Hi-Fi voice: observations on the distribution of energy in the singing voice spectrum above 5 kHz

Sten Ternström

Kungliga Tekniska Högskolan, Dept. of Speech, Music & Hearing, Lindstedtsvägen 24, SE-100 44 Stockholm, Sweden
stern@kth.se

Current audio technology enables the weak spectrum of the voice above 4-5 kHz to be studied reliably. It is known that energy in the 5-20 kHz range can be perceived even when it is 50 dB or more below the main voice spectrum peak. These upper frequencies are conventionally emphasized in broadcasting and production of popular vocal music; yet very few studies of the acoustic content of this range have been made. High fidelity recordings were made of vowels sustained by speakers and singers. A general characterization of the two highest octaves (5-20 kHz) in the spectrum was sought. The prevalence of high-frequency energy and the covariation with overall SPL were highly variable, but several landmark features were identified. In addition to the commonly observed zero at 4-5 kHz, spectral dips were often seen also at 10-12 kHz, so as to form clusters of resonances in the regions 5-10 kHz and 10-20 kHz. Harmonic energy was observed up to 20 kHz in some loud sung tones. It is suggested that octave numbers are useful for referring to these uppermost frequency bands.

1 Introduction

Speech research has conventionally dealt with the voice spectrum only up to about 5 kHz, for good reasons. The major bulk of the signal energy is below 5 kHz; removing high frequencies does not severely impair speech intelligibility, as is exploited in telephony; and the high-frequency acoustics of the vocal tract become awkward when plane propagation of the shorter sound waves can no longer be assumed. Yet, in music production and in broadcasting, speech and song are almost universally emphasized in the upper treble range. ‘Wide-band’ telephony is now being introduced (to 7 kHz rather than 3.5 kHz). From audio engineering, we know that a frequency response to 15 or 20 kHz is considered mandatory for high fidelity. There are also reasons to suppose that clinical voice analysis may benefit from a study of the highest frequencies, for example, in regard to the precision of vocal fold closure, and the relative content of turbulent noise. Finally, it may be argued that subtle *variations* in the high spectrum could contribute to the naturalness of synthetic voices.

We have found very few publications on the voice signal above 5 kHz (e.g., [1, 2]). In audiology, there are a number of studies on high-frequency audiometry (e.g., [3, 4]). In general, they report that even young adults have a hearing threshold raised by a moderate 10-20 dB at 6-12 kHz, and more drastically from 12 kHz and upwards. However, these audiological studies typically report the pure-tone thresholds rather than complex-tone thresholds that would be appropriate for voice sounds. Moore and Tan [5] reported that ten listeners’ ratings of the naturalness of band-limited speech dropped very little when the audio was low-passed at 11 kHz, but drastically with a 7 kHz filter; so the range 7-11 kHz, at least, is audible and important.

Octave	Frequencies	Vocal significance (general)
0	20-40 Hz	- not vocal -
1	40-80 Hz	- not vocal -
2	80-160 Hz	male fundamental F0
3	160-320 Hz	female fundamental F0
4	320-640 Hz	first formant F1
5	640-1250 Hz	F1-F2
6	1.25-2.5 kHz	F2-F3
7	2.5-5 kHz	F3-F5, singer’s formant cluster
8	5-10 kHz	distinct modes; audible to most
9	10-20 kHz	lumped modes; audible to some

Table 1: Octave bands can be appropriate for segmenting the high spectrum of the voice.

Human hearing spans 20-20,000 Hz, or 9.97≈10 octaves. The observations that will be reported here suggest that octaves fortuitously are convenient for describing the high-frequency acoustics of the voice (Table 1). According to the standard for musical octaves [6], octave 0 starts at 16.352 Hz rather than 20 Hz, but otherwise the numbering here is the same. In the speech and voice literature, octaves 2-5 are usually called ‘low’ frequencies, while octaves 6-7 are ‘high’. For frequencies >5 kHz, resorting to octave numbers helps us avoid terms such as ‘very high’ and ‘ultra high’ frequency (in broadcasting: VHF, UHF). Octaves 6-7 correspond roughly to the F1 region that tends to carry the overall sound pressure level. A band limit frequency of 1 kHz, often used with the alpha ratio [7] for spectral balance, is a bit low for most vowels and for female voices; so 1250 Hz is actually better; although it could be argued that 1500 Hz would be better still. Octaves 6-7 correspond to the F3-F5 region, including the singer’s formant cluster. The content of octaves 8 and 9 are the topic of this report.

