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

How to define formant level. A study of the mathematical model of voiced sounds

Fant, G. and Liljencrants, J.

journal: STL-QPSR
volume: 3
number: 2
year: 1962
pages: 001-009

KTH Computer Science and Communication

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

I. SPEECH ANALYSIS

A. HOW TO DEFINE FORMANT LEVEL. A STUDY OF THE MATHEMATICAL MODEL OF VOICED SOUNDS

The problem

The following theoretical study is an extension of the investigation reported on by Fant 1959 [1], p. 46-60. It is intended as a supplement to the experimental study of Fintoft, Lindblom, and Mártony reviewed in the present quarterly report. The particular problems to be dealt with are the following:

1. What are the relations between the various frequency domain and time domain quantities which may be adopted as measures of formant level or in other words formant amplitude?
2. What are the effects of a variation in voice fundamental frequency $F_0=1/T_0$ on the various formant amplitude parameters assuming a constant waveshape of each vocal pulse?

The standard model

The simplified transform of an ideal non-nasal voiced sound is generally expressed in the following form, Eq. 2.5-4 of ref. [1],

$$P(1,p)=S(p)\cdot H(p)\cdot R(p) = \frac{\varrho\cdot p\cdot K_T(p)\cdot U_0\cdot K_{rr}(p)\cdot \prod_1^{\nu} p_k\cdot \prod_1^{r} \hat{p}_n\cdot \hat{p}_n^*}{4\pi l[1-e^{-pT_0}]\prod_1^{\nu}(p-p_k)\prod_1^{r}(p-\hat{p}_n)(p-\hat{p}_n^*)} \qquad (1)$$

which conforms with our standard synthesis procedure. It is derived from the three main factors; source, vocal transmission, and radiation

$$R(p) = \frac{\varrho p}{4\pi l}\cdot K_T(p) \qquad (2)$$

is the transfer from volume velocity at the lips to sound pressure $P(1,p)$ at the distance of 1 cm from the speaker and $K_T(p)$ accounts for the baffle effect of the head and the deviation of radiation resistance from the simple square law dependency of frequency. This correction factor is generally neglected. The driving source function enters with the transform

$$S(p) = \frac{U_0 \prod_1^{\nu} p_k}{\prod_1^{\nu}(p-\hat{p}_k)} \cdot \frac{1}{[1-e^{pT_0}]} \tag{3}$$

The periodicity of the source function enters through the factor $[1-e^{pT_0}]^{-1}$ where T_0 is the time interval between two successive vocal pulses. The voice fundamental frequency is $F_0 = 1/T_0$. In more careful matchings of individual voice spectra we generally include a complex conjugate zero around 1000 c/s in the source function.

The vocal transfer function

$$H(p) = \frac{K_{rr}(p) \cdot \prod_1^{r} \hat{p}_n \hat{p}_n^*}{\prod_1^{r}(p-\hat{p}_n)(p-\hat{p}_n^*)} \tag{4}$$

comprises a function of r conjugate complex poles only and a correction for poles above pole number r. In nasalized vowels there enters an additional factor of conjugate complex zeros.

The inverse transform of P(1,p) is the speech wave time function which we will label w(t). When stationary conditions are reached a single fundamental period may be expressed as

$$w(t) = \sum_1^{\nu} A'_k e^{p_k t} + [-1]^n \sum^{n} A'_n \cdot e^{\sigma_n t} \cos(\omega_{nt} + \varphi'_n) \tag{5}$$

which may be considered as the superimposed sum of the contributions from damped exponentials and sinusoids excited once during each fundamental period starting with period $g = -\infty$ and ending with the period $g = 0$ under observation.

