Dept. for Speech, Music and Hearing
Quarterly Progress and Status Report

Perception of vowels with truncated intraperiod decay envelopes

Fant, G. and Liljencrants, J.

journal: STL-QPSR
volume: 20
number: 1
year: 1979
pages: 079-084

KTH Computer Science and Communication

http://www.speech.kth.se/qpsr

II. SPEECH PERCEPTION

A. PERCEPTION OF VOWELS WITH TRUNCATED INTRAPERIOD DECAY ENVELOPES

G. Fant and J. Liljencrants

Abstract

The glottal impedance is a highly time-variable component of the damping of vocal tract resonance modes which significantly affects the decay profile of formant oscillations. In extreme cases the rapid decay of energy in the glottal open part of the voice cycle accounts for a truncated decay contour. Experiments where subjects match such stimuli to vowels generated with conventional exponential decay envelopes of variable rate have been performed. By this technique it is possible to derive equivalent bandwidths. An interpretation of the results is given in terms of alternative models of auditory signal processing. It is found that a criterion of short time average integration of signal amplitude preserving a constant peak to mean value of the signal within a voice fundamental period provides a reasonable fit to the experimental data. A model with 3 dB down spectrum bandwidth provides too large bandwidths. On the other extreme a criterion of constant peak to rms ratio provides a too low estimate.

Non-uniform damping

The phenomenon of non-uniform formant damping is extensively treated in another article in this issue of STL-QPSR. It is illustrated in Fig. II-A-1 which shows the main excitation of the first resonance mode F1 at the instant of glottal closure. It is subjected to constant exponential decay until the glottis opens when the damping increases heavily. The top curve in Fig. II-A-1 pertains to a glottal maximal opening of 0.16 cm^2, a subglottal pressure of 6 cm H_20, a glottal slit of 3 mm depth, and a vocal tract configuration with a constricted pharynx as for the vowel [ɑ] . The glottal bandwidth is furthermore inversely proportional to particle velocity of the air flow or to $(P_S)^{-1/2}$, where P_S is the subglottal pressure.

An important point is that the bandwidth increases with the vocal tract supraglottal impedance level at the frequency of the resonance mode. This is not a dependency of the resistive component of the supraglottal impedance. It is rather a matter of the characteristic impedance $\varrho c/A$ of the back part of the tract, which indicates that a vowel produced with narrow pharynx is potentially more susceptable to glottal damping than vowels produced with a wide pharynx. In gen-

eral, the glottal damping increases with the level of the potential energy of the mode in the vi inity of the glottis. The fourth mode of a male resonator system is known to be highly dependent of the larynx tube which means that the F4-decay should display a typical truncation effect. The most simple model, however, is that of a Helmholtz resonator with volume V and capacitance $C = V/\varrho c^2$. Assuming a parallel resistance R_g the bandwidth contribution from R_g is

$$B_g = \frac{1}{2\pi R_g C} = \frac{\varrho c^2}{2\pi R_g \cdot V}$$

With $V = 62.5\ cm^3$ and $R_g = 0.5\varrho c$, $B_g = 180$ Hz which would be typical of the vowel [œ] or [e]. A reduction of the pharynx volume by a factor 3.8 would provide 3.8 times larger damping or $B_g = 690$ Hz, as in the top curve of Fig. II-A-1. A closer analysis of conditions appropriate for a vowel [ɑ] shows that an additional volume reduction of a factor 2 is needed for the same effect since F1 and F2 of [ɑ] have approximately the same dependency of the back and front part of the resonator system. From Fant (1960) we note that V_2 of the vowel [ɑ] is $8.4\ cm^3$ and of the vowel [o] $13\ cm^2$. The increase in F_1-frequency during the period of glottal opening caused by the glottal inductance is probably less significant because it is confined to a time interval where the F1-energy anyhow is largely reduced.

Perceptual experiments

The first experiment was to synthesize a neutral vowel of $F_1 = 500$ Hz, $F_2 = 1500$ Hz, $F_3 = 2500$ Hz etc. by means of a time domain approach. A wave train of five 1-ms long rectangular pulses with 1 ms spacing between pulses was modulated by an exponential decay of variable time constant $\tau = \pi B$ and periodically repeated at intervals from 11.2 to 6.4 ms providing an F_0-contour from about 90 to 155 Hz. This was the reference. The test stimuli were produced in the same way except for B = 0 and a truncation leaving four, three, two, or only one rectangular pulses within a full fundamental period. The subject manipulated a joy stick for selecting the best combination of the decay constant of the reference and its amplitude level to match the test item.