2 Method

2.1 General considerations

For analysing and explaining the high spectrum of the voice signal, a decomposition would be desirable into source and filter, and into the periodic versus the turbulent sources. Although it is invasive and difficult to do, a very small microphone or calibrated sound source might be introduced into the airway near the glottis, to estimate the vocal tract transfer function [8]. However, due to the close spacing of resonance nodes at short wavelengths, the estimated transfer function will be very sensitive to the exact position of the transducer [9]. The method of transcutaneously excited sine sweeps, successfully used up to 5 kHz by Fujimura and Lindqvist [10], would incur similar difficulties.

From room acoustics we know that the density of resonance modes is low at the lowest frequencies, but increases very rapidly with frequency f. If we crudely approximate the vocal tract with a very small rectangular room, of volume *V*, enclosing area *S* and total edge length *L*, the number of resonances per Hz n at frequency f can be estimated [11] as

$$n = \frac{4\pi V}{c^3} f^2 + \frac{\pi S}{2c^2} f + \frac{L}{8c} \quad (1)$$

Inserting reasonable values for *V*, *S* and *L*, and combining this with the formula

$$ERB = 6.23 f_{kHz}^{2} + 93.39 f_{kHz} + 28.52 \quad (2)$$

for the equivalent rectangular bandwidth *ERB* of critical bands [12], we find that for vocal tract volumes in the range 50-100 cm^3, there will be more than one resonance per critical band above 4-5 kHz. It remains to estimate the Schroeder frequency of the vocal tract, above which individual resonances no longer can be resolved even instrumentally. For this calculation, we need to know their typical resonance bandwidths.

So, at high frequencies, the spectrum level is highly variable, the acoustic energy is miniscule, and the auditory critical bands are wider than the resonance clusters. Therefore, a very detailed analysis of various static vowel spectra is not likely to be meaningful. Rather, it is only the major features, resolvable by our critical bands, that attract our attention in this first study. It was decided to record only radiated sound, and to use only the voice itself as the sound source.

2.2 Acquisition

Recordings were made in anechoic conditions, in two locations. Normally, omnidirectional microphones are preferable, but practical considerations led us to use cardioid condenser types (Neumann model KM140; Line Audio model CM3). A DPA 4066C miniature omnidirectional condenser was recorded on a parallel channel. The microphones were placed 30 cm in front of the mouth, and adjusted to each subject's height. An external sound card (RME Fireface 400; MOTU Traveler; both with built-in low-noise preamps) was connected to a laptop computer. The sampling rate was 44100 Hz throughout, with 16-bit resolution. On comparing the cardioid and omni signals, the proximity effect of the cardioids was found to be negligible, while the omni with its smaller diaphragm was a little noisier. Therefore, only the cardioid signals were subjected to analysis. The long-time average spectrum (LTAS) of the background noise was at least 20 dB below the LTAS at all frequencies for all voiced sounds. For fry and whisper, the signal at high frequencies would sometimes drop below the noise floor.

2.3 Vocal tasks and subjects

The recording protocol for each subject was as follows:

1. Calibration for SPL using a sound level meter and a sustained vowel.
2. Read a prose text [13], as if reading aloud to a group, for at least 60 seconds.
3. For the five vowels {u: O: a: E i: } , repeat, while attempting strictly to maintain the vowel articulation as constant as possible throughout:
   (a) sustain the vowel for at least five seconds at a comfortable phonation frequency and effort level,
   (b) perform ingressive fry phonation at as low a pulse rate as possible,
   (c) sustain a whisper for at least five seconds,
   (d) sing a free glissando, about an octave up and down,
   (e) sing an arpeggio on the major scale notes 1-3-5-8-10-12-11-9-7-5-4-2-1, where 1 is freely chosen.
   (f) expert singers only: sing a crescendo-decrescendo while sustaining the F0 and the vowel.