$$w(t) = \sum_{-\infty}^{0} \left\{ \sum_1^{\nu} A_k e^{p_k(t-gT_0)} + [-1]^n \sum_1^{r} A_n e^{\sigma_n(t-gT_0)} \cdot \right.$$

$$\left. \cdot \cos[\omega_n(t-gT_0) + \varphi_n] \right\} \tag{6}$$

The superposition effect

In the following study we will be concerned with the damped oscillations only and we note that A_n is the initial amplitude due to the excitation from a single voice pulse and that A'_n is the initial amplitude with due regard to the superposition effects. The frequency domain correspondence of A_n is the transform of a single formant, i.e. the pole function

$$H_n(p) = \frac{\hat{p}_n \hat{p}_n^*}{(p-\hat{p}_n)(p-\hat{p}_n^*)} = \frac{\omega_{on}^2}{(p+\sigma_n)^2 + \omega_n^2} \tag{7}$$

of frequency $F_n = \omega_n/2\pi$, bandwidth $B_n = -\sigma_n/\pi$, and $\omega_{on}^2 = \sigma_n^2 + \omega_n^2$. In the following treatment we will assume high $Q = -\omega_n/2\sigma_n$.

The ratio of A'_n/A_n will be the same as the ratio $h'_n(t)/h_n(t)$ where

$$h_n(t) = L^{-1} \cdot \left\{ H_n(p) \right\} = \omega_n \cdot e^{\sigma_n t} \sin\omega_n t \tag{8}$$

$$h'_n(t) = L^{-1} \left\{ H_n(p)/(1-e^{pT_0}) \right\} = (1+e^{+2\sigma_n T_0} - 2 \cdot e^{\sigma_n T_0} \cos\omega_n T_0)^{-\frac{1}{2}} \cdot$$

$$\cdot \omega_n e^{\sigma_n t} (\sin\omega_n t + \varphi'_n) \tag{9}$$

are the unit impulse responses of the elementary pole function assuming single and periodic excitation respectively.

Thus the effect of superposition on initial amplitudes is

$$\frac{A'_n}{A_n} = \left| \frac{h'_n(t)}{h_n(t)} \right| = (1+e^{-2Y_n} - 2e^{-Y_n} \cos 2Q_n Y_n)^{-\frac{1}{2}} \tag{10}$$

where

$$Y_n = -\sigma_n T_0 = \pi \cdot B_n/F_0 \tag{11}$$

The ratio A'_n/A_n is an oscillatory function of $Q_n \cdot Y_n$ which has the maximum value

$$(A'_n/A_n)_{max} = (1-e^{-Y_n})^{-1} \qquad (2Q_n Y_n = 2n \cdot \pi) \tag{12}$$

the minimum value

$$(A'_n/A_n)_{min} = (1+e^{-Y_n})^{-1} \qquad [2Q_nY_n = (2_n+1)\pi] \tag{13}$$

and intermediate value

$$(A'_n/A_n)_{med} = (1+e^{-2Y_n})^{-\frac{1}{2}} \qquad (2Q_nY_n = n\,\tfrac{\pi}{2}) \tag{14}$$

The frequency domain correspondence of maximum superposition is the condition of the spectral peak coinciding in frequency with one of the harmonics. Minimum superposition occurs when two harmonics are spaced $F_0/2$ from the formant peak.

It is possible to derive the spectrum envelope peak amplitude within the formant from the residues of the conjugate complex poles $p = \pm j\omega_n$. This operation applied to the transform

$$H'_n(p) = \frac{\omega_{on}^2}{(p-\sigma_n)^2+\omega_n^2} \cdot \frac{1}{(1-e^{pT_0})} \tag{15}$$

provides the expression

$$[h'_n(t)]_{p=\pm j\omega_n} = \frac{\omega_n}{\sigma_n T_0} \cdot \sin\omega_n t = \hat{A}'_s \cdot \sin\omega_n t \tag{16}$$

From eq. (8) and (16) we thus derive the ratio of the spectrum peak amplitude $\hat{A}'_s$ of a periodic sound to the initial amplitude A_n of the damped oscillation following a single excitation.

$$\hat{A}'_s/A_n = 1/Y_n = F_0/\pi \cdot B_n \tag{17}$$

This ratio rizes monotonically in direct proportion to F_0. This is a property of the source function alone. Given a source function $S(p)$ comprising periodically repeated pulses of amplitude U_0/t_0 and duration t_0 where t_0 is very small the spectrum envelope will be constant and equal to $2F_0 \cdot U_0$.[x] The F_0-proportionality prevails independent of the particular pulse shaping.