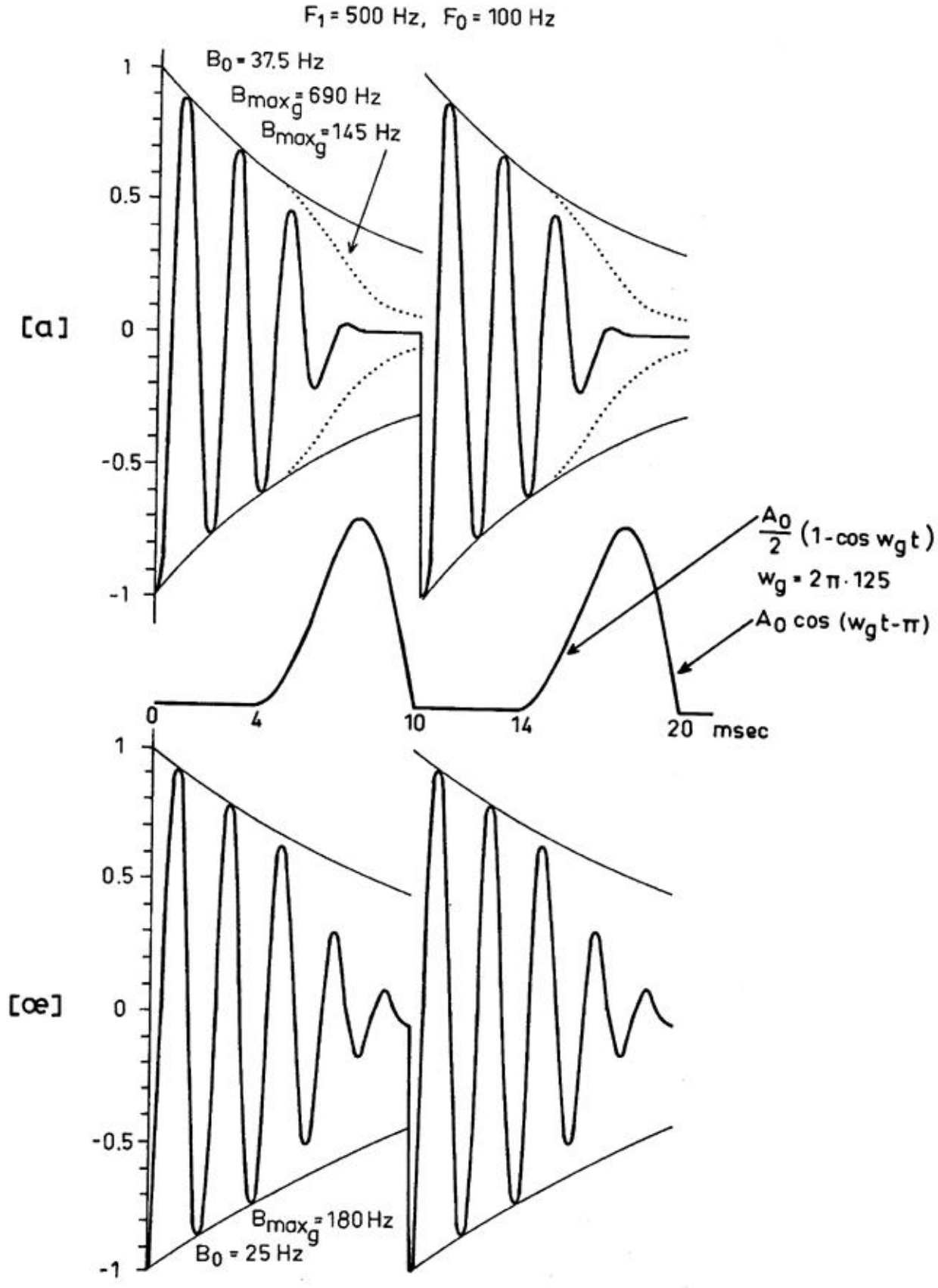

Fig. II-A-1. Calculated decay characteristics of F1-oscillations evoked at the instant of glottis closure. High and medium degree of glottal damping.

The data for one-pulse vowels were discarded because of the difficulty of the task and a limited range of varying the reference bandwidth. An upper limit of the reproduced audio range was set by a low-pass filter to 5.6 kHz and to 2.8 kHz in a second part of the experiment. The results are shown in the following tabulations.

Table II-A-Ia. Subjects' bandwidth selection.

Number of pulses in stimuli	RC	AL	GF	LN		JL	
	LP 5.6	LP 5.6	LP 2.8	LP 5.6	LP 2.8	LP 5.6	LP 2.8
2	53	50	84	92	99	76	88
3	55	32	53	34	26	41	47
4	56	42	45	15	-1	36	41

Table II-A-Ib. Average data and linear model data

Number of pulses in stimuli	LP 5.6	LP 2.8	Linear model
	Bandwidth setting		
2	68	90	80
3	40	42	40
4	37	28	15

The linear model predicts a bandwidth which provides the same peak to mean value as the test signal. The fit is good except for the condition with four pulses. This may be explained by the fact that the fourth pulse is not fully reproduced at the highest F_0.

