQ parameter appears to have little effect except when it is reduced to the fairly extreme value of 40%. In this instance there is a reduction in perceived breathiness. Reducing bandwidth values (B1/B2) causes only a very slight reduction of perceived breathiness.

A striking finding was that, in combination, OQ, SQ and BW (stimulus LOW) were surprisingly potent in determining perceived breathiness, even though on their own no one of these parameter had much effect. The very striking reduction in breathy responses to this stimulus, shows that they can offset the effect of high TL and AH values.

The other remaining stimuli, EE, ONOFF and the combination of these two in the EE+ONOFF stimulus appear to contribute little to perceived breathiness.

Increased f_0 did yield a slight reduction in perceived breathiness. However, responses to this stimulus showed a high degree of uncertainty, with some subjects rating the F0+7Hz stimulus as having a more breathy quality. As there was considerable internal variability for many subjects' responses, we conclude the difference is quite audible, but does not on its own contribute seriously to perceived breathiness.

4. DISCUSSION

The main results of this experiment are that TL emerges as the single most important determinant of breathy voice quality. Surprisingly, AH, which might have been expected to be the most salient perceptual cue, would appear to contribute relatively little.

Both of these results run contrary to those reported by Klatt and Klatt [4]. In that study it was found that increasing TL relative to a modal reference stimulus only resulted in a small increase in breathy judgements for some subjects, and was deemed unnatural or nasalised by others. They concluded that "By itself spectral tilt does not appear to be a strong cue to breathiness, perhaps because it does not occur naturally by itself, but rather only in conjunction with certain other cues to breathiness."

The results of these two studies may not be as irreconcilable as would appear at first glance. The Klatt and Klatt experiment was based on manipulations to a reference stimulus with modal parameter settings, whereas the present study involves deviations from a breathy-voiced reference. The present results suggest that although TL is perceptually very potent, its effects may be offset when other parameters (SQ, OQ and B1/B2) are at more modal-like values. Therefore, it is possible that the apparent lack of impact of TL in the Klatt and Klatt study may simply have resulted from its being overridden by the modal settings of these parameters.

Furthermore, a closer look at the numerical results presented in Table XV of Klatt and Klatt [4] suggests that the contribution of TL may have been underestimated in their conclusions and the contribution of AH overestimated. Although it is the case that TL on its own does not much enhance breathiness judgements, the potency of AH is clearly dependent on TL. AH on its own yields high breathiness ratings only when it is at the high value of 60 dB. When TL is increased however, breathiness ratings for considerably lower settings of AH are almost as high.

On their own, these Klatt and Klatt results might be interpreted as showing that AH is the important parameter, but that it can only be heard when spectral tilt is reasonably high, being acoustically masked when TL is low. Our results would point however to a different interpretation: when TL is high (25 dB) and aspiration presumably at its most audible, varying AH as between stimuli AH-12dB, AH0 and the reference stimulus makes rather little difference to perceived breathiness. On the other hand, regardless of the AH value, TL appears to have a major effect. We would reiterate, however, the point made earlier concerning inter-subject differences. It may be that some subjects tune in specifically to this parameter: 4 of our 12 subjects responded to AH variation to a higher degree than the others. Yet even for this subgroup, TL yielded the stronger effect in cueing breathiness.

One final point may also be relevant to the apparently different findings of these two studies where the TL parameter is concerned. When the spectral slope is increased, the loudness of the output is decreased. In our experiment, steps were taken to maintain a constant excitation strength (EE), but the overall loudness was allowed to vary with variation in the other parameters. In Klatt and Klatt, it seems as if loudness variation was compensated for: when TL was high, AV and AH were indirectly increased by increasing the overall gain.

5. CONCLUSIONS

The principal findings were that TL, spectral tilt, appears to be a major determinant of perceived breathy voice. AH, aspiration noise, contributed surprisingly little. On their own, OQ, SQ and B1/B2 have very little effect: however, when all of these are set to modal values they result in a large reduction in perceived breathiness. In this condition, an increase in TL will not in fact suffice to cue breathy voice. Other parameters tested included EE, f_0 differences and the relative abruptness of the vowel onset and offset: these were found to have little effect.

REFERENCES

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

[2] Gobl, C. and Ní Chasaide, A. 1992. Acoustic characteristics of voice quality. *Speech Communication*, 11, 481-490.

[3] Gobl, C. (1989). A preliminary study of acoustic voice quality correlates. *STL-QPSR*, Royal Institute of Technology, Stockholm, 4/1989, 9-21.

[4] Klatt, D.H. and Klatt, L.C. 1990. Analysis, synthesis, and perception of voice quality variations among female and male talkers. *Journal of the Acoustical Society of America*, 87, 820-857.

[5] Mahshie, J. and Gobl, C. 1999. Effects of varying LF parameters on KLSYN88 synthesis. *Proceedings of the XIVth International Congress of Phonetic Sciences*, San Francisco.