---

x) This is a sine wave peak amplitude measure.

Formant amplitude parameters

From a technical point of view it would apparently be desirable to measure formant amplitudes from spectrum envelope peak values since a small shift in F_0 would little affect such measures whereas the initial amplitude of the corresponding damped oscillation may be very sensitive to the same F_0 variation as suggested by Eq. (10) - (14).

Neither of these possible formant amplitude parameters can be measured exactly by technical means. However, if the formant is isolated by a broad filter and the formant frequency F_n is larger than the fundamental frequency F_0 it may be shown that the peak value of the wave would be expected to be not much below the initial amplitude A'_n.

The standard technique of formant amplitude measurements is to use a root mean square or a mean average value measuring circuitry. The corresponding measures will be labeled A'_e and A'_a respectively.

The exact expression for the r.m.s. value of the damped sinusoid would be

$$U'_e = \sqrt{\frac{1}{T_0}\int_0^{T_0} (A'_n)^2 \cdot e^{2\sigma_n t} \cos^2(\omega_n t + \varphi'_n)dt} \qquad (18)$$

We shall approximate the calculation by integrating over the time envelope alone, thus neglecting the oscillatory behavior of the instantaneous power.

$$U'_e = \sqrt{\frac{1}{T_0}\int_0^{T_0} A'^2_n e^{2\sigma_n t} \cdot \frac{1}{2} \cdot dt} = A'_n \cdot \frac{1}{2}\sqrt{\frac{1-e^{-2Y_n}}{Y_n}} \qquad (19)$$

where $Y_n = \pi B_n / F_0$ as before. The error inherent in this approximation is smaller than 0.5 dB.

The expression (19) should be used instead of Eq. (2.6-5) of ref.(1).[x]

The exact expression for the mean average amplitude

$$U'_a = \frac{1}{T_0} \int_0^{T_0} A'_n e^{\sigma_n t} \left| \cos(\omega_n t + \varphi'_n) \right| dt \qquad (20)$$

will be subjected to the same approximation. The mean average value of a sinusoid period is $2/\pi$ times the amplitude. Thus

$$U'_a = \frac{1}{T_0} \int_0^{T_0} A'_n \cdot \frac{2}{\pi} e^{\sigma_n t} dt = A'_n \cdot \frac{2}{\pi} \frac{(1-e^{-Y_n})}{Y_n} \qquad (21)$$

Summary of formulas

Summarizing the previous calculations we have derived the following formant amplitude parameters

1. Spectrum envelope peak amplitude

$$A'_s = A_n \cdot \frac{1}{Y_n} \qquad (22)$$

2. Time function mean average value

$$U'_a = A'_n \cdot \frac{2}{\pi} \cdot \frac{1-e^{-Y_n}}{Y_n} \qquad (23)$$

3. Time function r.m.s. value

$$U'_e = A'_n \cdot \frac{1}{2} \sqrt{\frac{1-e^{-2Y_n}}{Y_n}} \qquad (24)$$

---

x) In eq. (2.6-5) of ref. (1) there is a missprint in so far as A'_n should be replaced by A_n. Furthermore eq. (2.6-5) is based on the maximum superposition effect and the r.m.s. value A_e is defined as the equivalent peak amplitude of a sinewave.