In a second test the fundamental frequency was held constant at 100 Hz. One of the test items, nr 3A in Fig. II-A-2, had the same properties as described in the previous test, i.e. constant amplitude of the three pulses, whereas a constant decay set by a 50-Hz bandwidth was introduced in the sequence of pulses of other test items. The test signal 3C1 is identical to 3C except for a low-pass filtering at 1000 Hz which eliminates F_2 and higher formants.

Some of the subjects reported difficulties in performing the task and it appears that competing auditory criteria were involved. The spread was particularly large in the 3A case, i.e. with bandwidth B=0 during the non-truncated part of the voice cycle. The spread is not uniform but highly individual as may be seen from the following tabulation.

Table II-A-II. Subjects' bandwidth selection. Means of four matchings and standard deviation are tabulated.

Subject	2C mean	2C s.d.	3A mean	3A s.d.	3C mean	3C s.d.	3C1 mean	3C1 s.d.	4C mean	4C s.d.
JS	208	24	144	68	95	43	82	13	57	9
MB	165	9	27	7	84	10	92	11	61	4
LN	180	23	51	5	93	9	99	29	64	8
IK	113	25	62	27	75	17	97	13	80	23
BG	151	13	108	9	84	10	94	14	61	5
HT	180	21	163	31	98	15	111	24	77	9
	166	19	93	25	88	16	95	17	67	10

Discussion

The distribution of the 3A data is probably bimodal. The lower values represented by subjects MB, LN, and IK are of the order of 45 Hz as in the first experiment, whilst the remaining subjects matched around 140 Hz. The monotone pitch F_0 apparently made the matching task more difficult than in the first experiment. It is probable that most of the subjects used a criterion of equal balance between the "ringing" quality of formants and the partially masking background of the overall spectrum. Another difficulty in the matching is probably related to the discontinuity between the non-truncated and truncated parts of the fundamental period. The corresponding distortion in terms of sinx/x sidebands produces an overall "graininess". Further experiments should be based on more continuous time envelope modulation and further tests should be made with variable fundamental frequency.

Four different signal processing models have been tested against the experimental data. The best match is as before a criterion that