Tasks 3a-d were intended to give samples of (a) pulse train excitation, (b) single-pulse excitation, (c) noise excitation, (d) frequency sweep excitation, (e) variability over a large F0 range, and (f) spectrum slope vs. vocal effort. Prior to recording, subjects were rehearsed in the tasks.

The subjects were four males, S1-S4, and four females, S5-S8, aged 25-51, with singing experience ranging from experienced choir singer to national-level professional teacher. Subject S1 was not a singer, but was included for his rich speaking voice and for his ability to perform separated pulses in ingressive fry phonation.

2.4 Analysis

The signal files were analyzed using the Swell Soundfile Editor and its companion tools for making line spectra, LTAS and spectrograms (Soundswell Core 4.00, Hitech Development AB, Täby, Sweden). For the line spectra and LTAS, 2048-point FFT:s were used, with a 45 ms Hanning window, giving an frequency resolution of 44 Hz. Using the *Extract* tool (bandpass filters and thresholds with hysteresis), the passage of running speech from task 2 was split into voiced and unvoiced parts, and LTAS were made of the voiced part. For tasks 3a-3d, spectrograms to 20 kHz were first made of the five-second productions. The spectrograms invariably exhibited gradual fluctuations in octaves 8 and 9, which would be due to small shape changes of the subject's vocal tract (

Figure **2**). Because the LTAS tends to emphasize the stronger parts of a signal, any frequency shifting of a spectral dip will conceal it in the LTAS. Therefore, for each token, the most stable portion, usually of one or two seconds duration, was selected manually, and the LTAS of this portion only was computed. For task 3b, selected individual glottal pulse responses were edited out and analyzed with no windowing but with an FFT length matching the pulse length.

All spectral data were copy-pasted into Microsoft Excel, where they were displayed and grouped into octave-based frequency bands for data reduction, as needed.

3 Qualitative results

A dip in the vowel spectrum at 4-5 kHz is commonly seen; it is caused by a pair of antiresonances due to the cavities of the piriform fossa [14], [15]. This dip, henceforth called the PF notch, has historically been taken as an upper bound to the speech spectrum. Here, it makes a convenient landmark for the transition into octaves 8 and 9.

Looking first at the LTAS of the running speech task (Figure 1), it was noted that, when all the unvoiced segments had been removed, the spectrum level in octaves 9-10 became 10-20 dB lower. Also, local minima appeared, more or less clearly in all subjects' LTAS at about 5-6, 9-10 and 12-14 kHz, for males and females alike. The one at 5-6 kHz would be the remnant of the antiresonances at 4-5 kHz under the LTAS operation. These minima notwithstanding, the LTAS contour of the running speech was quite personal even in octaves 8-9. Taking the LTAS of the first or last 30 seconds of the read speech would give very similar results within subjects; while the contour was generally more different from one subject to the next.

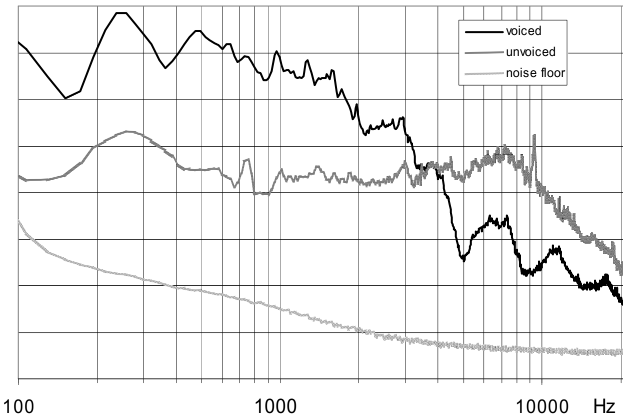

Figure 1. Example LTAS of 60 s running speech, subject S5. Vertical scale is 10 dB/div. (The spurious unvoiced peak at 9.5 kHz is due to a single whistling ‘s’ sound.)