4. Time function peak value (Approximated by the initial amplitude)

$$A'_n = A_n\left[1+e^{-2Y_n} - 2\cdot e^{-Y_n}\cos 2\,Q_nY_n\right]^{-\frac{1}{2}}$$

where A_n is the initial amplitude in the case of single excitation and A'_n the initial amplitude in the case of a stationary periodic excitation. The parameter $Y_n = \pi B_n/F_0$ and $Q_n = F_n/B_n$. Formant frequency is $F_n = \omega_n/2\pi$ and formant bandwidth is $B_n = -\sigma_n/\pi$.

For practical applications of these formulas it should be pointed out that they have not been adjusted to provide equal numerical value for the special case of a spectrum dominated by a single harmonic situated at the frequency of the formant. Under these circumstances of maximum superposition and very small Y-value, i.e. high F_0, the formant is essentially a sine wave, the peak value of which is referred to by A'_s, the r.m.s. by U'_e, and average value by U'_a. They can be calibrated to provide the same numerical value by dividing A'_s by $\sqrt{2}$ and multiplying U'_a by $\pi/2\sqrt{2}$.

Discussion

If F_0 is varied, everything else held constant, it is evident that the spectrum envelope peak amplitude will increase monotonically with increasing F_0 and in direct proportion to F_0. All other parameters oscillate as functions of F_0 due to the superposition effect. This is illustrated by Fig. I-1 which pertains to measurements performed with a Brüel & Kjaer automatic level recorder measuring the peak, r.m.s. and average amplitude of the output of a single resonant circuit connected to a periodic pulse source of continuously varying fundamental frequency F_0. This pulse train was derived from the variable frequency oscillator of the Brüel & Kjaer unit.

The general trends of the recorded curves conform well with eq. (21) to (24). Al low F_0-ranges we would expect the rising F_0 result in the following average slopes:

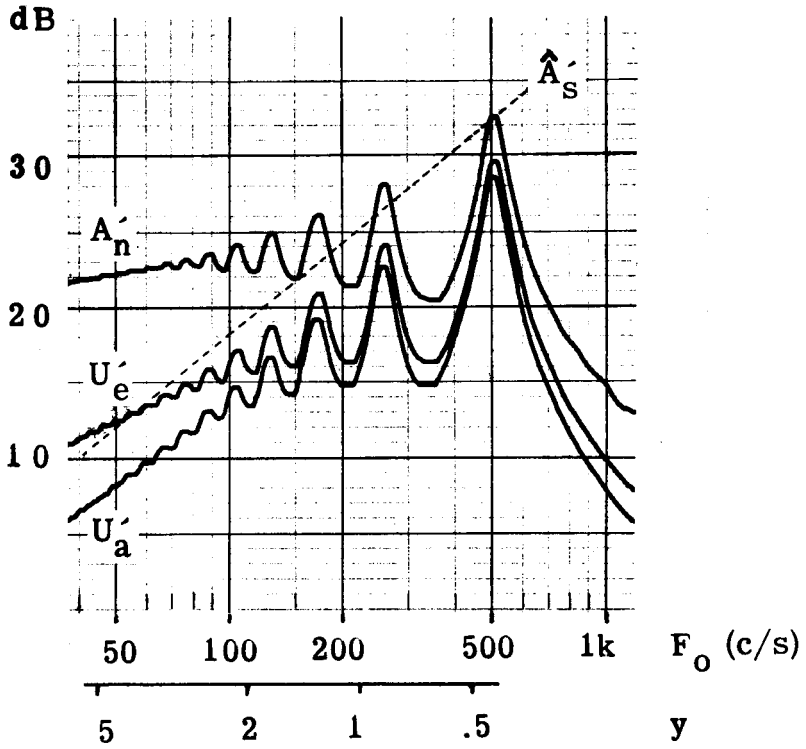

Fig. I-1. The variation of the three formant level parameters A'_n (peak value), U'_e (r.m.s. value), and U'_a (mean average value) as a function of voice fundamental frequency F_0 assuming constant waveform of source pulses. F_n=500 c/s and B_n=70 c/s.