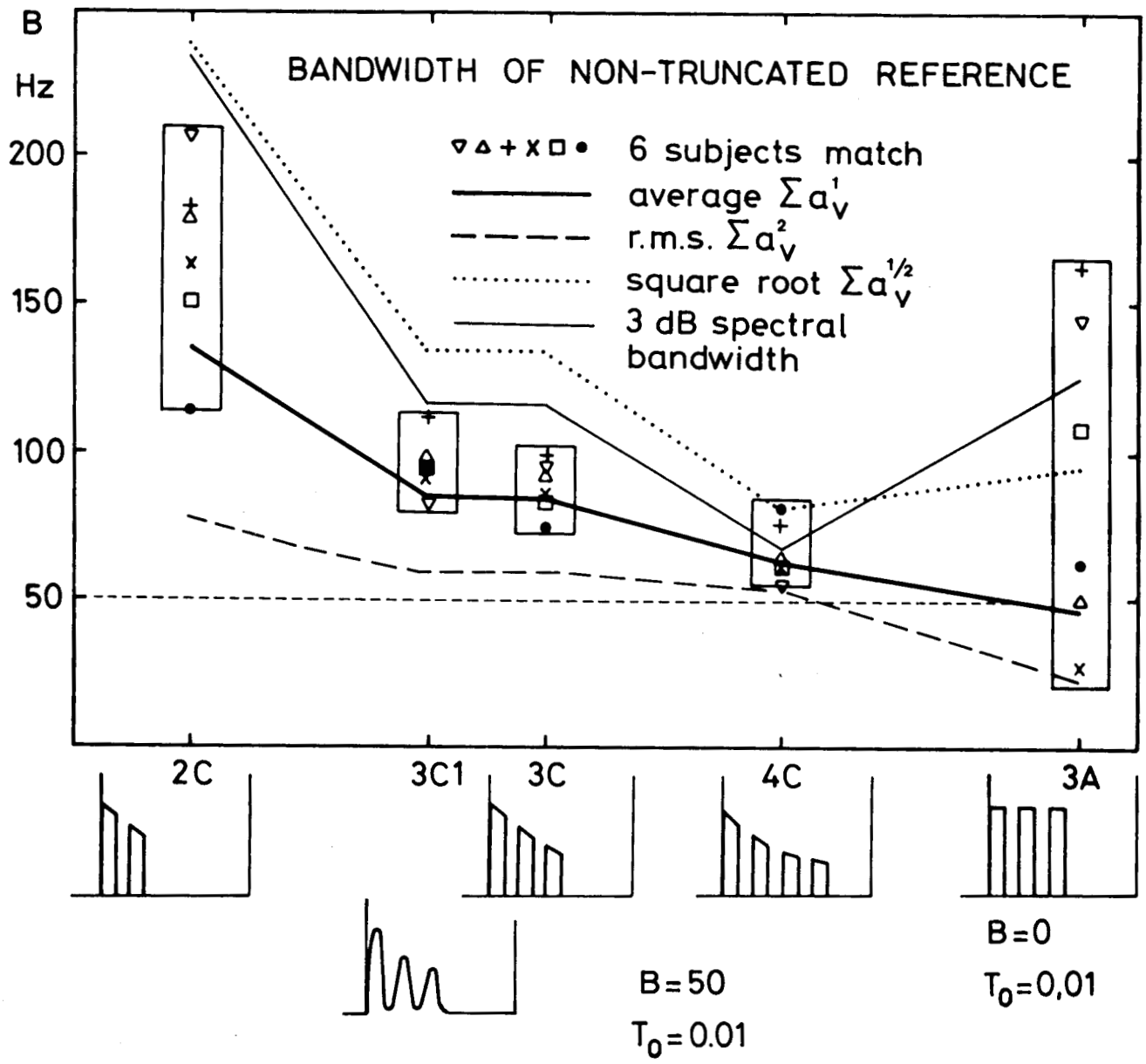

Fig. II-A-2. Equivalent bandwidth of exponentially damped formant oscillations matching truncated decay profiles.

the uniform exponentially decaying reference should have the same peak to rectified mean value as that of the test signal. A criterion of equal peak to rms value provides lower values than any of the matched data points. A 3-dB down spectral criterion, on the other hand, provides an upper bound.

One could speculate in an auditory processing involving short time signal convolution within critical bands with or without additional temporal integration of activity in subsequent stages. The time constant of the critical band filter sets a limit to the resolution of temporal fine structure. Rapidly varying temporal envelopes are thus smeared out according to some law of temporal integration of neutral activity first at the cochlear level and then at higher levels. A fourth root power law of signal intensity or square root of signal amplitude has been suggested by Schroeder, see Fourcin et al (1977). This model provides values that in general are somewhat greater than the spectral 3-dB down values of equivalent bandwidth and of the order of twice our experimentally determined values. The simple mean value integration of signal amplitudes provides a much better match but are consistently somewhat lower than the experimental data. The ambiguity element in our subjects' performance might be related to a judgement of on the one hand roughness which is related to modulation profiles within critical bands and on the other hand judgements of formant loudness where longer integration times occur, 100 ms versus the order of magnitude of 10 ms for critical bands in the first formant region. The small difference observed between the test stimuli 3C1 and 3C suggests a dominance of F1 in the percept.

Finally we may comment on the magnitude of the equivalent glottal damping of speech formants. The full scale from 2C to 4C is typical for resonance modes with high sensitivity to glottal damping. In this range corresponding to relative durations of glottal closure of 40% to 80%, the effective glottal bandwidth varies from 120 Hz down to the order of 20 Hz. The intermediate value of 60% effective closure time and 40 Hz bandwidth increase in 3C should be typical of F1 of the vowel [ɑ] produced by a male voice.

If the glottis never reaches a proper closure the effect is maximal. This would be the case in high pitched falsetto voices. The

glottal bandwidth also increases with any leakage during the closed phase. The typical range of glottal bandwidths 20-40 Hz derived by Flanagan et al (1975) from a simple mean value within the voice period is representative of our findings. It should be observed, however, that the true values may be more dependent on the relative opening and closing times than on the particular scale value of the glottal resistance and that the glottal inductance minimizes the effect on higher formants.

Acknowledgments

Inger Karlsson and Lennart Nord participated in the data processing. Thanks are also due to the subjects engaged in the test. This work was in part supported by a grant from the Bank of Sweden Tercentenary Foundation.

References

FANT, G. (1960): Acoustic Theory of Speech Production, Mouton, 's-Gravenhage (2nd edition 1970).

FOURCIN, A.J. & al. (1977): "Speech Processing by Man and Machine. Group Report", p. 325 in Recognition of Complex Acoustic Signals, (T.H. Bullock, ed.), Dahlem Konferenzen, Berlin.

FLANAGAN, J.L. ISHIZAKA, K. & SHIPLEY, K. (1975): "Synthesis of speech from a dynamic model of the vocal cords and vocal tract", Bell System Techn. J. 54, March, pp. 485-506.

WAKITA, H. & FANT, G. (1978): "Toward a better vocal tract model", STL-QPSR 1/1978, pp. 9-29.