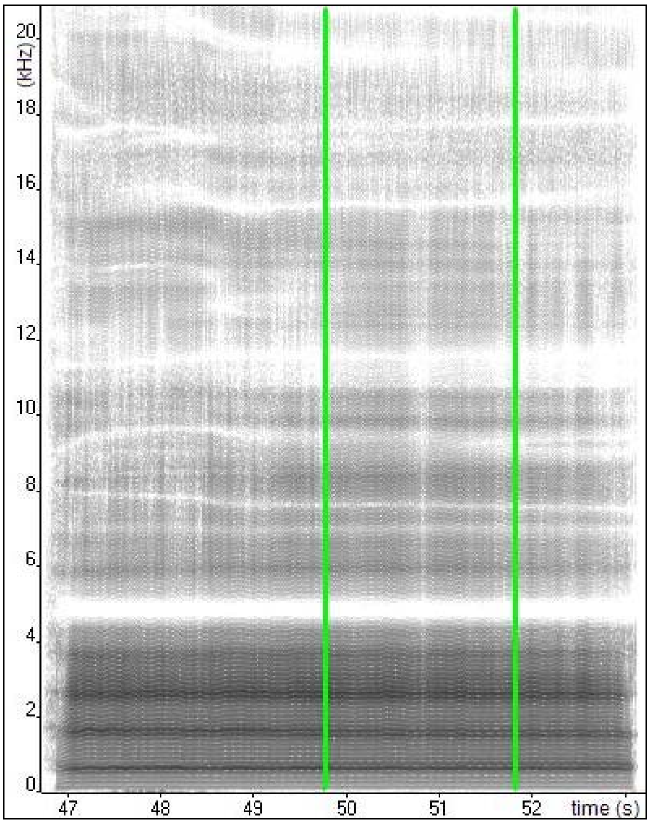

Figure 2. A spectrogram to 20 kHz: task 3a, vowel /ɛ:/ sustained for six seconds, subject 1, male. Note the fluctuations above 8 kHz; the unusually deep but typically located antiresonance notch at 4.9 kHz; and a less prominent trough at 11.5 kHz. Solid vertical lines enclose a relatively static portion, where the average spectrum can be taken.

In a spectrogram of a sustained vowel (**Figure 2**), it can be seen that the approximate spacing between formants is the familiar 1 kHz, up to the PF notch at 5 kHz or so. Then in octave 8, resonances come closer together but are still discernible, and antiresonances appear. In octave 9, the spacing between resonances becomes smaller still, and they start to smear into clusters. This is quite analogous to the behaviour of resonance modes that is known from room acoustics. The level at 6 kHz and higher is about 50 dB below the main spectrum peak at 640 Hz. The higher spectrum is seen to fluctuate slowly because of inevitable small movements of the vocal tract. Such movements would be indirectly due to changes in lung volume, subglottal pressure, etc. The perceived timbre was very stable, but not mechanically so.

A spectrum of the same vowel, pronounced immediately afterwards in ingressive fry by the same subject, is shown in Figure 3. One advantage of ingressive fry is that it is possible, with practice, to produce pulses at very low rates, even in isolation. The output spectrum of the vocal tract, if it were excited by a single unit impulse, would be that of the tract’s transfer function. Here, the exact shape of the excitatory ingressive pulse is not known; although it can safely be assumed (a) not to be a unit impulse in the mathematical sense; (b) not to contain significant periodic components. Therefore, it is essentially the overall slope of the obtained vocal tract transfer function that will be incorrect. However, the resonances and antiresonances show up very clearly. Since this method gives the vocal tract response to one impulse only, the contour smearing that is observed toward high frequencies is due not to variations in fundamental frequency, as in the LTAS, but to the increasing density of the resonance modes.

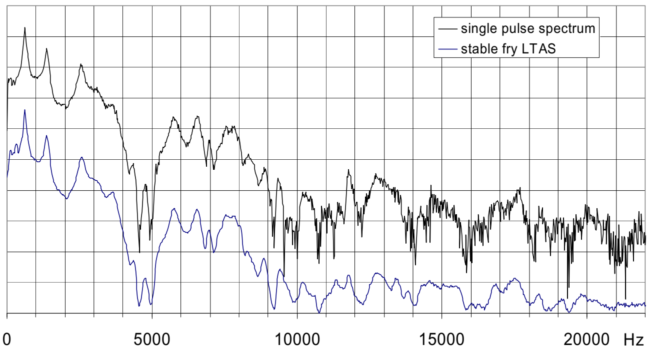

Figure 3. Spectrum of a single ingressive pulse (upper curve) and LTAS of 2 s of ingressive fry phonation (lower). Subject S1, male, vowel /ɛ:/. Vertical scale is 10 dB/div.