A'_n	constant level
U'_e	+3 dB/octave
U'_a	+6 dB/octave
A'_s	+6 dB/octave

Fig. I-2 shows the calculated relations between the various formant amplitude measures according to eq. (22-25), expressed as $20\log_{10} A'_s/U'_a$, $20\log_{10} U'_e/U'_a$, and $20\log_{10} A'_n/U'_a$. The mean average U'_a was chosen as a reference since it is the most commonly measured quantity. A $Q_n = 7$ was assumed.

It may be seen that at small values of Y_n, i.e. at small B_n and high F_0, there is no difference between the three time-domain measures except for constant factors. For normal values of Y_n, i.e. in the range $0.5 < Y_n < 2$, one would get rather small differences between U'_a and U'_e. There is thus not a pronounced motivation for adopting the r.m.s. measure.

The oscillatory nature of the curve labeled A'_s is due to the normalization with regard to U'_a. It is of interest to note that U'_a has the advantage of approximating A'_s at high values of Y_n, i.e. under conditions of very low F_0, as was also apparent from Fig. I-1.

A possible argument in favor of the peak measure A'_n is that it would have a direct correspondence to the overall scale factor of the driving source function U_0 under conditions of phonation with constant glottal pulse shape and varying F_0 in a low F_0-range. Providing F_0 is separately controlled it follows that the intensity radiations then are completely conditioned by F_0.

The oscillatory nature of all three time domain measures due to the superposition effect is very large and of the order of ±6 dB at $B_n = 50$ c/s and $F_0 = 250$ c/s. It should also be pointed out that these oscillations as a function of the parameter $Y_n = \pi B_n/F_0$ have nothing to do with the F_0 ripple in intensity curves under conditions of insufficient smoothing but appear as additional variations.

The implication for automatic analysis-synthesis procedures in a formant vocoder is that all three time domain methods of measuring