The formants F1-F3 are neatly resolved in Figure 3, and there are hints of F4-F6 in the slope down to the PF notch. This twin antiresonance just below 5 kHz is particularly clear here. It is followed by a characteristic cluster of resonances from 5-10 kHz which was seen in many tokens.

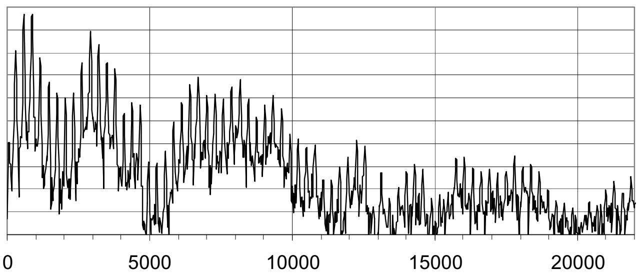

Figure 4. Spectrum of sung vowel /a:/, subject S2, male, at a fairly high fundamental of 291 Hz. Note that harmonics are visible up to 20 kHz. Vertical scale is 10 dB/div.

On loud sung notes, harmonic energy was in a few cases visible all the way up to 20 kHz. The example shown in Figure 4 had a perceptual ring that was attributed in part to the moderate singer’s formant cluster at 3 kHz, but especially to the cluster of resonances at 5-10 kHz. The harmonics above 11 kHz are probably inaudible.

4 Quantitative results

The energy in the highest octave bands, relative to the total SPL was measured for task 2 and task 3a, see Figure 5. Subjects performed task 3 at diverse SPL's ranging from 75-91 dB, which accounts for the large spread in the highest octaves. In octaves 2-5, the level was only slightly lower than the total SPL of the sound, as follows from the fact that this frequency band dominates the signal. The level in octaves 6-7 varied greatly with the frequency of especially the second formant, also as expected. In octaves 8 and 9 the relative level was typically -30 to -45 dB, with differences between vowels being somewhat smaller than for the band of octaves 6-7. The levels in octave 8 and 9 appeared to vary together, with octave 9 being on average 5-8 dB weaker than octave 8.

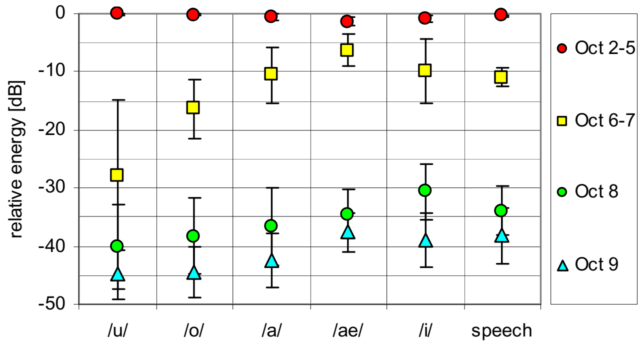

Figure 5. Energy of octave bands for five sustained vowels, and the voiced segments only of running speech, relative to the total SPL. Each point is a mean of the levels for eight subjects. Vertical bars give the standard deviation. Octave numbers are those defined in Table 1.

Task 3f was performed by only three subjects: 4 (male), 7 and 8 (females), who were the most highly trained singers. The covariation of the levels in the high spectrum with SPL was assessed as follows. The signal from the crescendo-decrescendo task (also known as *messa di voce*), of 5-10 s duration, was band-pass filtered into four channels corresponding to octaves 2-5, 6-7, 8 and 9. The levels in each of these bands were then plotted against the total SPL. Two examples are shown in Figure 6.

The slopes of the lines (dB in-band per dB SPL) in these plots were computed by linear regression. The results for five vowels are shown in Figure 7. In general, the slope was around 1.5 in octave band 6-7, which concurs with the literature; much the same in octave 8, and smaller, but usually greater than one, in octave 9. This means that the spectrum slope, on the whole, changed with SPL only up to 5 kHz. The level difference between octaves 6-7 and octave 8 remained much the same with changing SPL; while the level in octave 9 changed little more than the SPL itself, in most cases.

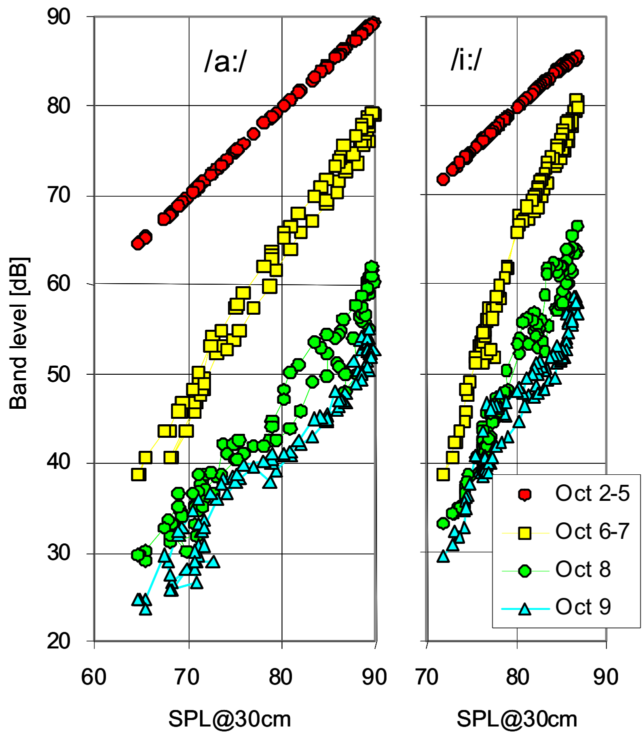

Figure 6. Example of high-band spectrum level variation with SPL. Subject 4, male, task 3f, vowels /a:/ and /i:/.

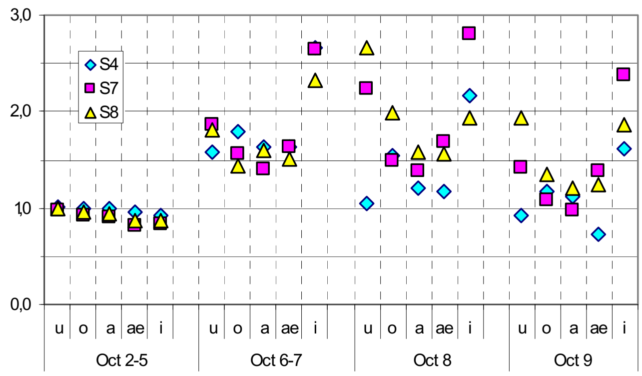

Figure 7. Slope values (vertical axis) for the linear regressions of in-band levels versus SPL, for subjects S4, S7 and S8.

5 Discussion and conclusion

Auditory masking: it can be seen in Figure 5 that the levels in octave 6-7 for sustained vowel sounds were 20-30 dB higher than that in octave 8. Models of auditory masking [16] indicate a slope of the masker toward higher frequencies of about -30 dB per octave, and the energy in octave band 6-7 will usually be greater in octave 6. Hence it may be expected that for most of these vowel sounds, octaves 8-9 will not be masked, and on average octave 8 will not mask octave 9. The audibility of the high bands will still depend on the individual's threshold of hearing. Future work will include listening tests with filtered recordings as stimuli.

With this rather limited selection of subjects and tasks, characteristic features in the high spectrum were found to be the PF notch, a cluster of resonances at 5-10 kHz (octave 9), and a smaller trough at 10-13 kHz. In the LTAS of reading there was also usually a small trough at about 9 kHz; however, this may be an artefact of the LTAS operation. In the non-reading tasks there was a weak trend toward a lower and broader cluster spanning all or part of octave 9. Two high clusters have been observed in male singers by Titze and Jin [2]. They suggested the interpretation that the clusters were occurring at odd multiples of the singer's formant cluster, given that the vocal tract acts like a quarter-wave pipe. In the present study, there were no operatic male singers, and no strong singer's formant cluster was manifest in any subject. Subject S4 is a rock singer who teaches at the conservatory level. Another interpretation is that the dip in the 10-12 kHz region could be a wavelength multiple of the PF notch. The data of Dang and Honda [15][14] extend only to 10 kHz, so a further study would be needed to test this. Sundberg [14] showed with an acoustic model that the frequency of the PF notch depends on the size of the sinus piriformes.