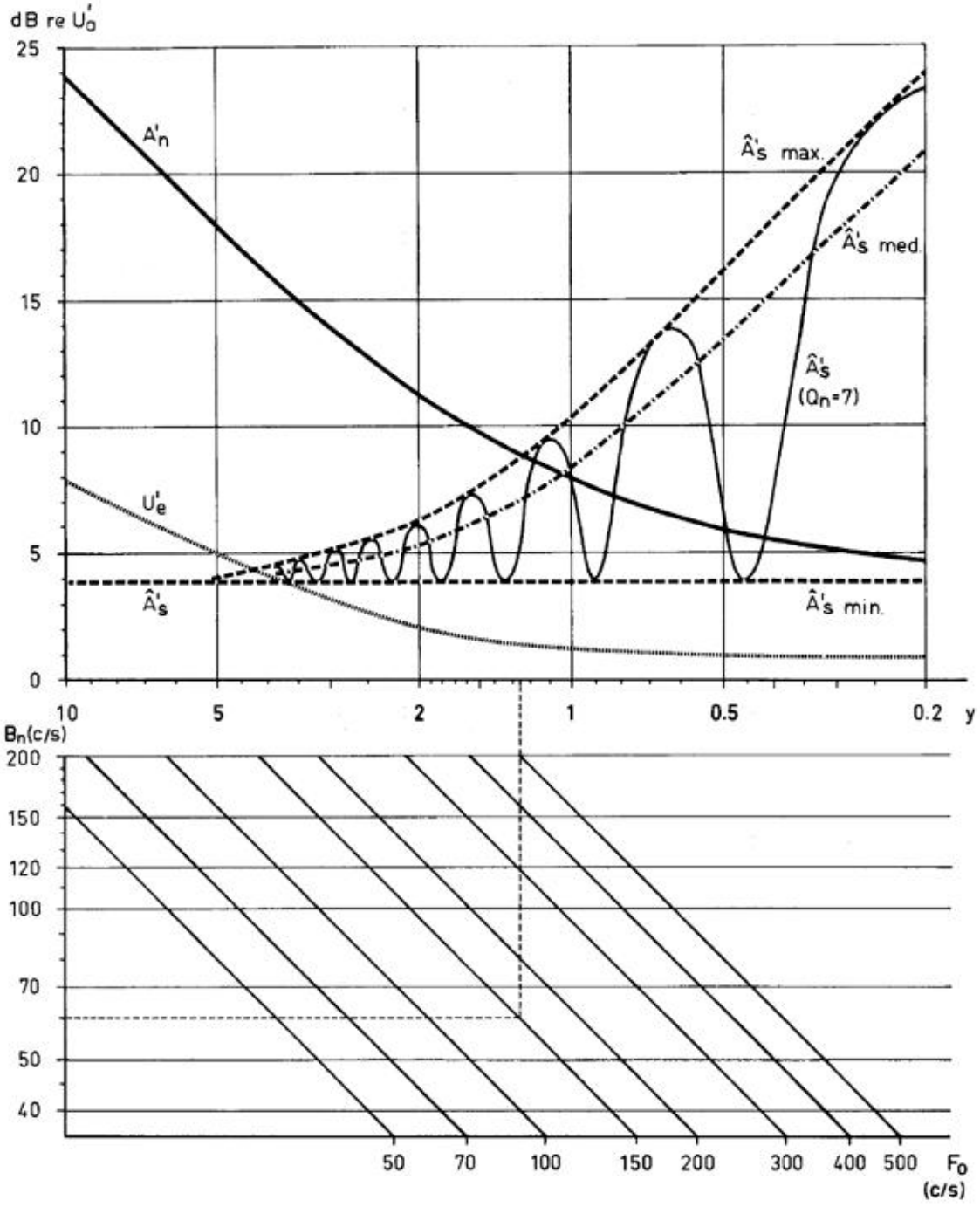

Fig. I-2. Diagram of formant levels A'_n (peak value), U'_e (r.m.s.) value), and $\hat{A}'_S$ (spectrum envelope peak value) relative to U'_a (mean average value) for an idealized one-formant "speech wave" in the steady state.
Below: Auxiliary chart for the evaluation of $Y_n = \pi B_n/F_0$. Dashed line exemplifies the determination of Y_n for B_n=60 c/s and F_0=150 c/s.

formant amplitude will potentially create false variations in the amplitude parameters controlling the synthesis process. The theoretically correct measure would have to be proportional to A'_s/F_0, i.e. the spectrum peak envelope divided by the voice fundamental frequency.

From a practical point of view, however, it remains to find out how prominent the superposition effect is in natural speech and what perceptual role it plays.

G. Fant, J. Liljencrants

References:

(1) Fant, G.: "Acoustic analysis and synthesis of speech with applications to Swedish", Ericsson Technics 15, No. 1 (1959) pp. 3-108.

(2) Fintoft, K., Lindblom, B., and Mártony, J.: "Measurements of formant level in human speech", article in this QPSR, pp. 9-17.

(3) House, A.S.: "A note on optimal vocal frequency", J. of Speech and Hearing Research 2 (1959) pp. 55-60.

B. MEASUREMENTS OF FORMANT LEVEL IN HUMAN SPEECH

Introduction

It has been suggested that formant be defined as a normal mode of vibration of the vocal tract (1). A formant is thus the manifestation of a transfer function conjugate complex pole. Spectrographically this manifestation usually has the form of a spectral energy maximum which may fail to appear, in the presence of, for instance, too high a fundamental frequency, zeros, large formant bandwidths, or voice source irregularities.

Current terminology, on the other hand, assigns three dimensions to a formant: frequency, bandwidth, and amplitude. A conjugate complex pole is characterized by two: it has a real part (bandwidth) and an imaginary part (frequency). It is common practice to