The high end of the voice spectrum is routinely amplified in music production and broadcasting. This is said to create a more 'open' and/or 'crisp' sound. A case in point is that most cardioid microphones for voice are designed with a slight treble boost around 10 kHz. For the microphones used here, the boost was very small: about +2 dB around 9 kHz for the Neumann KM140, and +1 dB at 10-15 kHz for the Line Audio CM3. Still, such deviations should be compensated for in a more precise quantitative analysis.

Our sense of hearing abhors constancy and dotes on variation. Hence the nature of the *variations* in the high spectrum envelope could be particularly interesting. This is an aspect which is rarely modelled, yet which might contribute to the naturalness of synthesized speech. In formant synthesis, for example, the high spectrum is often absent, or represented by static higher-pole compensation filters, or even faked using intentional digital aliasing. In signal compression, synthetic bandwidth expansion has been implemented, by which the high spectrum is guessed from the low spectrum, and this is now a standardized method. It would be interesting to ascertain whether a simple model of a suitably variable, if imprecise, high spectrum would improve naturalness.

It is an interesting coincidence that the auditory critical bands do not resolve individual resonances above that same frequency, 5 kHz, where the plane wave approximation for the vocal tract no longer holds.

Acknowledgments

The author is grateful to the eight subjects who participated enthusiastically and without compensation. Thanks are due to David Howard and Damian Murphy for their assistance during recordings at the University of York. The author's trip to York was funded by an EPSRC (UK) grant held by Dr. Murphy. Thanks also to the department of Linguistics at Stockholm University for access to their anechoic facility and technical assistance. Johan Sundberg gave valuable comments on the manuscript. This work is supported by the Swedish Research Council, contract 2007-4460.

References

[1] K. Shoji, E. Regenbogen, J. Daw Yu, S.M. Blaugrund. High-frequency components of normal voice. *J. Voice*, **5** (1), 29-35 (1991).

[2] I.R. Titze, Sung Min Jin. Is there evidence of a second singer's formant? *J. Singing* **59** (4), 329-331 (March/April 2003).

[3] D. Osterhammel, P. Osterhammel. High frequency audiometry – age and sex variations. *Scand. Audiol.* **8** 73-81 (1979).

[4] M.A. Schechter, A. Fausti, Z. Rappaport, H. Frey. Age categorization of high-frequency auditory threshold data. *J. Acoust. Soc. Am.* **79**(3), 767-771.

[5] B. J. C. Moore, C.-T. Tan: Perceived naturalness of spectrally distorted speech and music. *J. Acoust. Soc. Am.* **114** (1), 408-419 (2003).

[6] ANSI Standard S1.1-1994, item 13.18.

[7] P. Kitzing. LTAS criteria pertinent to the measurement of voice quality. *J. Phonetics* **14**, 477–482 (1986).

[8] Kob M (2002). *Physical Modeling of the Singing Voice*. Doctoral dissertation RTWH Aachen, Logos Verlag, Berlin. ISBN 3-89722-997-8.

[9] K. Motoki. Three-dimensional acoustic field in vocal-tract. (Tutorial). *Acoust. Sci. & Tech.*, 23 (4), 207-212 (2002). G. Greene. *Travels With My Aunt*, chapter 1.

[10] O. Fujimura, J. Lindqvist. Sweep-tone measurements of vocal-tract characteristics. *J. Acoust. Soc. Am.* **49** (2), 541-558 (1970).

[11] J. Liljencrants, S. Granqvist. Kompendium i Elektro-akustik. KTH TMH 2004.

[12] B. J. C. Moore, Glasberg, R. Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. *J. Acoust. Soc. Am.* **70**, 1003-1014.

[13] G. Greene. *Travels With My Aunt*, chapter 1.

[14] J. Sundberg. Articulatory interpretation of the "singing formant." *J. Acoust. Soc. Am.*, **55** (4), 838-844 (1974).

[15] J. Dang, K. Honda. Acoustic characteristics of the piriform fossa in models and humans. *J. Acoust. Soc. Am.* **101** (1), 456-465.

[16] E. Zwicker, H. Fastl. *Psychoacoustics – facts and models*. 2nd edition, p. 168, Springer Verlag, Berlin Heidelberg, 1999.